СОДЕРЖАНИЕ

Том 508, 2022

СПЕЦИАЛЬНЫЙ ВЫПУСК: ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ИССЛЕДОВАТЕЛЬСКИХ ЦЕНТРОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

EDN: LAWVQG
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 3–5

ИИ: ПОЧЕМУ МАТЕМАТИКА?
Предисловие главного редактора журнала «Доклады Российской академии наук. Математика, информатика, процессы управления» академика РАН Алексея Львовича Семенова

DOI: 10.31857/S2686954322070268

Инициатива данного выпуска принадлежит Сберу, и редакция «Докладов РАН. Математика, информатика, процессы управления» благодарна ему за это.

Однако далеко не совпадением является то, что Отделение математических наук РАН в течение последней пары лет постоянно обращается к проблематике искусственного интеллекта (ИИ), что мехмат МГУ запустил магистерскую программу “Цифровые технологии и искусственный интеллект”. Причины – во всем комплексе взаимоотношений математики и ИИ.

Если говорить об определении того, что такое ИИ, то я отношусь к категории специалистов, которые считают, в соответствии с буквальным пониманием русского термина, что ИИ – это средства автоматизации интеллектуальной деятельности человека. Далее можно выделять в ИИ часть, занимающуюся автоматизацией рациональной деятельности, какой является, например, решение задач по матанализу «из Демидовича», и интуитивную часть, которая, например, умеет распознавать лица на фотографии.

Как известно, уже в 1960-е гг. была, в основном (рационально), решена проблема построения систем компьютерной алгебры, принципиальный же прогресс в (интуитивном) решении широкого круга задач распознавания был достигнут только в XXI веке благодаря алгоритмам машинного обучения. Базовые принципы построения последних которых были разработаны уже к концу 1950-х гг., а существенная идейная часть (байесовский поход) еще значительно старше. Именно с этим, интуитивным ИИ многие сегодня связывают общее представление об ИИ, с одной стороны, и взгляды о небольшой “математико-емкости” алгоритмов ИИ, с другой стороны. Действительно, очередной прогресс последние годы был связан, в первую очередь, с ростом вычислительных мощностей и накоплением больших данных, во вторую – с новыми алгоритмами. При этом не используются какие-то новые теоремы, или даже нетривиальные вычислительные идеи. Можно сказать, что Человечество, вложившись в огромное количество “математических стартапов”, теперь зарабатывает на очень узком и простеньком сегменте, как это и бывает со стартапами. В определенной степени так и есть, однако, при более глубоком рассмотрении мы видим, что если бы математики XX века не существовало, то инженерам искусственного интеллекта “следовало бы ее выдумать”. С другой стороны, сейчас нарастает потребность в системах доверенного ИИ, сверхнадежного ИИ, объясняющего ИИ, которым посвящены многие статьи в нашем выпуске журнала. И есть серьезные основания полагать, что здесь потребуется новая математика: как “новая старая”, дополняющая уже использованную, например, элементами математической ло-

гики и математической лингвистики, так и “новая новая”, для которой еще нет даже названия. В частности, об этом математики, физики и философы рассказывали на AI Journey в прошлом году [1].

Если же говорить о направлении влияния от ИИ к математике, то в последние десятилетия последовательно расширяются области математики, в которых ИИ оказывается полезным инструментом. Очевидна критическая роль компьютера в окончательном решении или существенном продвижении для классических задач теории чисел [2]. Рациональный ИИ повышает надежность математических доказательств: например, возможно, самый сложный отдельный результат современной математики – классификация простых конечных групп, приобретает такую надежность благодаря усилиям по автоматизации ключевых конструкций [3]. Проблема надежности и воспринимаемости доказательств стала толчком и к построению Воеводским нового фундамента для всей математики [4], пожалуй, наиболее серьезной, после альтернативной теории множеств П. Вопенки [5] и теории топосов [6] попытки новых оснований. Видимо, закономерно, что в [7] математика сразу строится как человеко-машинный объект. Конечно, нельзя не упомянуть и машинную генерацию данных экспериментальной математики [8] и основанных на этих данных гипотез [9–13].

Мне представляется важным и еще одно направление использования ИИ в математике. Одним из первых подходов к доказательству в математике было “Смотри!” [14]. Представляется, что визуальная, графическая ФОРМУЛИРОВКА для некоторых математических утверждений сегодня может стать единственной, воспринимаемой человеком. Мне кажется, что так обстоит дело с утверждениями, относящимися к внешним биллиардам [15], где утверждение формулируется на картинке, картинка строится компьютером, утверждение о соответствии картинки математической реальности, как и доказательство правильности работы компьютерной программы строятся человеком.

Наконец, абсолютно принципиальную роль цифровые технологии, в частности ИИ, начинают играть в математическом образовании. Они, прямо с начальной школы, позволяют сделать школьную математику действительно интересным, интеллектуальным, экспериментальным, творческим и одновременно полезным предметом для всех учеников. Математический кружок и матшколы, ранее ориентированные только на высокомотивированных, могут стать моделью массового математического образования [16,17]. Мы надеемся подготовить отдельный дополнительный выпуск “Докладов” с полученными здесь результатами.

БЛАГОДАРНОСТИ

Пользуюсь случаем выразить личную благодарность за возможность совместной работы с Германом Оскаровичем Грефом, развивающим Сбер и все его созвездие, как сильнейшую научную и образовательную силу в стране. Также я и мои коллеги в МГУ, РАН и “Докладах” выражаем признательность: Александру Александровичу Ведяхину, Альберту Рувимовичу Ефимову, Максиму Алексеевичу Еременко и всей замечательной команде Сбера. Уверены, что AI Journey будет развивать свою роль как крупнейший форум в сфере ИИ.

СПИСОК ЛИТЕРАТУРЫ

1. *Анохин К.В., Новоселов К.С., Смирнов С.К. и др.* Искусственный интеллект для науки и наука для искусственного интеллекта // Вопросы философии. 2022. № 3. С. 93–105.
2. *Вавилов Н.А.* Компьютер как новая реальность математики. Части I, II, III // Компьютерные инструменты в образовании. 2020. № 2, 3, 4.
3. *Théry L.* Feit thomson proved in coq // Microsoft Research Inria Joint Centre. https://web.archive.org/web/20161119094-854/http://www.msr-inria.fr/news/feit-thomson-proved-in-coq/
4. *Vladimir Voevodsky.* An experimental library of formalized Mathematics based on the univalent foundations // Mathematical Structures in Computer Science. Cambridge University Press, 2015. V. 25. P. 1278–1294.
5. *Вопенка П.* Альтернативная теория множеств: новый взгляд на бесконечность. Пер. со словац. — Новосибирск: Изд-во Института математики, 2004. Оригинал: Petr Vopěnka. Mathematics in the Alternative Set Theory. Leipzig: Teubner Verl., 1979. ASIN B0006E3AXY
6. *Johnstone P.T.* Sketches of an Elephant: A Topos Theory Compendium // Oxford Science Publications, Oxford, 2002.
7. UniMath: This coq library aims to formalize a substantial body of mathematics using the univalent point of view // Univalent Mathematics. https://github.com/UniMath/UniMath
8. *Матиясевич Ю.В.* Асимптотическая структура собственных чисел и собственных векторов некоторых треугольных ганкелевых матриц // Чебышевский сборник, 21, 1, 2020. С. 259–272.
9. *Birch B.J., Swinnerton-Dyer H.P.F.* Notes on Elliptic Curves (II) // Journal for Pure and Applied Mathematics, vol. 1965, no. 218, 1965. P. 79–108. https://doi.org/10.1515/crll.1965.218.79
10. *Borwein J.M., Bailey D.* Mathematics by experiment: Plausible reasoning in the 21st century. A.K. Peters/CRC Press, 2008. 288 pp.
11. *Raayoni G., Gottlieb S., Manor Y., et al.* Generating conjectures on fundamental constants with the Ramanujan Machine // Nature, 2021 Feb., 590 (7844). P. 67–73.

https://doi.org/10.1038/s41586-021-03229-4.
https://pubmed.ncbi.nlm.nih.gov/33536657/

12. *Davies A., Veličković P., Buesing L.*, et all. Advancing mathematics by guiding human intuition with AI // Nature, 2021, 600 (7887). Pp. 70. https://doi.org/10.1038/s41586-021-04086-x
13. DeepMind (Google) – AlphaGo // https://www.deepmind.com/research/highlighted-research/alphago
14. *Успенский В.А.* Простейшие примеры математических доказательств // 2-е изд., стереотипное. М.: Изд-во МЦНМО, 2012. 56 с. ISBN 978-5-94057-879-6.
15. *Рухович Ф.Д.* Внешние биллиарды вне правильных многоугольников: ручной случай // Известия РАН. Сер. матем., 86, 3, 2022. С. 105–160. https://doi.org/10.4213/im8972
16. *Семенов А.Л., Булин-Соколова Е.И, Муранов А.А., и др.* Цифровые технологии в начальной школе. Вход в будущий мир // Информатизация образования и методика электронного обучения: цифровые технологии в образовании. Материалы VI Международной науч. конф., г. Красноярск, 20–23 сентября 2022 г. В 3 ч. Ч. 2 / под общ. ред. М.В. Носкова. – Красноярск : КГПУ им. В.П. Астафьева, 2022. С. 325–329. ISBN 978-5-907558-24-3.
17. *Константинов Н.Н., Семенов А.Л.* Результативное образование в математической школе // Чебышевский сборник, т. XXII, вып. 1(77), 2021. С. 413–446. https://doi.org/10.22405/2226-8383-2021-22-1-413-446

EDN: IIRWCP
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 6

ВСТУПИТЕЛЬНОЕ СЛОВО КОМАНДЫ AI JOURNEY

DOI: 10.31857/S2686954322070256

Дорогие друзья!

Представляем вашему вниманию научный сборник Международной конференции по искусственному интеллекту AI Journey 2022. Мы рады, что в этом году сборник выходит в рамках специального выпуска журнала “Доклады Российской академии наук. Математика, информатика, процессы управления”. Журнал публикует статьи о научных исследованиях в области математики, естественных и технических наук и является одним из ведущих профильных изданий в данной области.

В 2022 г. конференция AI Journey прошла уже в 7-й раз. Основной фокус программы был посвящен развитию науки в области искусственного интеллекта. Ведущие эксперты представили доклады на актуальные темы в области искусственного интеллекта и машинного обучения, многие из которых можно отнести к state-of-the-art направлениям, — мультимодальные и мультиязыковые модели, генеративные модели, трансформеры, новые архитектуры.

В данном сборнике мы собрали статьи об итогах развития AI в рамках деятельности ведущих исследовательских центров страны. Также в сборник вошли научные статьи с передовыми разработками в области искусственного интеллекта. Основными тематиками сборника стали глубокое и мультиагентное обучение, компьютерное зрение, большие языковые модели, сильный искусственный интеллект.

Искусственный интеллект сегодня является ключевой технологией, которая во многом задает вектор и динамику развития человечества. Объем научных публикаций в области искусственного интеллекта и машинного обучения растет из года в год. Поэтому для научного сообщества крайне важно быть в курсе последних достижений коллег, обмениваться опытом и лучшими практиками для создания прорывных AI-решений.

Приглашаем вас в путешествие в мир искусственного интеллекта!

С уважением,

Команда AI Journey

“Масштаб конференции AI Journey год от года растет — расширяется спектр представленных на ней докладов, добавляются тематические треки, к участию присоединяются новые представители мирового научного и бизнес-сообщества. Мы стремимся делать все для того, чтобы каждая новая конференция была все более интересной и значимой. Так, в этом году мы усилили научное направление конференции, запустив трек AI Journey Science и впервые опубликовав этот сборник. Мы планируем и дальше развивать AI Journey как международную научную площадку, и возможность публикации в нашем научном сборнике, уверен, станет для многих ученых дополнительным стимулом для участия в будущих конференциях”.

Александр Ведяхин, первый заместитель Председателя Правления Сбербанка

EDN: QOWDGI
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 7–12

РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ИССЛЕДОВАТЕЛЬСКИХ ЦЕНТРОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

УДК 004.8

МЕЖОТРАСЛЕВЫЕ ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ПОИСК И РЕАЛИЗАЦИЯ ЭФФЕКТИВНЫХ РЕШЕНИЙ

© 2022 г. А. В. Корнаев[1,*], И. А. Никанов[1], Р. Ф. Кулеев[1]

Представлено академиком РАН А.А. Шананиным
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

Большинство исследований в сфере искусственного интеллекта связано с разрешением следующего противоречия. С одной стороны, методы глубокого обучения обладают универсальностью и могут быть применены в различных областях знаний благодаря общности основных математических и алгоритмических идей, программных средств их реализации, возможности трансфера ранее полученных результатов обучения. С другой стороны, процесс обучения для решения конкретных задач требует специализированных качественно размеченных данных, а достижение высокой точности – применения оригинальных алгоритмических решений и правильной настройки множества гиперпараметров. Работа Исследовательского центра в сфере искусственного интеллекта Университета Иннополис направлена на разрешение этого противоречия путем создания алгоритмического ядра и соответствующих программно-аппаратных средств, объединяющих решение разноплановых межотраслевых задач. Научная работа центра направлена на создание достаточных оснований для решения практических задач. В данной статье представлены основные результаты научной и практической работы центра в 2022 г.

Ключевые слова: искусственный интеллект, фреймворк, обработка изображений, обучение с подкреплением, драг дизайн, дизайн материалов, сверточные нейронные сети, графовые нейронные сети

DOI: 10.31857/S2686954322070116

1. ВВЕДЕНИЕ

Реализация программы деятельности исследовательского центра в сфере искусственного интеллекта “Межотраслевые технологии искусственного интеллекта для задач цифровой трансформации приоритетных отраслей экономики” Университета Иннополис направлена на ускорение перехода приоритетных отраслей к цифровой экономике, их цифровой трансформации и решения стратегических задач индустриальных компаний за счет решения фундаментальных задач машинного обучения, разработки и коммерциализации аппаратно-программного обеспечения, основанного на новых технологиях искусственного интеллекта. Деятельность центра связана с развитием трех основных направлений:

– поддержка принятия врачебных решений;

– поддержка принятия решений в функционировании системы безопасности предприятия;

– поддержка принятия решений в поиске материалов с заданными свойствами.

2. СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ ВРАЧЕБНЫХ РЕШЕНИЙ

Искусственные нейронные сети являются эффективным инструментом выявления признаков и их применение в медицине связано, в первую очередь, с обработкой медицинских изображений. Ниже представлены основные научные результаты центра в 2022 г., а также сведения о практической реализации ранее полученных и новых научных результатов.

2.1. Поиск ключевых точек на медицинских изображениях для выявления аномалий суставов с применением метода обучения с подкреплением

Новизна исследования заключается в создании мультиагентной модели поиска ключевых точек на трехмерных изображениях.

Метод исследования основан на комбинированном применении искусственных нейронных сетей архитектур U-Net и FPN для сегментации трехмерных изображений и поиска областей

[1] *Исследовательский центр в сфере искусственного интеллекта, Университет Иннополис, Иннополис, Россия*
*****E-mail: a.kornaev@innopolis.ru*

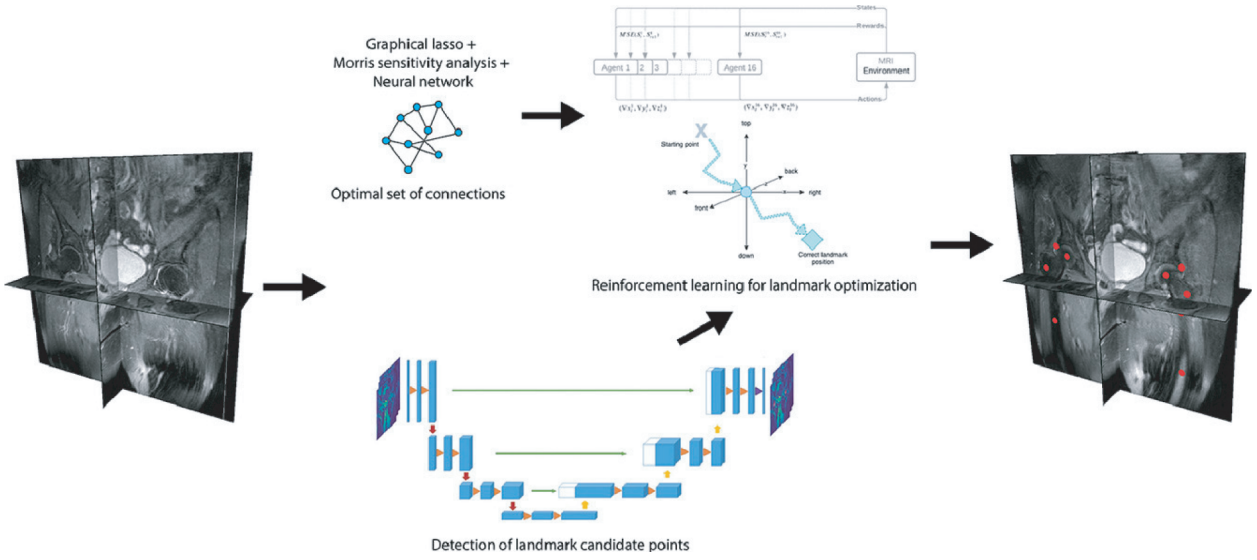

Рис. 1. Предлагаемый фреймворк поиска ключевых точек МРТ изображений, который включает 3 этапа: сегментации для поиска областей интереса, определения начальных положений ключевых точек и применения мультиагентной модели обучения с подкреплением для корректировки положения ключевых точек [1].

ожидаемых положений ключевых точек, и сетей архитектур DeepQN, DDPG, TD3, A2C для обучения с подкреплением нескольких агентов поиска уточненных положений ключевых точек (рис. 1) [1].

Значимость исследования связана с возможностью высокоточного определения положений ключевых точек в области тазобедренного сустава для последующей диагностики заболеваний суставов.

Перспективы развития тематики связаны с созданием быстродействующих высокоточных алгоритмов детектирования областей и определения координат особых точек на трехмерных изображениях по данным МРТ или КТ исследований.

2.2. Физически обоснованное машинное обучение для моделирования течений неньютоновских жидкостей

Новизна исследования заключается в создании теоретических основ и инструментария моделирования течений неньютоновских жидкостей, прежде всего, крови, а также реомагнитных жидкостей в каналах искусственных и естественных гидродинамических систем.

Метод исследования основан на минимизации предложенного авторами целевого функционала мощности внутренних сил, реализованной с помощью глубокой сверточной сети архитектуры U-Net (рис. 2). В процессе обучения сеть использует одно входное изображение и не нуждается в датасете. Точность сети связана с точностью численного дифференцирования и интегрирования по значениям интенсивности пикселей выходного изображения [2].

Значимость исследования связана с возможностью разработки универсального программного обеспечения для обработки сегментированных изображений, в том числе, медицинских изображений, и моделированию течений жидкостей. В частности, возможно решение задач о доставке лекарственных средств в составе физиологических жидкостей, обладающих неньютоновскими и реомагнитными свойствами.

Перспективы развития тематики связаны с созданием алгоритмов и моделей для обработки трехмерных изображений по данным МРТ или КТ исследований, а также разработка новых методов медицинских исследований типа ‘in-silico’.

2.3. Другие научные работы в области поддержки принятия медицинских решений

Помимо перечисленных работ, в центре проводились исследования по обработке гиперспектральных изображений для детектирования следов крови [3], поиску эффективных средств детектирования и классификации лейкемии [4] и некоторые другие.

2.4. Практическая реализация систем поддержки принятия врачебных решений

Внедрение сервисов в работу медицинских организаций обеспечивает качественный скачок как за счет выдвижения высоких требований к медицинским изделиям, так и за счет быстрого повышения качества и количества обучающих

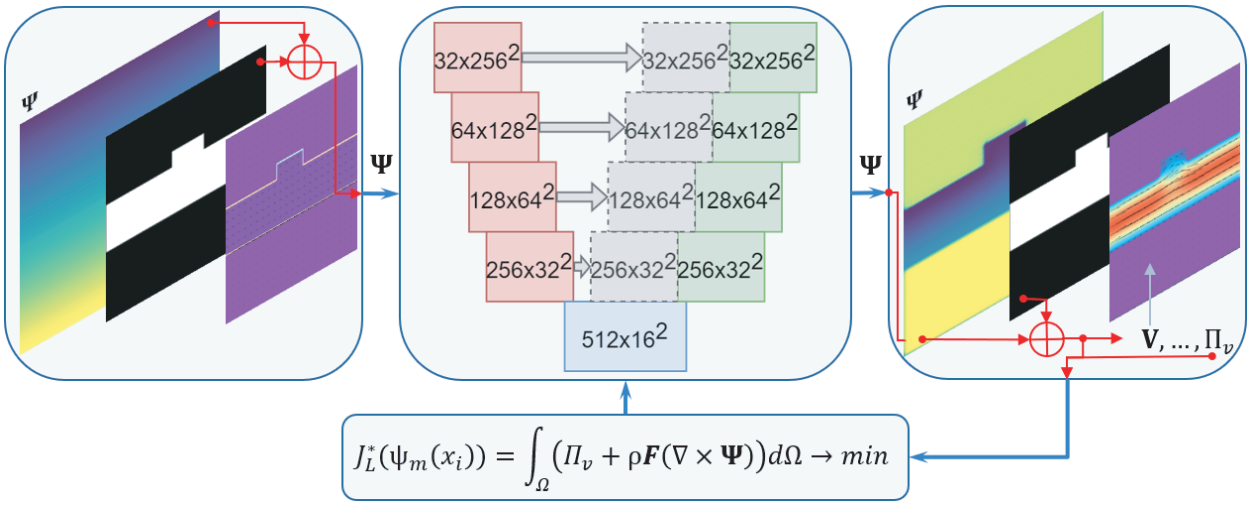

Рис. 2. Архитектура U-Net и алгоритм на основе представления области течения в виде изображения. Сеть U-Net получает на вход начальное распределение для неизвестной пси-функции с масками границ течения и рассчитывает уточненное распределение пси-функции путем минимизации функционала [2].

данных. Также это обеспечивает сокращение расходов на лечение за счет ранней диагностики заболеваний.

В настоящее время функционирует сервис распознавания патологий легких AI Radiology [5], ведется разработка сервисов маммографических и патоморфологических исследований. Новые практические разработки защищены как объекты интеллектуальной собственности [6].

3. СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ В ФУНКЦИОНИРОВАНИИ СИСТЕМЫ БЕЗОПАСНОСТИ ПРЕДПРИЯТИЯ

Обеспечение безопасности на предприятиях представляет собой многоуровневую задачу, связанную с анализом данных изображений различного масштаба: от спутниковых снимков до данных камер видеонаблюдения, безопасностью передачи и хранения данных, обеспечению безопасных и эффективных логистических процессов, мониторингом состояния персонала и оборудования предприятия.

Ниже описаны некоторые научные работы и практические результаты, полученные в ходе решения задач проекта.

3.1. Планирование траекторий летательных аппаратов

Новизна исследования заключается в разработке метода непрерывно оптимизируемой траектории летательного аппарата при движении в неизвестных условиях.

Метод основан на применении двух планировщиков: глобального и локального. Первый уточняет начальную опорную траекторию в случаях, когда траектория проходит через препятствие или вблизи него, и позволяет локальному планировщику рассчитать оптимальную траекторию, если глобальный планировщик не может выполнить задачу. Глобальный планировщик включает в себя два подхода к выпуклому программированию: Глобальный планировщик в основном фокусируется на производительности в реальном времени и обходе препятствий, в то время как предлагаемая формулировка локального планировщика на основе прогнозирующего управления с ограниченной нелинейной моделью обеспечивает безопасность, динамическую осуществимость и точность отслеживания базовой траектории для низкоскоростных маневров (рис. 3) [7].

Значимость исследования связана с возможностью обеспечения безопасных полетов беспилотных летательных аппаратов в условиях неопределенности и наличия препятствий, а также с разработкой быстродействующих алгоритмов, функционирующих на микрокомпьютерах.

Перспективы развития тематики связаны с совершенствованием алгоритмов локального планировщика, применением к высокоскоростным летательным аппаратам, обеспечивающим действие систем внешнего наблюдения за безопасностью на производстве.

В рамках научной работы центра также проводились исследования по обработке снимков ландшафта [8], защите данных при использовании облачных ресурсов [9], человеко-машинному взаимодействию [10], тестированию моделей искусственного интеллекта [11] и другим направлениям.

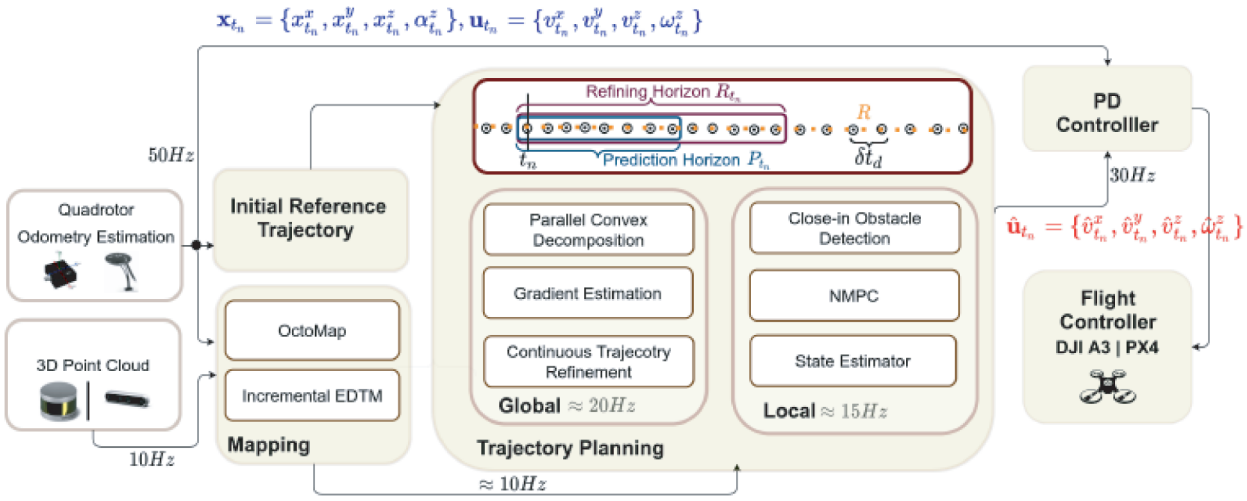

Рис. 3. Архитектура планировщика траекторий. Глобальный и локальный планировщики функционируют параллельно [7].

3.2. Практическая реализация систем поддержки принятия решений в функционировании системы безопасности предприятия

Основой разрабатываемых сервисов является обработка данных изображений и их последовательностей, данных сигналов мультисенсорных измерений, аудиоданных, а также табличных и текстовых данных.

Совместно с ПАО “Татнефть” разрабатывается интеллектуальная система поддержки принятия решений по обеспечению безопасности и контролю хода строительства, эксплуатации инфраструктуры с применением технологий компьютерного зрения. Реализована первая очередь аппаратно-программного комплекса в части сервиса распознавания разливов нефти и обнаружений строительных работ в охранных зонах. Предложены решения в разработке системы помощи при визуальной инспекции транспортных средств. Партнеры в разработке АО “Аэрофлот”, АО “Синара-Транспортные Машины”.

Совместно с АО “Почта России”, ООО “Вайлдберриз” разрабатываются технологии искусственного интеллекта для контроля внутрискладских операций в фулфилменте: контроль движения, снижение ошибок при комплектовании клиентских заказов, автоматизация работы с претензиями клиентов по поставленным товарам.

Совместно с Министерством промышленности и торговли РФ разрабатывается сервис предоставления услуг в электронном виде, осуществляется подбор аналогов в каталогах продукции.

На произведенную интеллектуальную собственность получены охранные документы [12–15].

4. СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ В ПОИСКЕ МАТЕРИАЛОВ С ЗАДАННЫМИ СВОЙСТВАМИ

Поиск новых материалов является одной из наиболее перспективных задач для искусственного интеллекта на ближайшие десятилетия. Первые важные результаты в этой области, создавшие достаточные основания для развития тематики в условиях работы центра, были получены в 2021 г. в ходе проведения открытого международного конкурса по поиску новых видов катализаторов, на котором команда Университета Иннополис заняла второе место, сразу после команды компании Microsoft [16].

Ниже описаны некоторые научные работы и практические результаты, полученные в ходе решения задач проекта в текущем году.

4.1. Пересмотр нейронных сетей графов передачи для разработки катализатора

Новизна исследования заключается в разработке архитектур физически информированных нейронных сетей для моделирования молекулярных структур.

Метод основан на применении нескольких вариантов графовых нейронных сетей, включая сверточную, для предсказания энергии. Предложенные архитектуры устойчивы к переобучению и (рис. 4) [17].

Значимость исследования связана с тем, что предложенные методики могут быть применены для предсказания экспериментальных и квантовых химических свойств широкого спектра материалов и молекул.

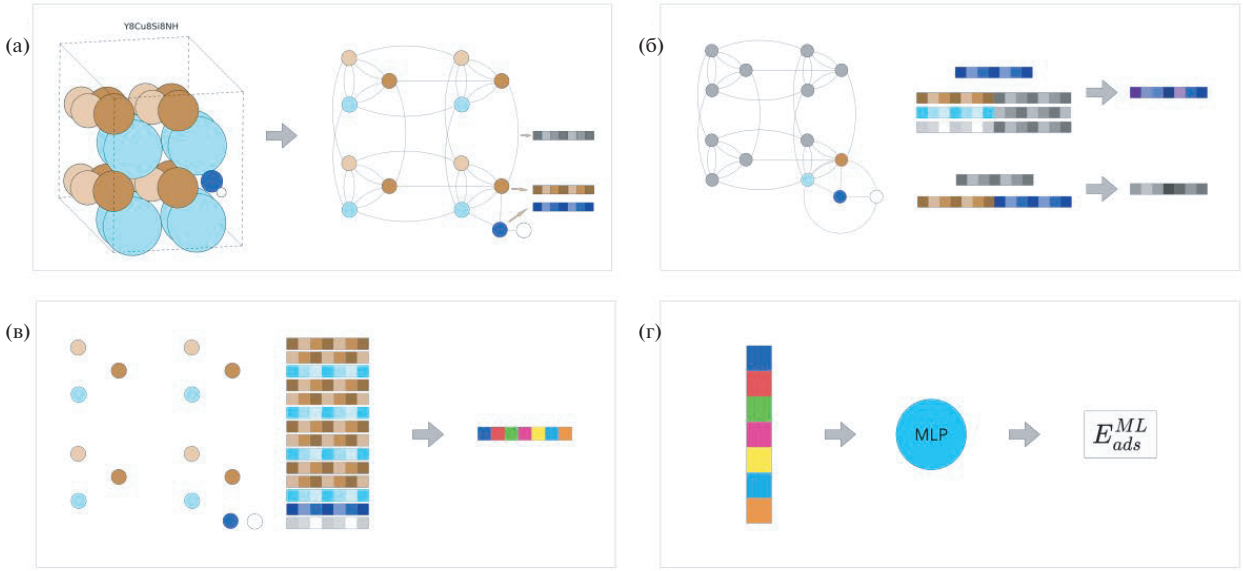

Рис. 4. Преобразование молекулярной структуры в граф с последующим предсказанием значения энергии [17].

Перспективы развития тематики связаны с разработкой новых материалов в химической и нефтехимической промышленности (полимеры, каталитические материалы, присадки, синтетические смазочные материалы). Химико-фармацевтической промышленности (новые синтетические лекарственные средства).

4.2. Практическая реализация систем поддержки принятия решений в поиске материалов с заданными свойствами

Совместно с ООО “СИБУР”, АО “ТАНЕКО”, ГК “ХимРар” ведется разработка платформы, помогающей исследовательским лабораториям и частным исследователям сократить цикл освоения новых продуктов за счет уменьшения числа экспериментов на 30–40% и сокращения потребления сырья на 10–20%. На произведенную интеллектуальную собственность получен охранный документ [18].

СПИСОК ЛИТЕРАТУРЫ

1. *Bekkouch I.E.I. et al.* Multi-landmark environment analysis with reinforcement learning for pelvic abnormality detection and quantification // Med Image Anal. Elsevier, 2022. Vol. 78. P. 102417.
2. *Kornaeva E. et al.* Physics-based loss and machine learning approach in application to non-Newtonian fluids flow modeling // 2022 IEEE Congress on Evolutionary Computation, CEC 2022 – Conference Proceedings. Institute of Electrical and Electronics Engineers Inc., 2022.
3. *Butt M.H.F. et al.* A Fast and Compact Hybrid CNN for Hyperspectral Imaging-based Bloodstain Classification // 2022 IEEE Congress on Evolutionary Computation, CEC 2022 – Conference Proceedings. Institute of Electrical and Electronics Engineers Inc., 2022.
4. *Das P.K. et al.* A Systematic Review on Recent Advancements in Deep and Machine Learning Based Detection and Classification of Acute Lymphoblastic Leukemia // IEEE Access. 2022. Vol. 10. P. 81741–81763.
5. AI Radiology [Electronic resource]. URL: https://ai.innopolis.university/airadiology/ (accessed: 27.10.2022).
6. *Колдашов А.С., Карпов И.А.* Программа автоматической аннотации медицинских изображений и формирования деперсонифицированных наборов данных для обучения искусственных нейронных сетей в составе продуктов с искусственным интеллектом: pat. 2021680668 USA. Свидетельство о регистрации программы для ЭВМ, 2021.
7. *Kulathunga G. et al.* Optimization-Based Trajectory Tracking Approach for Multi-Rotor Aerial Vehicles in Unknown Environments // IEEE Robot Autom Lett. Institute of Electrical and Electronics Engineers Inc., 2022. Vol. 7, № 2. P. 4598–4605.
8. *Ramadas M., Abraham A.* Segregating Satellite Imagery Based on Soil Moisture Level Using Advanced Differential Evolutionary Multilevel Segmentation // 2022 IEEE Congress on Evolutionary Computation, CEC 2022 – Conference Proceedings. Institute of Electrical and Electronics Engineers Inc., 2022.
9. *Hassan J. et al.* The Rise of Cloud Computing: Data Protection, Privacy, and Open Research Challenges – A Systematic Literature Review (SLR) // Comput Intell Neurosci. Hindawi Limited, 2022. Vol. 2022. P. 1–26.
10. *Kusal S. et al.* AI-based Conversational Agents: A Scoping Review from Technologies to Future Directions // IEEE Access. 2022. PP(99):1-1.
11. *Bajaj A. et al.* Test Case Prioritization, Selection, and Reduction Using Improved Quantum-Behaved Parti-

cle Swarm Optimization // Sensors. MDPI AG, 2022. Vol. 22, № 12. P. 4374.

12. *Искалиев Р.Д.* Программа для тестирования модели глубокого обучения для обнаружения дефектов турбин авиационных двигателей по фотографиям с использованием компьютерного зрения: pat. 2022662476 USA. Свидетельство о регистрации программы для ЭВМ, 2022.
13. *Гарипов Р.И.* Программа локализации видимых дефектов на боковой части транспортного средства: pat. 2022666320 USA. Свидетельство о регистрации программы для ЭВМ, 2022.
14. *Ахтямов Р.А., Сологуб Е.С.* Инструмент конфигурации микросервисов, использующих методы статистического анализа и машинного обучения. Свидетельство о регистрации программы для ЭВМ, 2022.
15. *Сологуб Е.С., Посашков И.Ф.* Программа для предобработки и хранения данных из внешних источников для последующего использования в аналитических сервисах с применением искусственного интеллекта: pat. 2022668980 USA. Свидетельство о регистрации программы для ЭВМ, 2022.
16. *Das A. et al.* The Open Catalyst Challenge 2021: Competition Report // Proceedings of Machine Learning Research. PMLR, 2022. Vol. 176. P. 29–40.
17. *Faleev M. et al.* Revising Message Passing Graph Neural Networks for Catalyst Design. 2022.
18. *Лукин Р.Ю., Фалеев М.А., Григорьев Р.А.* Программа для автоматического построения модели структура-активность для предсказания активности малых молекул в процессах ингибирования биологического таргета MCL-1: pat. 2021680832 USA. Свидетельство о регистрации программы для ЭВМ, 2021.

EDN: CVIVCS
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 13–18

РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ИССЛЕДОВАТЕЛЬСКИХ ЦЕНТРОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

УДК 004.8

ДОВЕРЕННЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: ВЫЗОВЫ И ПЕРСПЕКТИВНЫЕ РЕШЕНИЯ

© 2022 г. Д. Ю. Турдаков[1,4,6,*], академик РАН А. И. Аветисян[1,4,5,6], К. В. Архипенко[1], А. В. Анциферова[1], Д. С. Ватолин[4], С. С. Волков[1], А. В. Гасников[1,5], Д. А. Девяткин[4,7], М. Д. Дробышевский[1,5], А. П. Коваленко[1], М. И. Кривоносов[1], Н. В. Лукашевич[4], В. А. Малых[5], С. И. Николенко[6], И. В. Оселедец[2,3], А. И. Перминов[1,4], И. В. Соченков[2,7], М. М. Тихомиров[4], А. Н. Федотов[4], М. Ю. Хачай[8]

Поступило 28.10.2022 г.
После доработки 29.10.2022 г.
Принято к публикации 01.11.2022 г.

Широкое внедрение технологий искусственного интеллекта привело к возникновению новых угроз, эффективное противодействие которым не может быть реализовано текущими средствами разработки безопасного ПО. Для ответа на этот вызов в 2021 г. в рамках федерального проекта “Искусственный интеллект” на базе ИСП РАН был создан Исследовательский центр доверенного искусственного интеллекта, задачами которого является создание научно-технологической базы для обеспечения доверия к технологиям ИИ. В статье рассмотрены риски применения технологий искусственного интеллекта, а также представлены направления и промежуточные результаты работ Центра доверенного искусственного интеллекта ИСП РАН.

Ключевые слова: доверенный искусственный интеллект, атаки на машинное обучение, объяснимый искусственный интеллект, доверенные интеллектуальные системы

DOI: 10.31857/S2686954322070207

1. ВВЕДЕНИЕ

Современные интеллекутальные системы используют целый стек технологий, который состоит из методов и алгоритмов искусственного интеллекта, фреймворков машинного обучения (например, TensorFlow, PyTorch), а также инфраструктурных решений для их поддержки (облачные системы, специализированные аппаратные системы и др.). Обеспечение доверия к таким системам является долгосрочным вызовом, активным поиском ответа на который мировое сообщество занимается уже несколько лет. Однако на текущий момент не существует научно-технологической базы для разработки высоконадежных доверенных и одновременно эффективных систем, использующих технологии искусственного интеллекта (интеллектуальных систем), в том числе отсутствуют инструменты для поиска новых видов уязвимостей и противодействия новым типам угроз, специфичным для этих технологий.

Работа над ответом на этот вызов ведется в двух встречных направлениях. Первое направление – выработка требований к интеллектуальным системам и разработка государственных стандартов [1, 2].

Второе направление – создание научно-технологической базы, поддерживающей разрабатываемые стандарты. Без создания инструментальных средств, обеспечивающих безопасность функционирования интеллектуальных систем, невозможно говорить о полноценной реализации раз-

[1] *Институт системного программирования им. В.П. Иванникова Российской академии наук, Москва, Россия*
[2] *Сколковский институт науки и технологий, Москва, Россия*
[3] *Институт искусственного интеллекта AIRI, Москва, Россия*
[4] *Московский государственный университет имени М.В. Ломоносова, Москва, Россия*
[5] *Московский физико-технический институт, Долгопрудный, Россия*
[6] *Национальный исследовательский университет “Высшая школа экономики”, Москва, Россия*
[7] *Институт системного анализа Федерального исследовательского центра “Информатика и управление” Российской академии наук, Москва, Россия*
[8] *Институт математики и механики Уральского отделения Российской академии наук, Екатеринбург, Россия*
*****E-mail: turdakov@ispras.ru*

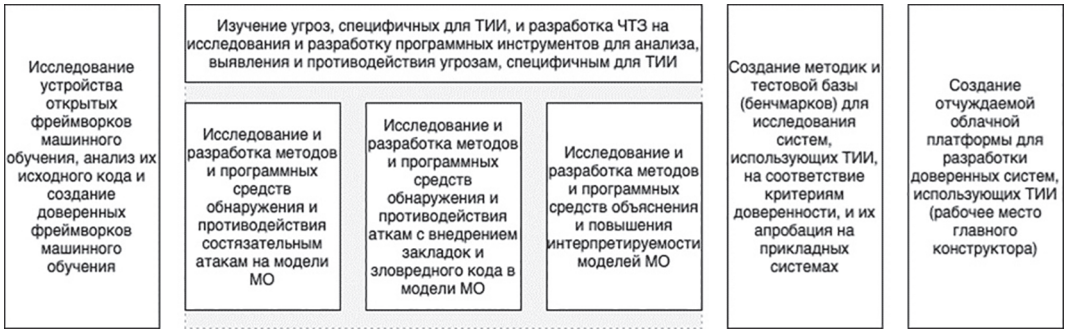

Рис. 1. Направления работ Центра.

работанных и перспективных стандартов. Работы в этом направлении ведутся исследовательским центром доверенного искусственного интеллекта ИСП РАН.

Программа Центра определяет следующие ключевые направления работ (рис. 1)

- Исследование устройства открытых фреймворков машинного обучения, анализ их исходного кода и создание доверенных фреймворков машинного обучения;
- Исследование и разработка программных инструментов для анализа, обнаружения и противодействия угрозам, специфичным для технологий искусственного интеллекта, включающее три раздела
- Исследование и разработка методов и программных средств обнаружения и противодействия состязательным атакам на модели машинного обучения;
- Исследование и разработка методов и программных средств обнаружения и противодействия атакам с внедрением закладок и зловредного кода в модели машинного обучения;
- Исследование и разработка методов и программных средств объяснения и повышения интерпретируемости моделей машинного обучения;
- Создание методик и тестовой базы (“бенчмарков”) для исследования интеллектуальных систем на соответствие критериям доверенности, и их апробация на прикладных системах;
- Создание отчуждаемой облачной платформы для разработки доверенных интеллектуальных систем.

В следующих разделах будут рассмотрены результаты текущих исследований и разработки по каждому направлению.

2. ДОВЕРЕННЫЕ ФРЕЙМВОРКИ МАШИННОГО ОБУЧЕНИЯ

Для продуктивной разработки интеллектуальных систем используется большое количество библиотек и фреймворков машинного обучения. Эти программные инструменты существенно ускоряют создание прикладных продуктов. Но как любое программное обеспечение они могут содержать ошибки и недокументированные возможности на уровне своего исходного кода или своих программных зависимостей. Такие уязвимости могут быть использованы злоумышленником для проведения атак на целевую систему. Таким образом, обеспечение доверия к интеллектуальным системам в целом невозможно без обеспечения доверия к этим ключевым системным компонентам.

Создание доверенного промышленного программного обеспечения регламентируется циклом безопасной разработки (SDL). Важным аспектом применения практик SDL является анализ не только разработанных компонент, но и заимствованного открытого программного обеспечения. Основными методами анализа, которые применяются в SDL, являются статический и динамический анализ (фаззинг). С помощью этих технологий возможно выявить критические дефекты в ПО и затем устранить их до выхода новой версии продукта.

В рамках работы Центра поставлена задача создания доверенных версий популярных фреймворков Tensorflow и PyTorch. Для этих фреймворков был проведен анализ поверхности атаки. Перспективным вектором атаки был выбран компонент загрузки моделей, так как обученные модели часто могут браться из недоверенного источника. Кроме того для фреймворка TensorFlow были восстановлены фаззинг-цели в проекте OSS-Fuzz, обеспечивающем непрерывный фаззинг для программного обеспечения с открытым исходным кодом [3].

Фазинг фреймворков производился с помощью инструмента Sydr [4]. В результате фаззинга было обнаружено 7 ошибок для фреймворка PyTorch [5, 6] и одна ошибка для фреймворка TensorFlow [7]. Также исходный код фреймворков был проанализирован инструментом статическо-

го анализа Svace [8]. В результате в фреймворке PyTorch было обнаружено 13 ошибок [9]. В фреймворке TensorFlow было обнаружено 8 ошибок [10].

Информация о найденных ошибках и предложения по исправлению некоторых из них были доведены до сообщества разработчиков. В целях дальнейшего использования все исправления были собраны в доверенных версиях фреймворков, созданных на базе PyTorch v1.11.0 и TensorFlow v2.8.2.

При этом сами фреймворки постоянно развиваются и появляются новые версии. Поэтому необходимы постоянная проверка новых изменений кода и синхронизация доверенных версий с оригинальными репозиториями. Для обеспечения этого непрерывного процесса создана аппаратно-программная инфраструктура обеспечения доверия к базовым фреймворкам машинного обучения, объединяющая инструменты SDL и инструменты для автоматизации их применения.

3. УГРОЗЫ, СПЕЦИФИЧНЫЕ ДЛЯ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ, И МЕТОДЫ ПРОТИВОДЕЙСТВИЯ

Одним из центральных вопросов разработки систем доверенного интеллекта является проблема неустойчивости отображений, выучиваемых моделями машинного обучения, к изменениям входных данных. Даже незначительное изменение входного объекта, например, добавление невидимого глазу шума к картинке, может существенно изменить предсказание модели на новом объекте, который как внешне, так и по метрикам схожести практически не отличается от исходного объекта. Этот феномен привел к возникновению так называемых состязательных атак (adversarial attack). Актуальность методов противодействия состязательным атакам на модели машинного обучения подтверждена экспериментами, проведенными сотрудниками Центра на моделях и наборах данных его научных партнеров (Университет Иннополис, ННГУ им. Н.И. Лобачевского). Проведена успешная атака белого ящика на систему сегментации рентгеновских снимков и на модели для предсказания возраста человека по экспрессии генов, основанные на бустинге решающих деревьев.

Подходы к построению моделей, устойчивых к атакам, делятся на 2 класса. В первом подходе модифицируется (сглаживается) сама модель, что приводит к более высокой эмпирической устойчивости. Возможно получить теоретические гарантии на величину атаки, при которой модель не будет менять свои предсказания. В рамках работы центра доверенного ИИ исследователями из Сколтеха впервые предложен [11] универсальный вероятностный подход к созданию моделей с сертифицированной устойчивостью, основанный на границах Чернова–Крамера. Подход позволяет формально оценить вероятность отказа модели, если атака выбрана из определенного распределения. Теоретические выводы подтверждены экспериментальными результатами на различных наборах данных.

Во втором подходе изменяется способ обучения. Сотрудниками Центра проведены исследования по этому направлению, в части задач оптимизации в условиях (злонамеренных) помех. Проведено исследование black-box моделей оптимизации, в которых атаки моделируются небольшим шумом к выдаваемому значению целевой функции [12]. Предложена общая схема, позволяющая сводить выпуклую задачу безградиентной оптимизации к выпуклой гладкой задаче оптимизации со стохастическим градиентным оракулом. Впервые удалось показать, что оптимальный метод с точки зрения числа оракульных вызовов (числа вычислений целевой функции), оказывается оптимальным и с точки зрения числа последовательных итераций, но, самое главное, что чувствительность такого метода к неточности в значении функции у него доказуемо наилучшая. Таким образом, был предложен подход, который одновременно оптимален по всем трем критериям (число оракульных вызовов, максимальный параллелизм и максимальный допустимый уровень шума). Кроме того удалось решить проблемы, связанные с формализацией атак на параметры задачи оптимизации таким образом, что исходная оптимизационная задача, в конечном итоге, заменяется седловой задачей. Предложены подходы к решению таких оптимизационных задач с доказанной оптимальностью [13, 14].

Кроме того, исследователями Центра протестирован прикладной подход противодействия состязательным атакам – состязательное обучение (adversarial training) моделей, т.е. обучение на атакованных некоторым методом атаки данных, а также методы аугментации данных, направленные на защиту от атак [15]. А также разработаны новые нейросетевые архитектуры, более устойчивые к состязательным атакам с общепринятыми видами возмущений [16]

4. ПОВЫШЕНИЕ ИНТЕРПРЕТИРУЕМОСТИ МОДЕЛЕЙ

Нейронные сети широко используются в качестве мощных инструментов моделирования, и большинство крупных поставщиков включили их в свое программное обеспечение для интеллектуального анализа данных. Моделирование, однако, является лишь частью процесса интеллектуального анализа данных. Дополнительно необходимо проанализировать влияние входных

переменных на результат модели. Результаты некачественной интерпретации применяемой модели могут быть использованы злоумышленником для проведения атак на целевую систему, например, через “отравление” данных. Таким образом, обеспечение доверия к интеллектуальным системам требует обеспечения интерпретируемости применяемых моделей.

В рамках работы Центра поставлены задачи повышения интерпретируемости и выявления некорректных результатов работы многослойных нейронных сетей. Разработан тестовый стенд для визуализации полносвязной нейронной сети с сублинейными функциями активации для задачи бинарной классификации в двумерном пространстве признаков. С его помощью проанализирован способ геометрической интерпретации обученной нейросетевой модели, где каждому нейрону соответствует разделяющая гиперплоскость в исходном признаковом пространстве. На основе нее предложен ряд новых методов анализа модели. Для входного примера в качестве интерпретации решения сети предъявляется k ближайших соседей с точки зрения модели из обучающей выборки. Предложен метод построения решающего дерева, имитирующего работу исходной нейросети, листовые вершины которого соответствуют упомянутым секторам или их объединениям. Статистический анализ в них может показать, например, стоит ли доверять решению сети в этой области, или что невозможно однозначно принять решение. Предложенные подходы могут быть применимы для полносвязных нейросетей с произвольным числом нейронов и слоев. Для сверточных нейросетей предложен метод восстановления интерполированных изображений из точек признакового пространства. Также разрабатываются методы анализа фильтров, значимых для определения заданного класса изображений.

Созданы [17] вычислительно-эффективные методы выявления некорректных результатов работы многослойных нейронных сетей с архитектурой “Трансформер”. Первый метод заключается в оценке принадлежности анализируемых объектов к определенным классам с помощью ансамбля нейронных сетей, в каждой из которых маскируются отдельные веса выходного слоя. Второй метод – оценка расстояния между скрытыми векторными представлениями (эмбеддингами) анализируемых объектов и ближайших к ним объектов из обучающей выборки. Результаты экспериментов на задачах классификации текстов и извлечения именованных сущностей показали, что применение предложенных методов позволяет существенно повысить надежность обнаружения ошибок классификации по сравнению с более вычислительно затратными методами.

Кроме того, исследован подход получения семантически интерпретируемых категорий для векторных представлений на основе семантических сетей типа WordNet. В качестве интерпретируемых размерностей используются семантические классы (суперпонятия), объединяющие множества слов.

5. СОЗДАНИЕ МЕТОДИК И ТЕСТОВОЙ БАЗЫ (БЕНЧМАРКОВ) ДЛЯ ОЦЕНКИ ДОВЕРИЯ К ПРИКЛАДНЫМ СИСТЕМАМ

Наличие или отсутствие злоумышленника, сценарии воздействия злоумышленника на процессы жизненного цикла прикладных интеллектуальных систем, разнообразие самих прикладных задач и интеллектуальных систем требуют тщательного анализа и систематизации этих сценариев и угроз. В результате проведения такого анализа были разработаны критерии доверия к интеллектуальным системам, которые в дальнейшем будут апробированы на реальных интеллектуальных системах.

При этом доверие к прикладным интеллектуальным системам не может рассматриваться в отрыве от эффективности таких систем. Так, тривиальной является разработка модели машинного обучения, неуязвимой, например, к состязательным атакам, если не задавать минимальные требования к точности и скорости распознавания для этой модели. Таким образом, важной задачей является создание тестовой базы из наборов данных и моделей на основе реальных задач из различных прикладных областей, и разработка методик для оценки соответствия интеллектуальных систем и их компонентов требованиям в области доверия и эффективности.

Были разработаны бенчмарки, решающие различные прикладные задачи, используя данные разного типа: текст, изображения, видео, графы, таблицы. При этом в качестве бенчмарков мы используем как простые (baseline) модели, так и создаем state-of-the-art решения, на которых можно демонстрировать эффективность методов обеспечения доверия в реальных условиях.

В частности, разработана новая процедура создания искусственного движения для обучения более устойчивых нейросетевых алгоритмов обработки видео. Предложенная процедура была использована для обучения нейросетевого алгоритма семантического матирования видео с людьми [18]. Акцент в данном алгоритме был сделан на стабильности результата работы во времени. При помощи предложенной процедуры удалось многократно повысить размер обучающей выборки, а также повысить устойчивость метода матирования к разным видам движения.

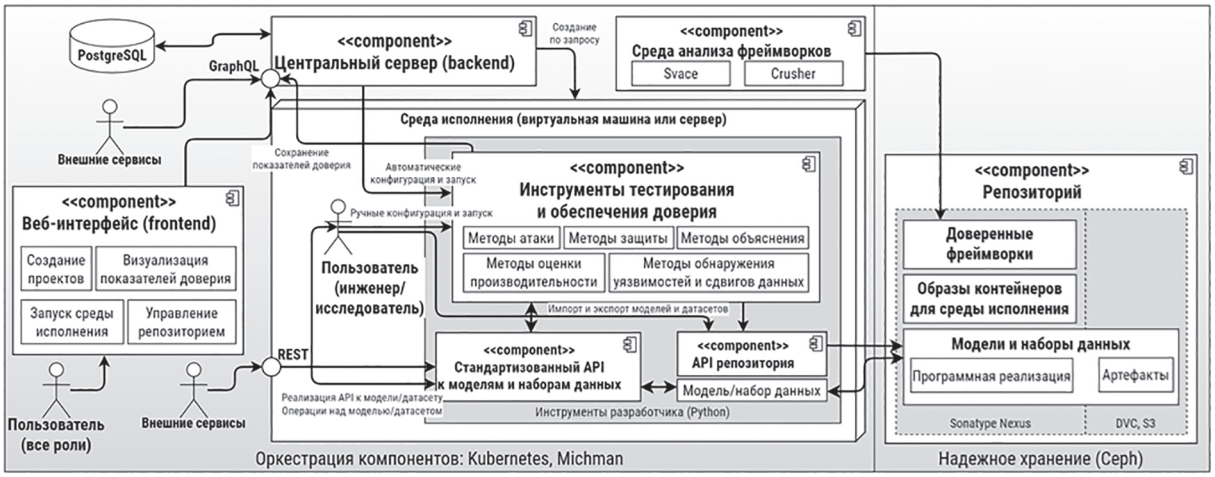

Рис. 2. Высокоуровневая архитектура Платформы.

Также разработан новый однопроходный метод для открытого извлечения информации из текстов, вдохновленный алгоритмами детектирования объектов из области компьютерного зрения [19]. Подход использует независимую от порядка следования отдельных слов функцию потерь, основанную на двудольном сопоставлении слов и отдельных элементов триплетов, которое обеспечивает однозначные предсказания модели, и архитектуру, использующую только кодировщик на основе “Трансформер” для маркировки последовательностей. Предлагаемый подход демонстрирует превосходящую или аналогичную производительность с точки зрения как показателей качества, так и времени вывода по сравнению с современными моделями на стандартных бенчмарках.

6. ОБЛАЧНАЯ ПЛАТФОРМА ДЛЯ РАЗРАБОТКИ ДОВЕРЕННЫХ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ

Основываясь на результатах проведенных исследований была разработана концепция облачной платформы для разработки интеллектуальных систем. Платформа будет

• содержать инструменты анализа наборов данных (датасетов) и анализа моделей машинного обучения, в том числе объяснения моделей, исследования их устойчивости к атакам и производительности;

• накапливать доверенные модели, алгоритмы и наборы данных для обучения моделей и проведения экспериментов (бенчмарки);

• предоставлять компоненты жизненного цикла разработки безопасного программного обеспечения.

Современные методы машинного обучения требуют большого количества вычислительных ресурсов в момент обучения модели. В остальное время эти вычислительные ресурсы могут использоваться неэффективно. Модель облачных вычислений позволяет решить эту проблему за счет обеспечения доступа к вычислительным ресурсам по требованию. Отсюда вытекает требование возможности развертывания Платформы в публичных и приватных облаках. Работа Платформы в связке с облаком будет продемонстрирована на примере облачной платформы Asperitas и оркестратора Michman [20], обеспечивающего развертывание необходимых программных компонентов по запросу.

Еще одно ключевое требование к платформе – расширяемость. Так как новые программные инструменты анализа и реализации атак и защиты появляются постоянно, будет обеспечена возможность их добавления в Платформу. В частности, разрабатываются программные интерфейсы (рис. 2) и методика добавления таких инструментальных средств.

7. ЗАКЛЮЧЕНИЕ

Исследовательский центр доверенного искусственного интеллекта поставил перед собой задачу разработки научно-технической базы, которая позволит создавать доверенные интеллектуальные системы. С этой целью сотрудниками и партнерами Центра проводились работы по следующим направлениям: создание доверенных фреймворков машинного обучения; исследование угроз, специфичных для искусственного интеллекта и методов противодействия им; повышение интерпретируемости моделей машинного обучения; создание методик и бенчмарков для оценки доверия; а также объединение прикладных инстру-

ментов в облачную платформы для разработки доверенных интеллектуальных систем. Научные исследования опубликованы в 9 статьях, представленных на конференциях A*, и 7 журнальных статьях из первого квартиля (Q1). Основным прикладным результатом в 2022 г. являются созданные сотрудниками Центра доверенные версии фреймворков TensorFlow и PyTorch, уже переданные в опытную эксплуатацию индустриальным партнерам Центра.

СПИСОК ЛИТЕРАТУРЫ

1. ГОСТ Р 59276–2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения.
2. ГОСТ Р 59921–2021. Системы искусственного интеллекта в клинической медицине.
3. Pull request на восстановление фаззинг целей для фреймоврка TensorFlow. https://github.com/google/oss-fuzz/pull/7704. Дата обращения: 2022-10-26.
4. V*ishnyakov A., Fedotov A., Kuts D., Novikov A., Parygina D., Kobrin E., Logunova V., Belecky P., Kurmangaleev S.* Sydr: Cutting edge dynamic symbolic execution. In 2020 Ivannikov ISPRAS Open Conference (ISPRAS) (pp. 46–54). IEEE. 2020 December.
5. Pull request в PyTorch [https://github.com/pytorch/pytorch/pull/79192]. Дата обращения: 2022-10-26.
6. Pull request в PyTorch [https://github.com/pytorch/pytorch/pull/84343]. Дата обращения: 2022-10-26.
7. Pull request, Fix endless loop in TF. [https://github.com/tensorflow/tensorflow/pull/]. Дата обращения: 2022-10-26.
8. *Ivannikov V.P., Belevantsev A.A., Borodin A.E., Ignatiev V.N., Zhurikhin D.M., Avetisyan A.I.* Static analyzer Svace for finding defects in a source program code. Programming and Computer Software. 2014. V. 40 (5). P. 265–275.
9. [https://github.com/pytorch/pytorch/pull/85705]. Дата обращения: 2022-10-26.
10. [https://github.com/tensorflow/tensorflow/pull/57892]. Дата обращения: 2022-10-26.
11. *Pautov M., Tursynbek N., Munkhoeva M., Muravev N., Petiushko A., Oseledets I.* (2022, June). CC-Cert: A probabilistic approach to certify general robustness of neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 7, pp. 7975–7983).
12. *Gasnikov A., Novitskii A., Novitskii V., Abdukhakimov F., Kamzolov D., Beznosikov A., Takáč M., Dvurechensky P., Gu B.* The power of first-order smooth optimization for black-box non-smooth problems. ICML 2022.
13. *Kovalev D., Gasnikov A., Richtárik P.* Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. NeurIPS 2022.
14. *Kovalev D., Gasnikov A.* The First Optimal Algorithm for Smooth and Strongly-Convex-Strongly-Concave Minimax Optimization. NeurIPS 2022.
15. *Chistyakova A., Cherepnina M., Arkhipenko K., Kuznetsov S.D., Oh C.S., Park S.* September. Evaluation of interpretability methods for adversarial robustness on real-world datasets. In 2021 Ivannikov Memorial Workshop (IVMEM) (pp. 6-10). IEEE. 2021.
16. *Курденкова Е.О., Черепнина М.С., Чистякова А.С., Архипенко К.В.* Влияние трансформаций на успешность состязательных атак для классификаторов изображений Clipped BagNet и ResNet. Иванниковские чтения, 2022.
17. *Vazhentsev A., Kuzmin G., Shelmanov A., Tsvigun A., Tsymbalov E., Fedyanin K., Zhukov L.* Uncertainty Estimation of Transformer Predictions for Misclassification Detection //Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022. С. 8237–8252.
18. *Molodetskikh I., Erofeev M., Moskalenko A., Vatolin D.* Temporally coherent person matting trained on fake-motion dataset. Digital Signal Processing. 2022. V. 126. P. 103521.
19. *Vasilkovsky M., Alekseev A., Malykh V., Shenbin I., Tutubalina E., Salikhov D., Stepnov M., Chertok A., Nikolenko S.* DetIE: Multilingual Open Information Extraction Inspired by Object Detection. In Proceedings of the 36th AAAI Conference on Artificial Intelligence. 2022.
20. *Aksenova E., Lazarev N., Badalyan D., Borisenko O. and Pastukhov R.* December. Michman: an Orchestrator to deploy distributed services in cloud environments. In 2020 Ivannikov Ispras Open Conference (ISPRAS) (pp. 57–63). IEEE. 2020.

EDN: DBRSQV
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 19–27

РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ИССЛЕДОВАТЕЛЬСКИХ ЦЕНТРОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

УДК 004.8

ФУНДАМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ И РАЗРАБОТКИ В ОБЛАСТИ ПРИКЛАДНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

© 2022 г. Е. В. Бурнаев[1,2,*], А. В. Бернштейн[1], В. В. Вановский[1], А. А. Зайцев[1], А. М. Булкин[1], В. Ю. Игнатьев[1], Д. Г. Шадрин[1], С. В. Илларионова[1], И. В. Оселедец[1,2], А. Ю. Михалев[1], А. А. Осипцов[1], А. А. Артемов[1], М. Г. Шараев[1], И. Е. Трофимов[1]

Представлено академиком РАН А.П. Кулешовым
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

Настоящий этап развития искусственного интеллекта (ИИ) характеризуется развитием технологий, методов и алгоритмов машинного обучения (МО), в том числе глубокого машинного обучения, интеллектуального анализа данных и других фундаментальных научных направлений и созданием на их основе прикладных решений практически во всех сферах цифровой экономики. Однако расширение сферы приложений ИИ, усложнение класса решаемых задач и спектра и объема данных, используемых для создания прикладных ИИ-моделей и интеллектуальных систем на базе ИИ, потребовали существенного расширения теоретической и алгоритмической базы ИИ, включая необходимость развития методов МО с использованием математических и физических моделей объектов и явлений предметных областей, методов консолидации мультимодальных данных, методов создания геометрических и топологических компонентов нейронных глубоких сетей, методов моделирования изучаемых 3D-объектов и др. Для ответа на эти вызовы в 2021 г. в рамках федерального проекта “Искусственный интеллект” на базе Сколтеха был создан Исследовательский центр прикладного искусственного интеллекта, задачами которого является создание научно-технологической базы для решения широкого спектра актуальных прикладных задач для целей устойчивого развития экономики РФ, включая задачи оптимизации управленческих решений в целях снижения углеродного следа и другие актуальные задачи направления ESG; задачи мониторинга окружающей среды с целью выявления аномалий и прогнозирования развития экстремальных ситуаций; оценка экономических и социальных рисков и их динамики, вызванных климатическими изменениями; задачи предиктивной аналитики и др. В статье описаны развиваемые в Центре новые технологии, модели, методы и алгоритмы ИИ, основные прикладные направления исследований Центра и уже достигнутые научные и прикладные результаты.

Ключевые слова: прикладной искусственный интеллект, устойчивое развитие, глубокое обучение, машинное обучение, физически информированные нейронные сети, анализ данных

DOI: 10.31857/S2686954322070049

1. ВВЕДЕНИЕ

Исследовательский Центр прикладного искусственного интеллекта был создан в 2021 г. в рамках федерального проекта “Искусственный интеллект” национальной программы “Цифровая экономика Российской Федерации” на базе нескольких научных групп Сколковского института науки и технологии.

Целью Центра являются проведение ориентированных фундаментальных и прикладных исследований в области искусственного интеллекта и использование полученных результатов для поддержки управленческих решений на базе разрабатываемых инструментов для мультимасштабного мониторинга и управления климатическими и экологическими рисками для реализации Национальной стратегии развития искусственного интеллекта (далее ИИ) на период до 2030 г. и Энергетической стратегии РФ до 2035 г.

Основные направления фундаментальных и прикладных исследований Центра в области ИИ включают в себя:

- развитие теоретических и прикладных методов машинного обучения, таких как методы
- создания глубоких нейронных сетей,

[1] *Сколковский институт науки и технологий, Москва, Россия*
[2] *Научно-исследовательский институт искусственного интеллекта, Москва, Россия*
*****E-mail: e.burnaev@skoltech.ru*

• создания крупномасштабных генеративных моделей [1, 3],

• создания интерпретируемых и устойчивых физически информированных нейронных сетей с использованием знаний и моделей предметных областей (Physics Informed ML),

• машинного обучения для решения обратных задач, в том числе для обнаружения ошибок входных данных и их корректировки,

• исключения систематических ошибок моделирования, связанных с неполным соответствием реальных явлений и процессов их аналитическим представлениям, и минимизации ошибок прогноза,

• адаптации предиктивных моделей под изменяющиеся со временем границы пространства условий и входных данных и др.;

• развитие теоретических и прикладных методов повышения вычислительной эффективности больших нейросетевых моделей (быстрое обучение, сжатие моделей, поиск новых архитектур) для снижения вычислительной нагрузки на симуляцию на основе математических моделей процессов, включая

• алгоритмические методы экономии памяти (использование attention-слоев, вычисление неточных градиентов линейных слоев и нелинейных слоев активаций, малопараметрическое представление линейных слоев при помощи тензоров в TT-формате и др.), позволяющие уменьшить нагрузку на память (с возможностью одновременной работы с большим количеством входных данных) и снизить итоговое время обучения,

• новые стохастические методы оптимизации, основанные на тонкой настройке внутренних гиперпараметров оптимизатора для каждой нейросетевой модели,

• параллельные алгоритмы на основе task-based парадигмы программирования, основанной на описании всех вычислений в виде направленного графа без циклов. Этот граф вычислений используется для распределения вычислений и асинхронной пересылки данных для роста общей “утилизации” всех устройств;

• развитие теоретических и прикладных методов интеллектуального анализа данных, таких как

• дифференциально-геометрические, топологические, графовые и стохастические методы анализа данных [2],

• методы обработки и анализа сигналов и изображений,

• методы консолидации мультимодальных данных различной физической природы и извлечения из них релевантной информации,

• методы 3D компьютерного зрения [18, 20],

• методы обнаружения и идентификации аномалий в данных и разладок в динамических системах и др.;

• развитие прикладных технологий построения моделей предиктивной аналитики и создания систем поддержки принятия решений, учитывающих знания и модели предметной области (включая особенности носителя обрабатываемых многомерных данных), их стохастического характера, дизайна процессов получения выборки эмпирических данных и степень их неопределенности.

На базе разрабатываемых методов и алгоритмов создается программный инструментарий ИИ в виде программных средств и платформенных решений (универсальных программных фреймворков и библиотек) для решения с их помощью прикладных задач развития научно-технологического комплекса РФ и устойчивого развития российской промышленности и экономики.

Программный инструментарий ИИ, развиваемый в Центре, используется для разработки ряда прикладных пилотных проектов для повышения эффективности и формирования новых направлений деятельности индустриальных партнеров Центра и ключевых отраслевых предприятий экономики РФ (в соответствии с Программой Центра), включая

• формирование нового направления финансового мониторинга и учета ESG рисков при кредитовании промышленных предприятий (для Сбера),

• повышение эффективности анализа корпоративной информации за счет применения методов ускорения обучения и сжатия больших нейросетевых моделей (для Сбера),

• оптимизацию управленческих решений на повышение нефтеотдачи и снижения экологического ущерба; разработку самообучающейся модели нефтегазоносного пласта (для Газпромнефти),

• повышение эффективности, надежности и масштабирование системы мониторинга качества атмосферного воздуха (для СитиЭйр),

а также для разработки актуальных прикладных решений и сервисов для других объектов реальной экономики и социальной сферы, включая ФОИВы, министерства и др.

В следующих разделах будут более подробно описаны нескольких текущих проектов (в рамках перечисленных выше направлений):

• прогнозирование экономических последствий от наступления физических и климатических рисков (для Сбера),

• снижение ресурсоемкости обучения больших нейросетевых моделей (для Сбера),

• прогнозирование ледовой обстановки в Арктике (для Газпромнефти),

• самообучающаяся модель пласта (для Газпромнефти), иллюстрирующих актуальность и востребованность исследований, необходимость разработки и использования технологий ИИ, а также достигнутые промежуточные результаты.

2. АНАЛИЗ ФИЗИЧЕСКИХ И ФИНАНСОВЫХ РИСКОВ, СОЗДАВАЕМЫХ КЛИМАТИЧЕСКИМИ ИЗМЕНЕНИЯМИ

2.1. Предпосылки проекта

Территория Российской Федерации находится в зоне высокого риска стихийных бедствий, таких как паводки, штормы, засухи, лесные пожары и др., а также рисков, связанных с изменениями климата, такими как таяние мерзлоты в Арктической зоне Российской Федерации, и др. [4, 5]. Таким образом возникает задача расчета стоимости ESG рисков (по E-компоненте) и их влияния на принятие управленческих решений в промышленной, социальной и финансовой областях (например, при кредитовании промышленных предприятий). Поэтому необходима разработка технологий расчета вероятности реализации различных экстремальных событий в конкретных областях РФ, прогнозирования релевантных последствий от наступления экстремальных событий и изменения климата на различных горизонтах планирования.

Использование новых методов машинного обучения и анализа данных для решения перечисленных выше задач определяется:

• необходимостью обработки и анализа больших массивов взаимосвязанных климатических данных, в связи с тем, что методы машинного обучения значительно превосходят по вычислительной эффективности и точности стандартные методы многомерного статистического анализа, упрощая процессы, связанные с построением моделей экстремальных климатических событий и использованием климатических сценариев;

• наличием неопределенности в климатических проекциях моделей экстремальных рисков при различных климатических сценариях;

• наличием моделей, основанных на первых принципах (напр., физические уравнения теплопроводности по Кудрявцеву [7]), с использованием которых возможно строить физически информированные модели машинного обучения. Например, модели, обученные на данных сети глобального мониторинга криолитозоны (данные GTNP) с использованием современных методов искусственного интеллекта (нейронные обыкновенные дифференциальные уравнения, модели на основе гауссовских процессов, градиентного бустинга и др.), уточняют прогнозы моделей на основе физических процессов и ускоряют расчет полученных прогнозов.

2.2. Результаты

В рамках проекта поставлены и решены следующие научные задачи:

• построены различные прогнозы (протаивания грунта, влияния климата на условия землепользования и др.), а также построены модели глубины протаивания и температуры грунта,

• вычислены вероятности реализации различных экстремальных событий в конкретных точках Российской Федерации (в масштабе 27 × 27 км), такие как сильный ветер (более 20 м/с), наводнения (разливы рек), град (вероятность возникновения конвективных явлений), засухи (индекс Палмера [6]). Например, на рис. 1 показан предсказанный уровень наводнений в зависимости от климатического сценария (сценария выбросов),

• построена модель кредитного риска с поправкой на климатические риски, в которой использование модели волатильности определенного типа позволило существенно реже калибровать модель.

3. СНИЖЕНИЕ РЕСУРСОЕМКОСТИ ОБУЧЕНИЯ БОЛЬШИХ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ

3.1. Предпосылки проекта

Основной задачей проекта является ускорение обучения больших нейросетевых моделей [8]. Уменьшение времени обучения фактически экономит денежные средства и сокращает углеродный след от вычислений в одинаковых пропорциях. Например, полное обучение всемирно известной GPT-3 модели со 175 миллиардами параметров потребовало бы примерно 3640 Петафлоп/с-дней при стопроцентной утилизации вычислительных устройств. Это эквивалентно 8650 мес вычислений на одной видеокарте Nvidia V100. Экономия даже 10 процентов времени обучения имеет большое экономическое и экологическое значение. В данный момент по проекту ведутся работы в трех основных направлениях: экономия памяти, стохастические методы оптимизации и параллельные алгоритмы на основе т.н. task-based парадигмы программирования.

Экономия памяти при обучении моделей экономит итоговое время обучения, но эта зависимость не очевидна. Уменьшение нагрузки на память открывает возможность одновременной работы с большим количеством входных данных. Чем больше данных обрабатывается одновременно, тем точнее вычисляется градиент по параметрам модели. Таким образом, параметры модели примут свои окончательные значения за меньшее

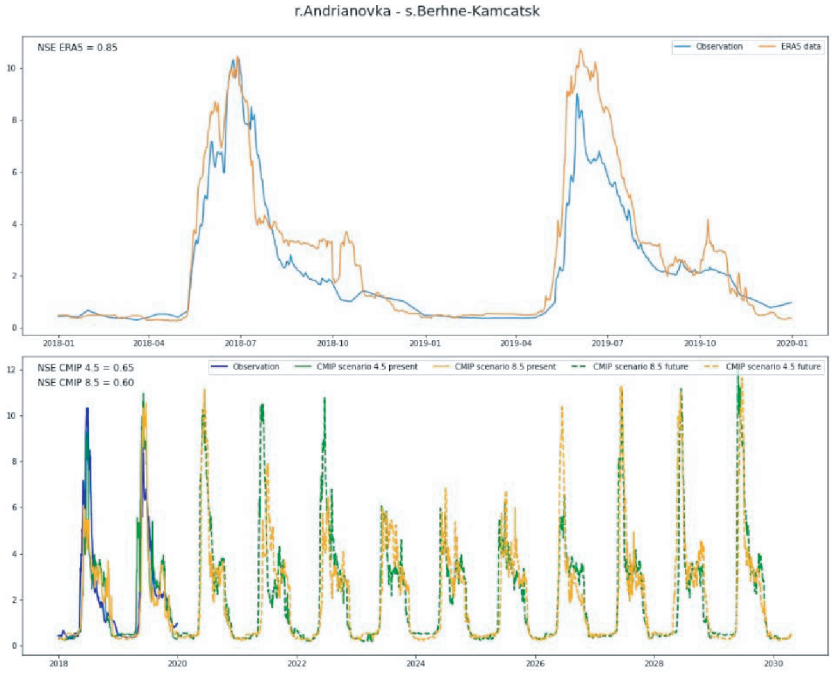

Рис. 1. Предсказанный уровень наводнений в зависимости от климатического сценария (сценария выбросов парниковых газов).

количество эпох обучения и итоговое время снизится.

Без параллельного обучения невозможно обучить хоть какую-то действительно большую нейронную сеть. В настоящее время большинство, если не все, методы параллелизации основаны на т.н. Bulk synchronous парадигме программирования. Этот подход подразумевает чередование параллельных вычислений и обмена данными. В рамках проекта ведется разработка параллельных алгоритмов на основе т.н. task-based парадигмы, основанной на описании всех вычислений в виде направленного графа без циклов. Этот граф вычислений передается специальной библиотеке, в нашем случае это StarPU, которая сама распределяет вычисления и организует пересылки данных асинхронно. За счет асинхронности вычислители не простаивают во время пересылки данных, и общая утилизация всех устройств растет. К сожалению, данный подход требует полного переписывания программного кода с нуля, что является времязатратным процессом.

3.2. Результаты

Инструменты, разработанные в ходе реализации проекта, позволяют экономить память при обучении нейросетевых моделей при помощи следующих алгоритмических улучшений: attention-слой, использующий в три раза меньше временных данных, вычисление неточных градиентов линейных слоев [9, 10] и нелинейных слоев активаций и малопараметрическое представление линейных слоев при помощи тензоров в TT-формате.

Разработан новый метод стохастической оптимизации. Безусловно, существует много различных методов оптимизации такого типа, однако, в рассматриваемом случае основной упор делается на тонкую настройку внутренних гиперпараметров оптимизатора для каждой нейросетевой модели, что позволяет получить конкурентноспособный продукт [11].

Созданные подходы уже применены для обучения больших нейросетевых моделей (ruDALLE, “Малевич”), получены снижение памяти, ускорение вычислений, затраченной энергии и углеродного следа на 15% [9, 12].

4. ПРОГНОЗИРОВАНИЕ ЛЕДОВОЙ ОБСТАНОВКИ В АРКТИКЕ

4.1. Предпосылки проекта

Глобальное потепление сделало Арктику доступной для морских операций и создало потребность в надежных оперативных прогнозах движения морского льда для обеспечения их безопасности. Для прогнозирования ледовой обстановки представляют интерес такие регионы, как Баренцево и Карское моря (1), море Лабрадор (2), море Лаптевых (3). Красной и зеленой изолиниями на рис. 2 ограничена зона маргинального льда (кон-

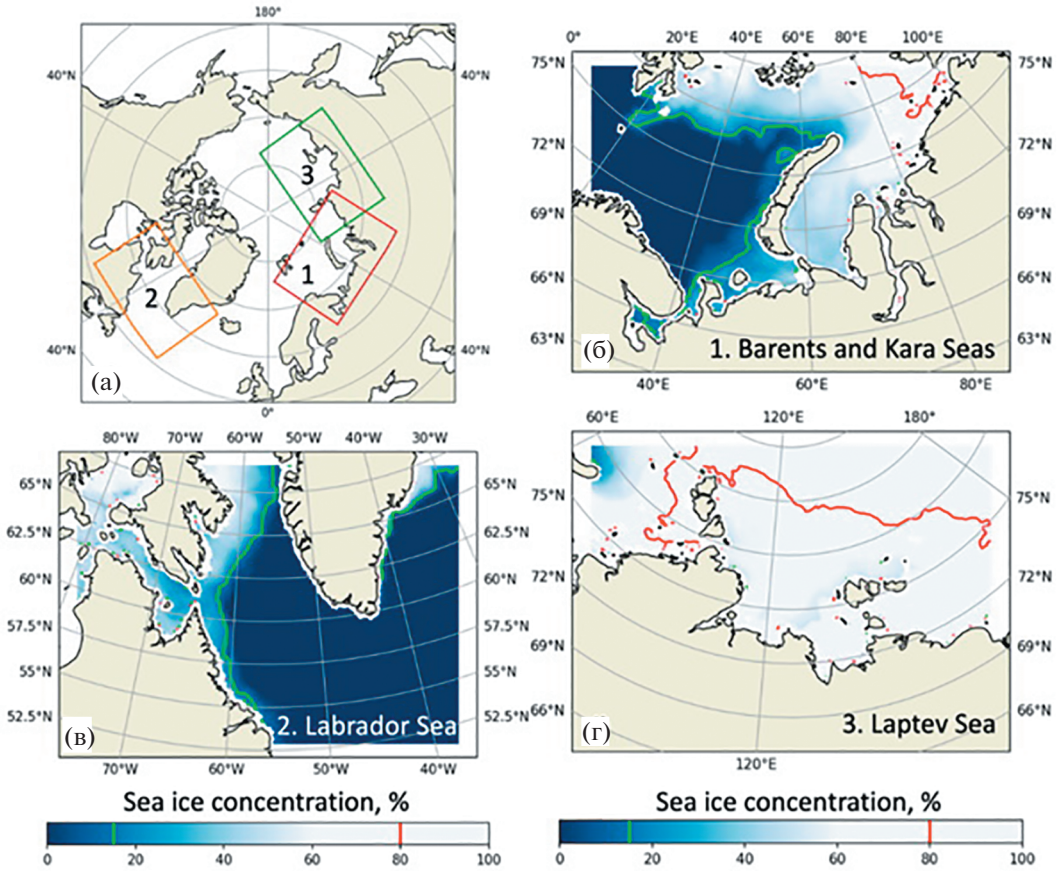

Рис. 2. Примеры карт концентрации морского льда.

центрация 15–80%), которая дает основной вклад в ошибку прогноза из-за высокой изменчивости ледового покрова.

В то время как численные модели океанского льда требуют больших вычислительных ресурсов, относительно легковесные методы на основе машинного обучения могут показывать себя более эффективно в этой задаче. Необходимость использования новых методов машинного обучения определяется тем, что лишь немногие из существующих исследований сосредоточены на разработке систем реального времени для построения ежедневных оперативных прогнозов, учитывающих доступные данные.

4.2. Результаты

В рамках проекта была усовершенствована технология обучения глубоких сетей с архитектурой U-Net (обучение проводилось в двух режимах), позволившая строить краткосрочные прогнозы морского льда на срок до 10 дней [13]. Было показано, что построенная модель глубокого обучения значительно превосходит простые бейзлайны, а использование дополнительных данных об оперативных прогнозах погоды позволяет дополнительно улучшить качества работы модели. Обучение модели на данных сразу нескольких регионов способствовало улучшению ее обобщающей способности при использовании в новых регионах. В результате получен быстрый и гибкий инструмент для оперативных прогнозов состояния морского льда в регионах Баренцева моря, Лабрадорского моря и моря Лаптевых. На рис. 3 показаны примеры прогнозов наилучшей конфигурации обученной модели U-Net. Представлены результаты для трех различных периодов прогноза (от одного до трех дней). Красно-синей шкалой отражены ошибки прогнозирования модели – прогноз избыточной и недостаточной концентрации морского льда соответственно.

5. САМООБУЧАЮЩАЯСЯ МОДЕЛЬ ПЛАСТА

5.1. Предпосылки проекта

Нефтегазовая отрасль остается и в XXI веке одной из самых наукоемких и цифровизированных и имеет хорошо развитую систему подходов и численных моделей для моделирования и управления нефтедобычей. Обилие методов и продук-

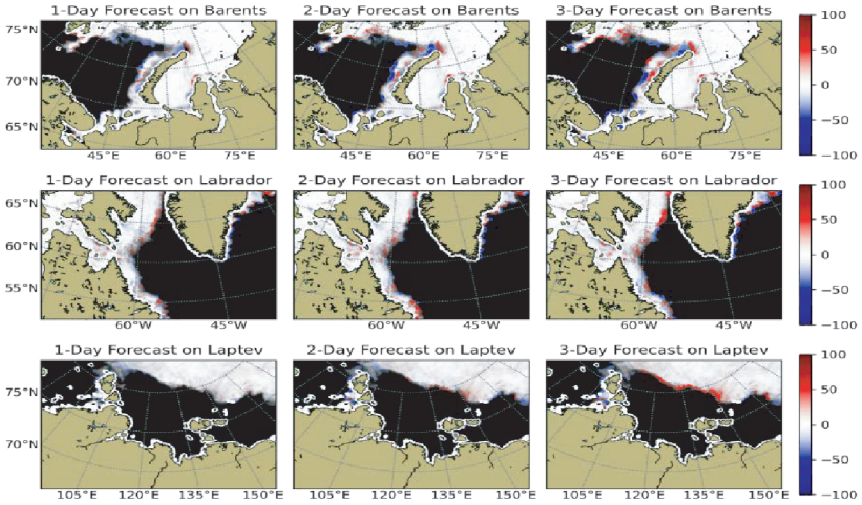

Рис. 3. Распределение ошибок прогнозирования модели.

тов для моделирования нефтегазовых пластов не решает главной проблемы – неопределенности и скудности имеющейся информации, вытекающей из несовершенства методов ее получения, а также вечного противоборства стратегий исследования неизученных областей и применения имеющихся знаний (exploration vs exploitation), что приводит к естественному желанию максимально эффективно использовать имеющуюся информацию. Неопределенности информации бывают разных типов, это, как и относительно понятные неопределенности измерений, как геологических, так и промысловых, численные погрешности гидродинамических симуляций, погрешности процесса адаптации геологогидродинамической модели, так и более сложные и концептуальные источники неопределенностей, как то: некорректность полуэмпирических зависимостей и формул, закладываемых в расчеты, отсутствие или несоответствие геологического реализма моделируемых подземных структур реальным.

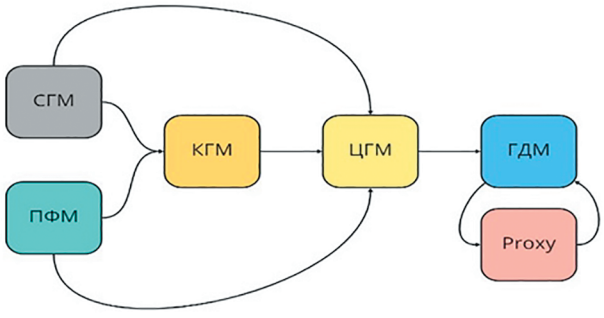

Рис. 4. Общая схема взаимодействия различных моделей пласта.

Общая схема взаимодействия различных взаимосвязанных между собой моделей пласта, описывающих различные аспекты функционирования пласта (Петрофизических моделей (ПФМ), Гидродинамических моделей (ГДМ), Сейсмогеологических моделей (СГМ), Концептуальных геологических моделей КГМ, Цифровых геологических моделей (ЦГМ)), при котором выходные данные одной модели зачастую являются входными для другой, изображена на рис. 4. Итоговая Гидродинамическая модель (ГДМ) объединяет все полученные знания о месторождении, адаптируется на реальные данные из месторождения (используя иногда прокси-модели типа IsobarProxy). Далее, происходит долгий ручной процесс адаптации модели на добычу, а весь процесс построения модели месторождения может занимать до полугода и требовать большого количества ресурсов.

Необходимость использования новых методов машинного обучения и анализа данных для построения модели месторождения определяется следующими обстоятельствами:

• данные измерений имеют очень разную локальность и степень надежности, их требуется объединять процедурой, которая будет настраиваться по целевым метрикам,

• данные сейсмических измерений не имеют однозначной интерпретации, процедура интерпретации в идеале должна быть настраиваемой на каждом месторождении,

• решение обратной задачи крайне неэффективно проводить в исходном пространстве кубов гидродинамических параметров месторождения размерностью порядка миллионов или даже миллиардов, поэтому требуются алгоритмы представ-

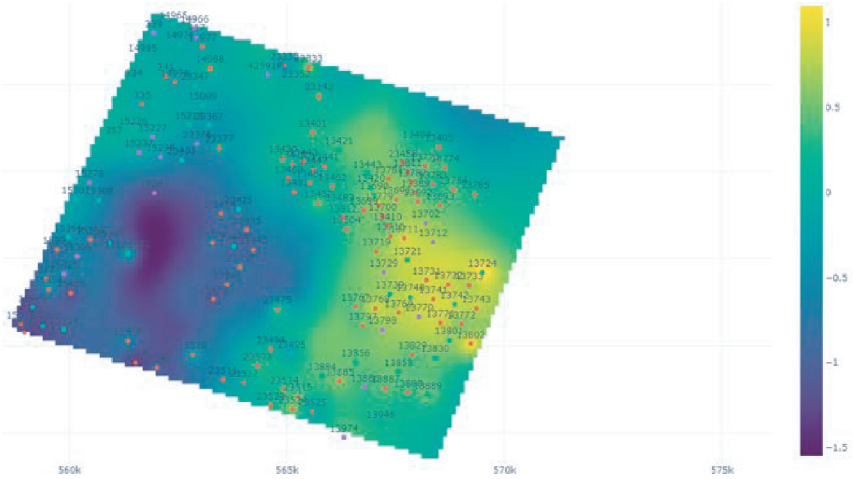

Рис. 5. Оценка карты проницаемости на основе комплексированных данных.

ления месторождения в пространстве параметров меньшей размерности,

• для ускорения процесса адаптации гидродинамической модели требуется значительно ускорить симуляции с возможной потерей точности,

• для адекватного учета рисков и определения конечных коридоров неопределенности по прогнозам добычи требуется проброс неопределенности из исходных данных измерений в конечные прогнозы.

Основной задачей комплексной исследовательской программы СМП (Самообучающаяся Модель Пласта) является одновременный учет с помощью технологий ИИ максимального числа имеющихся данных с их неопределенностями для построения цифровой модели месторождения и оценки будущих показателей добычи, извлекаемых запасов и других важных параметров с доверительными интервалами для принятия обоснованных управленческих решений и, в конечном итоге, повышении экономической эффективности разработки месторождения за счет увеличения количества извлекаемых углеводородов, уменьшения числа неоптимально пробуренных скважин и уменьшения трудозатрат специалистов для построения цифровой модели месторождения. Самообучающаяся модель пласта представляет собой иерархию моделей и методов, принимающую на вход разного типа исходные данные с их неопределенностями, алгоритмы расчета добычи по исходным данным, историю добычи, проводящую автоматическую адаптацию моделей на добычу, уточняющую оценку неопределенностей и выдающую прогнозы добычи с доверительными интервалами.

5.2. Результаты

В рамках проекта поставлены и решены следующие задачи:

Задача объединения данных о месторождении [16]. Стандартные алгоритмы объединения имеющихся на месторождении данных дают не слишком хорошие результаты в силу разной локальности и достоверности данных, а также наличия большого числа выбросов и пропусков в данных. В рамках проекта был разработан новый алгоритм комплексирования данных с помощью алгоритмов непараметрической регрессии с адаптивным ядром, которое настраивается по целевым метрикам для автоматического учета различного качества имеющихся данных. На рис. 5 изображены результаты такого комплексирования на одном из тестовых месторождений (карта проницаемости, полученная комплексированием данных ГИС, ГДИС и сейсмики на участке реального месторождения).

В метрике leave-one-out предложенный подход значительно превзошел результаты спектрального моделирования и кригинга.

Задача обусловленной генерации карт параметров месторождения. Существующие алгоритмы обусловленной генерации карт параметров месторождения плохо справляются с обусловливанием на данные гидродинамических исследований. Разработан новый способ привлечения экспертных знаний геолога о месторождении (таких как типы осадконакопления, преимущественное направление анизотропии и т.д.) и последующей генерации карт месторождения с обусловливанием на эти экспертные знания, а также на все данные, использованные для решения первой задачи. Данный способ реализован с помощью архитектур генеративных состязательных сетей WGAN и PatchGAN, также ведутся эксперименты

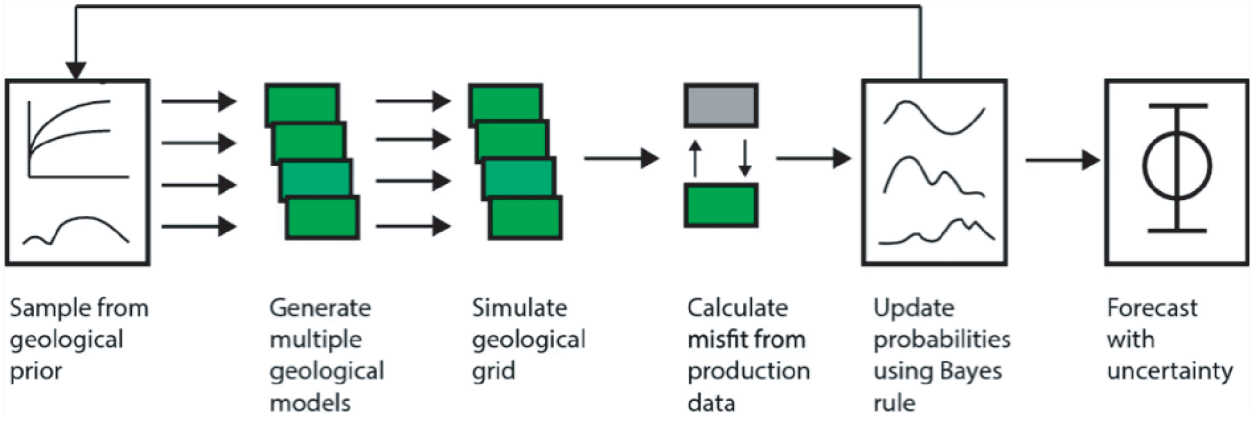

Рис. 6. Последовательность шагов при адаптации модели месторождения.

с технологиями оптимального транспорта [1, 3, 19]. Сотрудники Научно-Исследовательского Института Искусственного Интеллекта (AIRI) и Сколковского Института Науки и Технологий уже использовали разработанные в проекте инструменты оптимального транспорта для повышения разрешения реальных изображений [17]. Полученные результаты подтверждают эффективность методов оптимального транспорта при обработке изображений и позволяют рассчитывать и на хорошую точность при обработке карт распределений физических свойств.

*Задача ускорения и уточнения процесса прокси-адаптации гидродинамической модели***.** Процесс адаптации являет собой наиболее сложную для реализации часть проводимых работ по построению единой модели месторождения. На рис. 6 изображена общепринятая последовательность действий для байесовской адаптации с учетом неопределенности [14, 15]. Сложность процесса адаптации состоит в решении обратной задачи коррекции модели для соответствия расчетов добычи данным измерений, при этом прямая задача расчетов добычи решается обычно с помощью вычислительно сложных недифференцируемых гидродинамических симуляторов по типу Schlumberger Eclipse или tNavigator. Поэтому для решения данной задачи применимы в основном методы на основе Монте-Карло, а также для ускорения процесса – представления в сжатом пространстве признаков и разного рода методы глобальной оптимизации.

Однако в любом случае остается проблема медленности расчета симулятора и сложности всего процесса адаптации в многомерном пространстве параметров. Одним из возможных решений является предварительная адаптация с помощью быстрых упрощенных (прокси) моделей, и затем точная адаптация с помощью симулятора. Ведутся работы по прокси-адаптации гидродинамической модели месторождения с учетом неопределенностей исходных данных. Также важным направлением исследований являются ускорение и уточнение расчетов прокси-моделей с помощью физически информированных нейронных сетей, это направление является одним из наиболее многообещающих и ключевых для работы центра.

Конечной целью программы исследований СМП является сокращение времени построения модели месторождения и ее адаптации от месяцев до нескольких недель, причем большая часть действий будет выполняться в автоматическом режиме с помощью развитых для этого технологий ИИ. Для достижения этой цели требуется не только развивать отдельные части вычислительной цепочки (см. рис. 6), но также выстраивать верхнеуровневое управление процессом построения модели месторождения с помощью автоматизированных инструментов, интегрирующее в себя как классические методы, так и разные инновационные решения, доступные для использования, и позволяющие с помощью умного планирования выбрать путь решения отдельных задач в зависимости от требуемых метрик качества финального результата и имеющихся ограничений. Для решения этой задачи в центре развиваются методы иерархического моделирования и мультиагентного взаимодействия на основе технологий общего научного и инженерного ИИ.

6. ЗАКЛЮЧЕНИЕ

Миссия Исследовательского Центра Прикладного ИИ состоит в создании моделей и фреймворков ИИ для решения задач устойчивого развития промышленности и экономики РФ. Для разработки модулей ИИ и решения соответствующих прикладных задач в Центре разрабатывается программный инструментарий ИИ в виде платформы с универсальными фреймворками для повышения вычислительной эффективности

нейросетевых решений ИИ, учета физики исследуемых процессов в моделях машинного обучения, и консолидация мультимодальных данных от разнородных источников. В свою очередь, программный инструментарий ИИ имплементирует разрабатываемые в Центре методы и алгоритмы машинного обучения и интеллектуального анализа данных. В работе описаны основные прикладные направления исследований Центра Прикладного ИИ и приведены уже достигнутые научные и прикладные результаты.

СПИСОК ЛИТЕРАТУРЫ

1. *Rout L., Korotin A., Burnaev E.* Generative Modeling with Optimal Transport Maps. ICLR, 2022.
2. *Barannikov S., Trofimov I., Balabin N., Burnaev E.* Representation Topology Divergence: A Method for Comparing Neural Network Representations. ICML, 2022.
3. *Korotin A., Kolesov A., Burnaev E.* Kantorovich Strikes Back! Wasserstein GANs are not Optimal Transport? Neurips datasets track, 2022.
4. *Lloyd E., Shepherd T.* Environmental catastrophes, climate change, and attribution. Annals of the New York Academy of Sciences, 1469, 02 2020.
5. Эксперты раскрыли данные МЧС по регионам с самыми частыми затоплениями. https://www.rbc.ru/society/26/08/2021/612639f29a79473d011e9e1, 2021. Online;accessed 13-June-2022.
6. *Alley W.M.* The Palmer drought severity index: limitations and assumptions. Journal of Applied Meteorology and Climatology. 1984. Vol. 23. No. 7. pp. 1100–1109.
7. *Anisimov O.A., Shiklomanov N.I., Nelson F.E.* Variability of seasonal thaw depth in permafrost regions: a stochastic modeling approach. Ecological modelling. – 2002. – Vol. 153. – No. 3. – pp. 217–227.
8. *Gusak J., Cherniuk D., Shilova A., Katrutsa A., Bershatsky D., Zhao X., Eyraud-Dubois L., Shlyazhko O., Dimitrov D., Oseledets I.* Beaumont O.Survey on Large Scale Neural Network Training. Proc. of the 31st Int. Joint Conf. on Artificial Intelligence and the 25th European Conf. on Artificial Intelligence (IJCAI-ECAI), 2022.
9. *Novikov G., Bershatsky D., Gusak J., Shonenkov A., Dimitrov D., Oseledets I.* Few-Bit Backward: Quantized Gradients of Activation Functions for Memory Footprint Reduction. arXiv:2202.00441, 2022.
10. *Bershatsky D., Mikhalev A., Katrutsa A., Gusak J., Merkulov D., Oseledets I.* Memory-Efficient Backpropagation through Large Linear Layers. arXiv:2201.13195, 2022.
11. *Leplat V., Merkulov D., Katrutsa A., Bershatsky D., Oseledets I.* NAG-GS: Semi-Implicit, Accelerated and Robust Stochastic Optimizers. arXiv:2209.14937, 2022.
12. *Budennyy S., Lazarev V., Zakharenko N., Korovin A., Plosskaya O., Dimitrov D., Arkhipkin V., Oseledets I., Barsola I., Egorov I., Kosterina A., Zhukov L.* Eco2AI: Carbon Emissions tracking of Machine Learning models as the first step towards sustainable AI. arXiv:2208.00406, 2022.
13. *Grigoryev T., Verezemskaya P., Krinitskiy M., Anikin N., Gavrikov A., Trofimov I., Balabin N., Shpilman A., Eremchenko A., Gulev S., Burnaev E., Vanovskiy V.* Data-Driven Short-Term Daily Operational Sea Ice Regional Forecasting. arXiv: 2210.08877, 2022.
14. *Arnold D. et al.* Uncertainty quantification in reservoir prediction: part 1 – model realism in history matching using geological prior definitions. Mathematical Geosciences. – 2019. – Vol. 51. – No. 2. – pp. 209–240.
15. *Demyanov V. et al.* Uncertainty quantification in reservoir prediction: part 2 – handling uncertainty in the geological scenario. Mathematical Geosciences. – 2019. – Vol. 51. – No. 2. – pp. 241–264.
16. *Вановский В.В., Дупляков В.М., Попков Д.О., Морозов А.Д., Вайнштейн А.Л., Осипцов А.А., Бурнаев Е.В.* Построение куба проницаемости с адаптацией на ГИС ГДИС и сейсмические исследования. Тезисы конференции “Интеллектуальный анализ данных в нефтегазовой отрасли. Третья научно-практическая конференция”, 21–23 сентября 2022 г., Новосибирск, Россия, 2022.
17. *Gazdieva M., Rout L., Korotin A., Kravchenko A., Filippov A., Burnaev E.* An Optimal Transport Perspective on Unpaired Image Super-Resolution. arXiv:2202.01116, 2022.
18. *Rakhimov R., Ardelean A.-T., Lempitsky V., Burnaev E.* NPBG++: Accelerating Neural Point-Based Graphics. CVPR, 2022.
19. *Asadulaev A., Korotin A., Egiazarian V., Burnaev E.* Neural Optimal Transport with General Cost Functionals. arXiv:2205.15403, 2022.
20. *Matveev A., Rakhimov R., Artemov A., Bobrovskikh G., Egiazarian V., Bogomolov E., Panozzo D., Zorin D., Burnaev E.* DEF: Deep Estimation of Sharp Geometric Features in 3D Shapes. ACM Transactions on Graphics. Volume 41, Issue 4July 2022, Article No.: 108 pp. 1–22.

EDN: TSZPQT
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 28–32

РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ИССЛЕДОВАТЕЛЬСКИХ ЦЕНТРОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

УДК 004.8

О РАЗРАБОТКЕ ПРИКЛАДНЫХ РЕШЕНИЙ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ОБЕСПЕЧЕНИЯ ТЕХНОЛОГИЧЕСКОЙ БЕЗОПАСНОСТИ

© 2022 г. А. А. Масютин[1,*], А. В. Савченко[1], А. А. Наумов[1], С. В. Самсонов[1], Д. Н. Тяпкин[1], Д. В. Беломестный[1], Д. С. Морозова[1], Д. А. Бадьина[1]

Представлено академиком РАН Г.И. Савиным
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

Основной миссией Исследовательского центра в сфере искусственного интеллекта НИУ ВШЭ (Центра ИИ) являются развитие и внедрение технологий искусственного интеллекта в различные сферы жизни человека и общества, отрасли науки и секторы экономики. В рамках деятельности Центра ИИ разрабатываются новые технологии искусственного интеллекта, позволяющие расширить область применения искусственного интеллекта; создаются программные инструменты и средства для применения искусственного интеллекта в отраслях науки и бизнеса, разрабатывается открытая программная библиотека методов искусственного интеллекта для решения задач, имеющих высокую социальную значимость.

Ключевые слова: технологии искусственного интеллекта, программные инструменты и средства для применения искусственного интеллекта, библиотека методов искусственного интеллекта

DOI: 10.31857/S2686954322070165

1. ВВЕДЕНИЕ

Деятельность Исследовательского центра в сфере искусственного интеллекта НИУ ВШЭ (Центра ИИ) реализуется в рамках федерального проекта “Искусственный интеллект”. Сроки реализации: 2021–2024 гг.

Главная задача Центра ИИ состоит в разработке прикладных решений на основе искусственного интеллекта для обеспечения технологической безопасности и технологического суверенитета Российской Федерации.

Основные направления исследований Центра ИИ: обработка естественного языка и интеллектуальные ассистенты; оптимизация краудсорсинговых платформ; ИИ в промышленности; рекомендательные системы; повышение качества аудио- и видеоданных; технологии для медиа; ИИ в образовании; ИИ в биоинформатике; ИИ и право. В общей сложности в Центре ИИ реализуется более 25 проектов, которые условно можно разделить на 2 большие группы: проекты для бизнеса и проекты для социально-значимой сферы (рис. 1). Каждый проект соответствует мировым фронтирам в конкретной технологической области.

Индустриальными партнерами Центра ИИ являются Сбербанк, Яндекс, MTS AI.

В качестве примеров проектов для бизнеса можно привести следующие:

Для Сбербанка:

– создание и оптимизация новых языковых моделей,

– разработка интеллектуальных систем с широкой областью применения: от задач повышения качества аудио до выявления признаков подозрительных транзакций.

Для Яндекса:

– алгоритмы оптимизации взаимодействия пользователей на различных платформах,

– решение, направленное на выявление субъективно неоднозначного (спорного) контента.

Для МТС:

– решения для телеком-рынка (например, оптимизация чат-ботов),

– применение генеративных моделей для восстановления качества изображений и аудио, а также возможности колоризации.

Для социально-значимой сферы в Центре разрабатываются решения для повышения качества

[1] *Национальный исследовательский университет “Высшая школа экономики”, Москва, Россия*
*****E-mail: amasyutin@hse.ru*

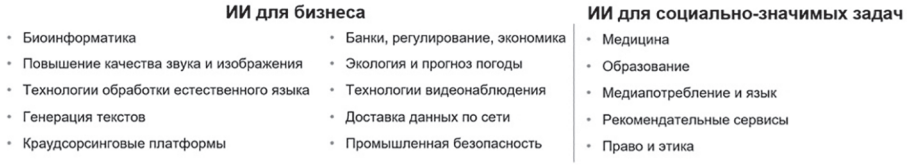

Рис. 1. Направления деятельности Центра ИИ.

образовательного процесса; модели машинного обучения для медицины, позволяющие снизить издержки при разработке медицинских препаратов; решения для судопроизводства и ответственного медиапотребления, а также ведутся исследования, направленные на решение вопросов в области этики ИИ.

Помимо трех индустриальных партнеров, ведется работа над привлечением дополнительных партнеров, которым интересны разработки Центра ИИ. Взаимодействие научных проектов и индустрии в дальнейшем приведет к появлению новых актуальных решений и продуктов на рынке.

Например, в 2022 г. подписаны четыре соглашения о сотрудничестве Центра ИИ с российскими компаниями, в том числе в сфере медицинских исследований, юридических услуг, туризма. Также ведутся переговоры с компаниями топливно-энергетического комплекса, с одним из крупнейших на российском рынке производителей лекарственных препаратов и другими компаниями.

За первое полугодие работы центра в этом году опубликован ряд статей, затрагивающих широкий спектр вопросов в области ИИ. Далее мы более подробно рассмотрим основные научно-прикладные результаты деятельности исследовательских проектов Центра ИИ.

2. ОБУЧЕНИЕ, ПОНИМАНИЕ И ОПТИМИЗАЦИЯ МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Данный крупный проект “Обучение, понимание и оптимизация моделей искусственного интеллекта” имеется несколько важных направлений.

В рамках исследования по теме “Разработка и верификация алгоритмов и дополнительных математических моделей эффективного сэмплирования данных” был предложен эффективный подход к уменьшению дисперсии для аддитивных функционалов от цепей Маркова, основанный на новом представлении мартингала с дискретным временем. Исследуемый подход не требует знания стационарного распределения или конкретной структуры целевой плотности. Проведенный тщательный анализ теоретических свойств предлагаемого алгоритм показал, что такой алгоритм позволяет добиться улучшения качества работы наивного Монте-Карло алгоритма при фиксированном вычислительном бюджете. Численная эффективность нового метода продемонстрирована для методов Монте-Карло по схеме марковской цепи (MCMC), основанных на динамике Ланжевена [1].

Также в рамках исследования цепей Маркова решалась проблема поиска оптимальной политики для марковских процессов принятия решений (MDP) с бесконечным горизонтом планирования. Для этой цели предлагается вариант стохастического зеркального спуска для задач выпуклого программирования с непрерывными Липшицевыми функционалами [2]. Важной деталью является возможность использовать неточные значения функциональных ограничений и вычислять значение двойственных переменных. Этот алгоритм был проанализирован в общем случае и получена оценка скорости сходимости, которая не накапливает ошибок во время работы метода.

Так был получен первый параллельный алгоритм для эргодичных MDP с усредненным вознаграждением, использующий генеративную модель, без сведения к дисконтированным MDP. Одной из ключевых особенностей представленного метода являются низкие коммуникационные затраты в контексте распределенного программирования в централизованной, даже очень большой, сети.

Ранее в других исследованиях было замечено, что обычный (суб)-градиентный метод может обрабатывать функциональные ограничения без дополнительных затрат с точки зрения суммарного количества итераций [3]. Способность к параллельным вычислениям имеет решающее значение для любого крупномасштабного приложения. В дальнейшем исследование развивалось в следующих направлениях:

– использование стохастических (суб)-градиентов;

– вычисление двойственных переменных для двойственной задачи Лагранжа (прямо-двойственность).

Первое направление имеет важное значение для масштабируемости высокопроизводитель-

ных приложений, поскольку вычисление точных градиентов может быть невозможным или долгим. Ценность второго типа улучшений сильно зависит от конкретного приложения, но всегда дает возможность использовать критерии остановки, основанные на зазоре двойственности.

Благодаря результатам проекта, есть возможность объединить оба свойства и предложить стохастический субградиентный метод, который вычисляет двойные переменные без дополнительных затрат. Кроме того, в работе применяется предложенный прямо-двойственный стохастический зеркальный спуск к проблеме максимизации среднего вознаграждения в эргодичных марковских процессах принятия решений.

Следует отметить, что в Центре ИИ также ведутся исследования в области обучения с подкреплением, которые отражают актуальность общей деятельности Центра ИИ. В свою очередь, обучение с подкреплением (Reinforcement learning, RL) важно не только как объект исследований, но и как инструмент для решения важных практических задач. Обучение с подкреплением – один из видов машинного обучения. Ключевое отличие данного подхода от классического машинного обучения – это постоянное взаимодействие агента (алгоритма) со средой, от которой он получает обратную связь в виде поощрений и наказаний. Цель агента – максимизировать сумму наград, которые среда дает ему за “правильное” взаимодействие с ней. Для достижения этой цели агент должен балансировать между исследованием окружающей среды и использованием текущих знаний о ней для максимизации вознаграждений.

В рамках исследования RL [4, 5] был предложен алгоритм Bayes-UCBVI для обучения с подкреплением в табулярном эпизодическом MDP (марковском процессе принятия решений). Данный алгоритм является естественным обобщением алгоритма Bayes-UCBVI Кауфман и др., 2012 для многоруких бандитов. Данный метод использует апостериорную квантиль Q-функции как верхнюю доверительную границу для оптимальной Q-функции.

В данном проекте еще одно исследование было направлено на неасимптотический анализ алгоритмов линейной стохастической аппроксимации (LSA) с фиксированным размером шага. Это семейство методов возникает во многих задачах машинного обучения и используется для получения приближенных решений линейной системы $Ax = b$. В данной задаче величины A и b не наблюдаемы, однако доступна последовательность случайных наблюдений $\{(A_n, b_n)\}$, оценивающих A и b соответственно. Анализ был основан на новых результатах, касающихся моментов и оценок больших уклонений для произведений случайных матриц, которые, как показано, являются точными. Данные оценки позволяют получить более точные оценки ошибки алгоритмов линейной стохастической аппроксимации при более слабых условиях на последовательность наблюдений $\{(A_n, b_n)\}$, чем в предыдущих работах. В частности, результаты не требуют симметричности случайных матриц $\{A_n\}$ [6].

Алгоритмы линейной стохастической аппроксимации имеют массу практических применений, в частности, в области обработки сигналов, шумоподавления, оптимизации, обучения с подкреплением. С помощью полученных результатов возможно получить новые оценки производительности RL-агентов в задаче оценивания политики, например, с помощью алгоритмов семейства TD (temporal differences).

3. НЕЙРОСЕТЕВЫЕ АЛГОРИТМЫ АНАЛИЗА ДИНАМИКИ ЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ И ВОВЛЕЧЕННОСТИ УЧЕНИКОВ НА ОСНОВЕ ДАННЫХ ВИДЕОНАБЛЮДЕНИЯ

В рамках проекта разрабатывается новый метод обработки видео для анализа эмоционального состояния участников видеоконференций, в том числе, учащихся в среде e-learning. На первом этапе применяются алгоритмы распознавания лиц, трекинга и кластеризации для извлечения последовательностей лиц каждого учащегося. Затем эмоциональные признаки лиц в каждом кадре извлекаются с помощью специальной вычислительно эффективной нейронной сети, которая предварительно обучена идентификации лиц и дообучена для классификации эмоций на фотографиях из набора AffectNet [7]. При этом в процессе обучения применяется специально разработанный робастный метод оптимизации. Показано, что полученные в результате векторы признаков лиц могут быть использованы для быстрого решения сразу нескольких задач: предсказания степени вовлеченности участников, их индивидуальных и групповых эмоций.

Разработанная нейросетевая модель может быть использована для обработки видео в режиме реального времени даже на мобильном устройстве пользователя без необходимости отправки видео их лиц на удаленный сервер или компьютер. Кроме того, продемонстрирована возможность подготовить краткое изложение хода онлайн-мероприятия с помощью создания коротких видеоклипов с фрагментами наиболее характерных эмоций каждого участника.

Экспериментальное исследование для наборов данных из конкурсов EmotiW показало, что предлагаемая сеть значительно превосходит существующие одиночные (не-ансамблевые) модели.

Основной вклад проекта состоит в следующем:

– ключевым компонентом предлагаемого подхода являются “легковесные” нейронные модели для распознавания выражений лиц на основе архитектур EfficientNet и MobileNet, которые могут использоваться для извлечения эмоциональных признаков лиц на фото и видео;

– предложена эффективная нейросетевая модель для одновременного обнаружения вовлеченности и распознавания эмоций на индивидуальном и групповом уровнях по видеоданным;

– представлена новая технология анализа лиц на видео в реальном времени.

Анализ вовлеченности и внимания может повысить эффективность и производительность проведения онлайн-встреч, а также обучения. Более того, была выявлена корреляция анализа выражения лиц и понимания.

Таким образом, целью проекта является разработка быстрой и точной методики для классификации эмоций и вовлеченности, которая может быть реализована в программном обеспечении на ноутбуках без мощных графических процессоров и или мобильных устройствах пользователей.

4. ДИАГНОСТИЧЕСКИЕ И АССИСТИВНЫЕ РЕЧЕВЫЕ ТЕХНОЛОГИИ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Целью нейрохирургических вмешательств при опухолях головного мозга, рефрактерной эпилепсии, артериовенозных мальформаций является удаление патологической ткани при сохранении функционально необходимых областей, чтобы соответствующие функции, в том числе когнитивные [8], не были нарушены после нейрохирургии [9]. Одной из важнейших функций является обработка речи, которая лежит в основе человеческого общения, и ее нарушение негативно влияет на возвращение к работе, социальную интеграцию и общее качество жизни.

В рамках работы проекта “Диагностические и ассистивные речевые технологии на основе искусственного интеллекта” была исследована парадигма для предоперационного языкового картирования с использованием МРТ у русскоязычных людей и предоставлены методологические доказательства ее надежности при повторном тестировании.

Интраоперационное картирование языковой функции в головном мозге может быть дополнено предоперационным картированием с помощью функциональной магнитно-резонансной томографии (ФМРТ/fMRI). Валидность парадигмы fMRI “языковой локализатор” в решающей степени зависит от выбора оптимальной языковой задачи и исходных условий. В этом исследовании представлен новый fMRI “языковой локализатор” на русском языке, использующий явное завершение предложения, задачу, которая всесторонне задействует языковую функцию, включая как понимание, так и воспроизведение на уровне слов и предложений [10].

Парадигма была подтверждена на 18 неврологически здоровых добровольцах, которые участвовали в двух сеансах сканирования для оценки надежности повторного тестирования. На групповом уровне активизировались как передние, так и задние области, связанные с языком.

В целом исследование демонстрирует валидность и надежность задачи завершения предложения для отображения языковой функции в мозге. Кроме того, оно вносит вклад в общие данные о надежности fMRI при повторном тестировании.

5. РЕЗУЛЬТАТЫ ЦЕНТРА ИИ ЗА 2022 ГОД

Одним из ключевых показателей эффективности работы Центра ИИ является готовность некоторых проектов зарегистрировать результаты работ в виде программных обеспечений для ЭВМ и патентов на результаты интеллектуальной деятельности.

Так, например, в рамках проекта “Искусственный интеллект в макромоделировании и прогнозировании экономических процессов и финансовых взаимосвязей с учетом сентимента участников рынка” была разработана программа для анализа общерыночного сентимента инвесторов в трех тональностях по российскому рынку акций по сообщениям в каналах мессенджера, которая обеспечивает автоматизированный сбор, обработку и выявление сентимента (тональности) текстовых материалов инвестиционной тематики за указанный пользователем период времени при заданном минимальном периоде в один день.

Классификация материалов осуществляется с применением предобученной модели двунаправленной нейронной сети Google BERT, которая была усовершенствована с целью повышения качества классификации новостных текстов финансово экономического характера на наборе данных. Программа способна рассчитать 11 различных метрик общерыночного сентимента инвесторов по российскому рынку акций.

В данный момент продолжается работа по подготовке заявок еще несколькими проектами Центра ИИ в области компьютерного зрения, лингвистического анализа текстов и т.д.

6. ЗАКЛЮЧЕНИЕ

Результаты работы Центра ИИ подтверждают свою актуальность и позволяют решать различ-

ные задачи, стоящие перед российской экономикой:

– Языковые модели развиваются крайне динамично, в том числе и для русского языка. Для индустриальных партнеров в Центре ИИ ведутся работы по разработке моделей, оптимизирующих работу виртуальных ассистентов, чат-ботов. Результаты проектов позволят снизить рутинную нагрузку на службы поддержки клиентов, что в свою очередь положительно сказывается на снижении затрат на содержание колл-центров.

– В условиях ограничения доступа к технологическим рынкам повышается актуальность рекомендательных сервисов выявления технологических трендов и научных фронтиров. Именно поэтому в Центре ИИ имеется ряд проектов, которые занимаются разработкой рекомендательных систем.

– Так, в рамках Центра была разработана программная библиотека, рекомендующая применение той или иной языковой модели, что позволяет при неизменном бюджете повышать качество решений NLP задач бизнеса, за счет выбора наиболее перспективных языковых моделей для экспериментов.

– Разработаны первые версии алгоритмов обработки открытых текстовых данных по компаниям, которые позволяют банковским организациям снижать операционные расходы на риск-анализ за счет нового потока данных о ESG-практиках компаний.

– Так как имеется ряд ограничений по доступу к образовательным продуктам, требуется создание уникальных импортозамещающих отечественных курсов. В 2022 г. было разработано 4 курса семинаров и лекций по тематике Центра ИИ, а также проведены школы по машинному обучению. К тому же одним из перспективных направлений Центра ИИ является использование искусственного интеллекта в образовании, в частности, разработка моделей персонализированного адаптивного обучения.

– Вследствие усложнения логистических цепочек, в том числе в сфере импортных лекарственных средств, повышается актуальность разработки отечественных лекарственных препаратов. В рамках Центра разрабатывается ПО для прогноза формы определенных участков антител. Решение позволит фарм-компаниям сократить число физических экспериментов, что позитивно скажется как на сроках разработки лекарств, так и на их стоимости.

Таким образом, результаты работы Центра должны лечь в основу создания и масштабирования прикладных решений в области искусственного интеллекта, которые находят применения в различных отраслях российской экономики.

СПИСОК ЛИТЕРАТУРЫ

1. *Belomestny D., Moulines E., Samsonov S.* Variance reduction for additive functional of Markov chains via martingale representations // Statistics and Computing. 2022. V. 32. № 1. Article 16. https://doi.org/10.1007/s11222-021-10073-z
2. *Tiapkin D., Gasnikov A.* Primal-Dual Stochastic Mirror Descent for MDPs // Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022). 2022. V. 151. P. 9723–9740. https://doi.org/10.48550/arXiv.2103.00299
3. *Nemirovski A. and Yudin D.* Problem Complexity and Method Efficiency in Optimization // A Wiley-Interscience publication. Wiley. 1983.
4. *Tiapkin D., Belomestny D., Moulines É., Naumov A., Samsonov S., Tang Y., Valko M., Ménard P.* From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses, in Proceedings of the 39th International Conference on Machine Learning. 2022. V. 162. P. 21380–21431. https://doi.org/10.48550/arXiv.2205.07704
5. *Donald B. Rubin*. The bayesian bootstrap // The annals of statistics. 1981. P. 130–134.
6. *Durmus A., Moulines E., Naumov A., Samsonov S., Scaman K., Wai H.* Tight High Probability Bounds for Linear Stochastic Approximation with Fixed Stepsize // Advances in Neural Information Processing Systems 34 (NeurIPS 2021), pp. 1–12. https://doi.org/10.48550/arXiv.2106.01257
7. *Savchenko A., Savchenko L., Makarov I.* Classifying emotions and engagement in online learning based on a single facial expression recognition neural network// IEEE Transactions on Affective Computing, July 2022, pp. 1–12. https://doi.org/10.1109/TAFFC.2022.3188390
8. *Satoer D., Visch-Brink E., Dirven C., Vincent A.* Glioma surgery in eloquent areas: can we preserve cognition? Acta Neurochi. 2016. V. 158. P. 35–50. https://doi.org/10.1007/s00701-015-2601-7
9. *Duffau H.* The challenge to remove diffuse low-grade gliomas while preserving brain functions. Acta Neurochir. 2012. V. 154. P. 569–574. https://doi.org/10.1007/s00701- 012-
10. *Elin K., Malyutina S., Bronov O., Stupina E., Marinets A., Zhuravleva A., Dragoy O.* A New Functional Magnetic Resonance Imaging Localizer for Preoperative Language Mapping Using a Sentence Completion Task: Validity, Choice of Baseline Condition, and Test–Retest Reliability // Frontiers in Human Neuroscience. 2022. V. 16. P. 1–21. https://doi.org/10.3389/fnhum.2022.791577

EDN: AKVQUB
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 33–40

РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ИССЛЕДОВАТЕЛЬСКИХ ЦЕНТРОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

УДК 004.8

ИНТЕЛЛЕКТУАЛЬНЫЕ ТЕХНОЛОГИИ ЦИФРОВОЙ ТРАНСФОРМАЦИИ ПРОМЫШЛЕННЫХ ПРОИЗВОДСТВ

© 2022 г. А. В. Бухановский[1,*]

Представлено академиком РАН В.Б. Бетелиным
Поступило 28.10.2022 г.
После доработки 31.10.2022 г.
Принято к публикации 01.11.2022 г.

Представлены облик и ключевые результаты деятельности исследовательского центра в сфере искусственного интеллекта “Сильный ИИ в промышленности” на базе Университета ИТМО. Изложены концепция и целеполагание центра, ключевые научные результаты, открытые фреймворки и библиотеки, а также практические результаты, демонстрирующие внедрения в различных областях промышленности. Отдельное внимание уделено активностям в развитии таких перспективных технологий ИИ, как автоматическое машинное обучение, генеративный дизайн и планирование целесообразной деятельности в условиях неопределенности и неполноты данных.

Ключевые слова: искусственный интеллект, автоматическое машинное обучение, генеративный дизайн, баейсовы сети, эвристическая оптимизация, открытый фреймворк

DOI: 10.31857/S2686954322070037

1. ВВЕДЕНИЕ

Исследовательский центр “Сильный искусственный интеллект в промышленности” создан при Университете ИТМО в 2021 г. в рамках реализации федерального проекта “Искусственный интеллект” и постановления Правительства Российской Федерации от 5 июля 2021 г. № 1120. Целью деятельности Центра является создание программного обеспечения (ПО) на базе технологий сильного ИИ для воспроизведения творческих профессиональных функций отраслевого специалиста – в целях разработки систем поддержки принятия решений (СППР), обеспечивающих процессы цифровой трансформации и интеллектуализации промышленных производств. В данном случае под сильным ИИ, в логике [1], понимается алгоритмическое воспроизведение высших когнитивных функций человека при решении творческих задач, связанных с извлечением и оперированием смыслами для широкого класса приложений на основе перспективных технологий ИИ. Это отличает его от традиционного (“слабого”) ИИ, работающего с базовыми когнитивными функциями (распознавание речи, машинное зрение, вывод на априорных знаниях конкретной области и пр.).

Стратегическим партнером Центра выступает Ассоциация “Искусственный интеллект в промышленности” [2], учрежденная в 2021 г. ПАО “Газпром нефть” и Правительством Санкт-Петербурга в формате государственно-частного партнерства. Она координирует проектную и образовательную деятельность вузов и индустриальных партнеров из различных отраслей промышленности. Несмотря на значимое проникновение технологий ИИ в деятельность промышленных предприятий, они в большинстве своем обеспечивают лишь фрагментарную автоматизацию отдельных операций или процессов за счет решения таких задач, как структурирование данных, анализ временных рядов технических параметров и изображений, а также диагностика и прогнозирование технологических процессов и состояния оборудования. Однако для системной цифровой трансформации производств необходимо решение задач, связанных с поддержкой или даже автоматизацией воспроизведения высших когнитивных функций специалистов, связанных с креативным принятием решений. Особенную значимость это приобретает в ситуациях с неопределенностью и неполнотой данных, например, в ходе концептуального инжиниринга, на этапах технико-экономического обоснования, организации и строительства производств, когда

[1] *Федеральное государственное автономное образовательное учреждение высшего образования “Национальный исследовательский университет ИТМО”, Санкт-Петербург, Россия*
*****E-mail: avbukhanovskii@itmo.ru*

последействие любой ошибки может быть значимо, а цена высока.

Промышленность является благодарной областью приложения технологий ИИ, поскольку не только опирается на эффективные средства сбора и измерения (первичной интерпретации) данных с контролируемым качеством, но и имеет альтернативные источники априорных знаний (причинно-следственные математические модели, инженерное ПО, экспертные обобщения, справочники, руководства, регламенты и пр.). Потому это позволяет гибко сочетать подходы ИИ на данных с технологиями символьного ИИ на знаниях, что в ряде случаев является ключевым для приближения к возможностям сильного ИИ.

Концепция и продукт Центра. Центр ориентирован на решение и практическое воплощение в форме отраслевого ПО новых классов задач, являющихся объектом творческой деятельности предметного специалиста (конструктора, технолога, управленца), связанных с объектами техники и технологий, и характерных именно для деятельности промышленных предприятий. К ним относятся:

Автоматическое создание и обучение новых интеллектуальных цифровых объектов (предсказательных моделей и иных сущностей ИИ). Специфика промышленных производств состоит в существенно большей вариативности технологических и обеспечивающих процессов, чем в традиционных областях применения ИИ (банки, ритейл и пр.). Как следствие, прямое тиражирование прикладных систем ИИ даже между разными филиалами одного предприятия требует значимого труда специалистов в области машинного обучения. Потому применение технологий автоматического машинного обучения (МО), позволяющего существенно сократить временные и ресурсные затраты, в данном случае является приоритетным. Особенно это важно для решения задач автоматического моделирования, когда модели ИИ на данных комбинируются с объектами прикладной математики, реализуемыми классическими численными методами моделирования и оптимизации.

Автоматическое проектирование объектов реального мира и абстрактных цифровых структур, определяющих процессы целесообразной деятельности. Несмотря на обширный отраслевой задел в области систем автоматизированного проектирования, в подавляющем большинстве сейчас такие решения носят советующий характер и основаны на воспроизведении лучших практик специалистов-конструкторов. Однако переход к парадигме генеративного дизайна, имитирующего логику конструктора не на основе априорных знаний, а за счет использования альтернативных процессов рационального выбора (например, эволюционных вычислений), позволяет формировать нестандартные конструкторские решения, часто превышающие когнитивные возможности человека.

Автоматическая валидация и верификация проектных, технических и управленческих решений. Цифровая трансформация производств неизбежно приводит к усложнению условий формирования и принятия решений за счет увеличения объема учитываемых факторов. В этих условиях даже решения, принимаемые квалифицированным специалистом (и даже группами специалистов, не говоря про системы ИИ), могут считаться лишь условно адекватными и достоверными. Потому принципиальным является развитие алгоритмов ИИ, которые (а) обеспечивают контролируемое качество предварительной аналитической обработки данных для определения фактов, и (б) решают обратную задачу оценки решений путем автоматической генерации минимальной системы критериев, необходимых для проверки их содержательности и непротиворечивости. Таким образом, это дает основу для создания самообъясняющихся систем ИИ, способных интерпретировать решения, как принятые человеком, так и рекомендованные иными системами ИИ.

Автоматическая настройка вычислительно-сложных многопараметрических математических моделей. Практика внедрения промышленных цифровых двойников связана с использованием крайне ресурсоемких вычислительных моделей (например, на основе уравнений трехмерной гидроаэродинамики, тепломассопереноса и пр.). В силу объективного наличия экспериментальных замыканий такие модели обычно имеют большое число эмпирических параметров, настраиваемых по данным. Однако сами данные, как правило, ограничены в объеме и одинаковых условиях эксперимента. Как следствие, применить для настройки таких моделей классические методы оптимизации не всегда возможно в силу (а) вычислительной ресурсоемкости, (б) неполноты данных (работа с малой выборкой), (в) разнообразия условий получения данных (их гетерогенности). Однако эта проблема решается паллиативным аппаратом – интеллектуальными (эвристическими) методами оптимизации, сочетающими классические методы с алгоритмами ИИ, имитирующими процесс поиска решения специалистом-предметником или использующими иные (например, био-инспирированные) принципы.

Автоматическое комплексирование цифровых решений с элементами ИИ, созданными на основе различных подходов и парадигм. Процессы цифровой трансформации крупных предприятий, как правило, ставят своей целью объединение возможностей различных (в том числе, уже имею-

Рис. 1. Облик продукта центра (на горизонте 2024 г.): информационная технология управления жизненным циклом крупных распределенных промышленных предприятий.

щихся) цифровых систем, формирующих процессы подготовки и принятия решений. Это включает в себя совместное использование традиционного моделирования, методов оптимизации и прикладных систем ИИ, в т.ч. построенных на совершенно разных принципах (от экспертных систем – до глубокого обучения). Как следствие, это требует развития технологий композитного ИИ, позволяющих бесшовно комбинировать различные системы ИИ, а также строить на их основе новые решения с заданным функциональным назначением.

Полноценное решение перечисленных выше задач, по-видимому, невозможно путем использования существующих библиотек ИИ, средств разработки и отраслевого инженерного ПО. Потому Центр в своей деятельности ориентируется на создание комплексной информационной технологии управления жизненным циклом крупных распределенных промышленных предприятий, которая включает в себя библиотеки и фреймворки алгоритмов сильного ИИ, средства разработки и оценки моделей ИИ, концептуальные СППР для различных видов отраслевой деятельности, а также конкретные решения (ПО для различных задач промышленности), кастомизируемые для разных индустриальных партнеров. На рис. 1 приведен облик цифрового продукта Центра.

Научные результаты. В 2022 г. научная повестка Центра была ориентирована на разработку новых методов и алгоритмов, реализующих перспективные технологии, потенциально пригодные для решения задач сильного ИИ, включая автоматическое машинное обучение, генеративный дизайн, работу с малыми или частично-размеченными наборами данных и др.

В части автоматического моделирования и автоматического машинного обучения основное внимание уделено развитию подхода для работы с композитными моделями МО, обеспечивающего их адаптацию под сложность конкретной прикладной задачи и/или изменяющихся условий среды за счет алгоритмических манипуляций со структурой модели. Для порождения таких моделей используются различные механизмы эволюционных вычислений. Они позволяют оперировать не только моделями классического МО, но и нейросетевыми моделями, и даже – гибридными моделями, включающими в себя структурные элементы на основе априорных знаний (в т.ч. уравнения в частных производных). Альтернативным направлением, обеспечивающим требуемую гибкость, является оперирование структурой ансамблей моделей МО. Так, в Центре разработан алгоритм ABB (AutoBalanceBoosting), основанный на различных комбинациях ансамблирования (бэггинг и бустинг), в котором все параметры подбираются автоматически. На рис. 2 на несбалансированных данных приведено сравнение результатов работы алгоритма ABB и наиболее популярных аналогов, к которым относятся SOTA-классификаторы (например, RF+Smote) и специальные алгоритмы (BCC, SBAC и пр.).

Из рис. 2 видно, что по отдельным наборам данных ABB стабильно входит в пятерку лучших, а в среднем заметно лучше всех конкурирующих решений.

В области генеративного дизайна существенное внимание уделено развитию гибридных мето-

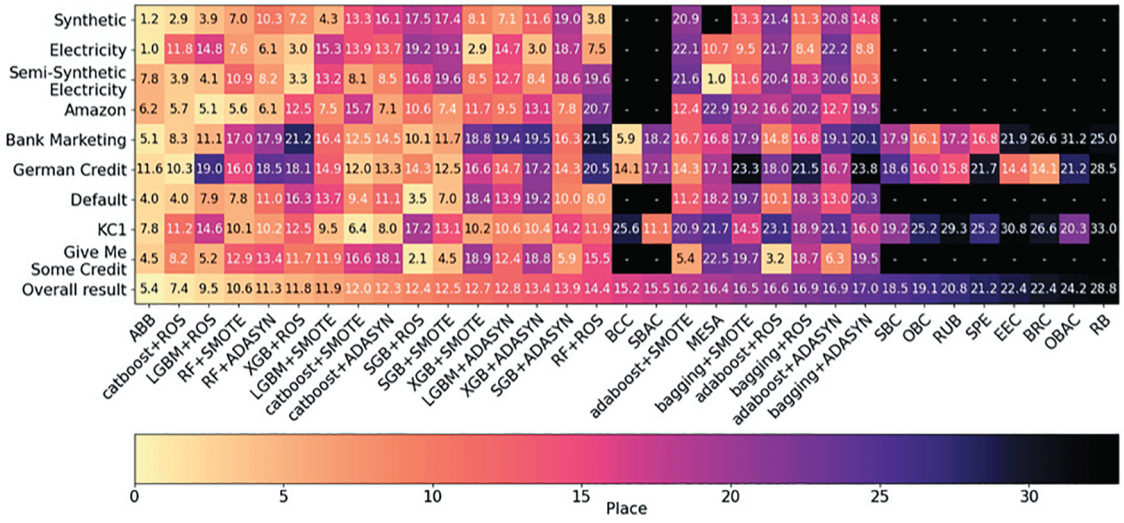

Рис. 2. Сравнение качества результатов классификации алгоритмом ABB и аналогами на несбалансированных наборах данных.

дов, позволяющих эффективно сочетать априорные знания предметной области (формализованные, например, в виде графов знаний), нейронные сети и эвристические методы оптимизации. Их сопряжение в условиях жестких системных ограничений обеспечивается за счет обучения на основе генеративно-состязательных сетей. Отработаны подходы к применению технологий генеративного дизайна с учетом специфики данных и особенностей предметной области для задач проектирования локомоционных робототехнических устройств, волнозащитных сооружений и даже цифровых систем кодирования.

Отдельное внимание уделено автоматическому обучению композитных моделей для многомерных распределений на основе адаптивных байесовых сетей (БС). В частности, предложен алгоритм BigBraveBN, который позволяет обучать большие БС за счет сокращения пространства поиска по принципу отбрасывания “слабых” связей, число связей измеряется с помощью коэффициента Браве. На рис. 3 приведено его сравнение на бенчмарках пакета BNLEARN с аналогами – sparcebn и BiDAG. Видно, что предложенный Центром алгоритм обеспечивает качество по метрике SHD в среднем выше на 20% за в два раза меньшее время.

Одной из научных задач Центра являются развитие алгоритмов для работы с графовыми нейронными сетями (ГНС) и оснащение их новыми механизмами, включая автономное объяснимое обучение, имплементацию времени для моделирования динамики эволюции графов знаний, автоматическую генерацию и структурное обучение композитных графовых моделей МО. Разработано семейство методов, которые на основе ГНС обеспечивают работу с моделями комплексных сетей на данных. Это включает в себя: восстановление пропущенных связей, узлов и атрибутов комплексной сети, предсказательное моделирование макро-характеристик комплексной сети, а также предсказание (экстраполяция) эволюции отдельных узлов и связей комплексной сети во времени. Данный аппарат пригоден как для классических задач на графах (распространение информации, логистика и пр.), так и для работы со сложными структурами данных на основе разнородных источников.

Также в Центре выполняются исследования в области алгоритмов эвристической оптимизации в условиях неопределенности и неполноты данных, развития технологий объяснимого ИИ для обработки изображений, а также генеративного ИИ для поддержки принятия решений. Результаты исследований представлены в 2022 г. на конференциях уровня A* в области ИИ: ICASSP, CEC, ACM SIGKDD, AAAI, NeurIPS, ICRA, Interspeech, см., например, [3–6].

Открытые фреймворки, библиотеки и средства разработки. Научные результаты Центра формируют основу отечественных библиотек и фреймворков, обеспечивающих их распространение в сообществе разработчиков ИИ. Для этого в Университете ИТМО развернута система экспортного контроля научно-технических материалов в области ИИ, которая обеспечивает легитимность их открытого опубликования и продвижения на веб-хостингах. В 2022 г. Центр развивает несколько ключевых продуктов.

Фреймворк автоматического машинного обучения FEDOT.Industrial (https://github.com/ITMO-NSS-team/Fedot.Industrial). Позволяет автомати-

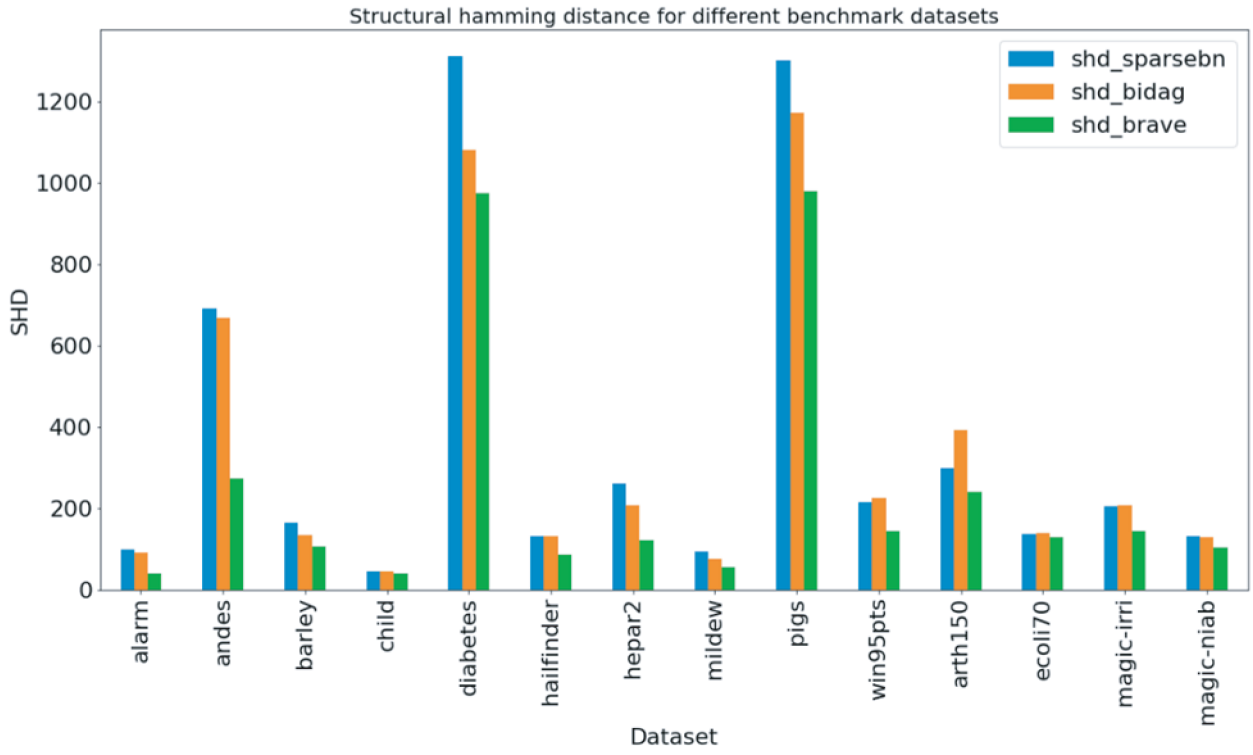

Рис. 3. Сравнение результатов обучения больших БС на разных данных по метрике SHD (меньше – лучше).

чески создавать композитные математические модели МО на промышленных данных. Он предназначен для ускорения процессов создания и обучения моделей на данных для задач предиктивной аналитики (прогноза временных рядов, классификации их фрагментов, выявления выбросов и аномалий) без потери качества. Областью применения являются различные отраслевые задачи диагностики оборудования, обеспечивающие непрерывный сбор данных в виде временных рядов. Фреймворк основан на аппарате автоматического машинного обучения на основе эволюционных вычислений и символьной регрессии для построения композитных моделей на данных (включающих как ИНС, так и классические процедуры машинного обучения). На его базе отработано решение таких типовых задач, как детектирование аномалий в промышленных трубопроводах по данным магнитометрии, оценка рентабельности обогащения полезных ископаемых, определение аномальных режимов работы ротационного оборудования. В целом использование фреймворка позволяет ускорить процесс разработки моделей МО в 10–18 раз, при этом сравнение с SOTA показывает, что в 75% случаев результат не хуже, чем решения, созданные лучшими специалистами вручную.

Фреймворк генеративного дизайна GEFEST (https://github.com/ITMO-NSS-team/GEFEST). Предназначен для решения задач генеративного дизайна геометрических объектов в сплошных средах, динамика которых описывается внешними моделями. Позволяет посредством эволюционных вычислений воспроизводить оптимальную конфигурации и атрибуты геометрического объекта, исходя из заданных критериев качества. При этом сохраняется вся эволюционная цепочка объектов, что является основой для интерпретации и объяснения результатов работы алгоритма. На основе фреймворка отработано решение таких типовых задач, как автоматическое проектирование гидротехнических сооружений на шельфе и проектирование объектов наносенсорики. В целом использование фреймворка позволяет ускорить процесс выработки проектных решений в 15–25 раз по сравнению с квалифицированным специалистом; это значение зависит от степени стандартизации задачи и наличия успешных прототипов.

Помимо фреймворков, коллективами Центра развивается ряд открытых библиотек в области ИИ, в том числе, библиотека автоматического моделирования текстов AutoTM (https://github.com/ngc436/AutoTM) и библиотека предсказательного моделирования комплексных сетей на основе ГНС (https://gitlab.actcognitive.org/anpolol/graphpred).

Для удобства работы с фреймворками и библиотеками Центр разрабатывает инструментальную облачную платформу для проектирования, быстрой разработки и обучения прикладных систем ИИ. В отличие от существующих аналогов, ориентированных, в первую очередь, на удобства разработчика, данная платформа реализует компромиссную стратегию, обеспечивая интересы как разработчика, так и заказчика разработки систем ИИ. При этом разработчик получает поддержку полного цикла создания и эксплуатации моделей ИИ, возможность работы как с традиционными (причинно-следственными) моделями, так и с моделями на данных, простой способ построения моделей без программирования на языках низкого уровня (low-code-диаграммы), а также возможность доступа к удаленным суперкомпьютерным системам и облачным ресурсам. Заказчик, в свою очередь, приобретает прозрачность всего процесса разработки моделей ИИ, доступность всех создаваемых цифровых решений и массивов данных, а также возможности выполнить испытания или исследования конкретной разработки модели ИИ “ad hoc”.

Платформа реализует прозрачную и нативно понятную отраслевому специалисту (не программисту) организацию структуры проектов по разработке моделей ИИ, которая предполагает четкое разделение этапов и формирования по ним отдельных результатов, автоматизацию анализа кода и практическую инкапсуляцию отдельных артефактов для переиспользования и независимого развития, идентификацию наиболее важных и критических точек проекта. При этом ключевой особенностью платформы является использование механизмов автоматического моделирования и машинного обучения для поддержки пользователей в ходе разработки и применения моделей ИИ. К ним относятся: собственно построение моделей с помощью автоматического машинного обучения, достройка (развитие) элементов готовых моделей с целью их улучшения, интеллектуальный анализ содержательной структуры моделей, а также работа с семантическим пространством артефактов, которое отображает единое представление всех элементов, участвующих в проектах данного класса (включая модели, результаты, данные, пользователей, проекты и т.д.). Платформа позволяет публиковать созданные модели ИИ путем генерации артефактов в виде сервиса, библиотеки или программного модуля, а также дорабатывать отдельные логические блоки (программные модули) в рамках низкоуровневых сценариев, программируя на языке Python.

Дополнительно в состав платформы входит цифровой полигон для оценки качества систем ИИ. Он предназначен для автоматизации процедуры оценки характеристик качества систем ИИ, основанных на методах статистического обучения на данных, в соответствии с требованиями ГОСТ Р 59898-2021: определение точностных характеристик, устойчивости работы систем ИИ и границ их работоспособности. Полигон применяется для валидации систем ИИ на данных в процессе разработки, проверки релевантности обучения систем ИИ актуальным данным, оценки, сравнения и реинжиниринг работы сторонних систем ИИ на данных. Полигон может использоваться для моделей, построенных на различных видах данных (табличные, текст, изображения) в разнообразных отраслях. При этом он также является системой ИИ, т.к. использует автоматическое машинное обучение для построения псевдо-эталонов для моделей в виде “черного ящика”, бутстреп и генеративные методы синтеза данных для оценки устойчивости, а также реализует функции экспертной системы на основе анализа вычислительного графа модели ИИ для управления ее качеством. Полигон успешно апробирован на задачах ранжирования и анализа моделей распознавания медицинских изображений, а также оценки качества реализаций алгоритмов предиктивной аналитики технического оборудования. Применение полигона способствует увеличению покрытия систем ИИ тестами в 3–10 раз по сравнению с существующими практиками, что необходимо для обеспечения априори заданного качества разработки при ограниченных компетенциях персонала, отвечающего за их разработку и эксплуатацию.

2. ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ

В 2021 г. Центром в интересах индустриальных партнеров проведен ряд пилотных проектов, в рамках которых подтверждена и апробирована на практике применимость технологий ИИ для построения оптимальных план-графиков работ по освоению месторождения, оптимальных планов обслуживания скважин, гибких организационных структур для решения задач добычи и бурения, суррогатных моделей для предиктивной аналитики нефтегазового оборудования, а также создания стратегий управления человеческим капиталом предприятий. На их основе запущена разработка семейства прикладных интеллектуальных систем, в т.ч.:

система комплексного планирования обустройства месторождений на основе усвоения инженерного опыта и генеративных технологий (обеспечивающая повышение эффективности планов до 30%, сокращение трудозатрат по их созданию в 16–20 раз);

система генеративного дизайна пространственных объектов промышленной и логистической инфраструктуры в условиях Крайнего Севера (обеспечивающая работу с территориями пло-

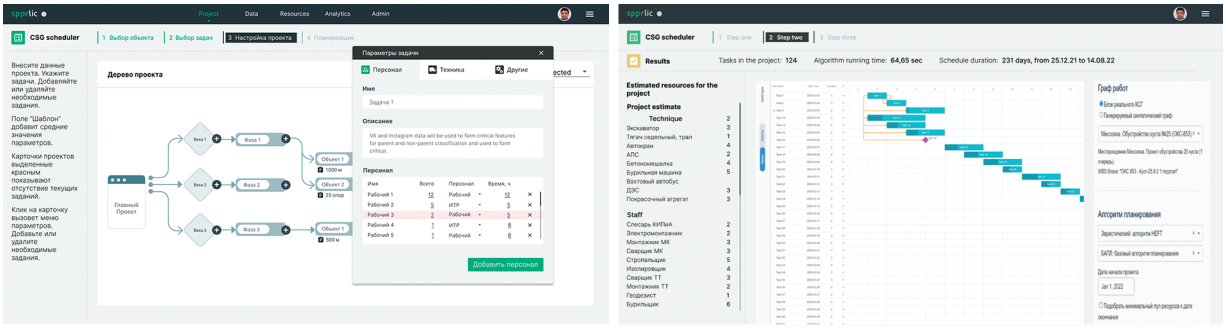

Рис. 4. Интерфейсы системы комплексного планирования обустройства месторождений на основе усвоения инженерного опыта и генеративных технологий.

щадью до 1000 Га и ускорение процесса выпуска проектной документации до 10 раз);

система планирования процессов добычи нефти и газа на основе интеллектуальной мультиагентной системы оптимизации технических решений (повышение качества планов на 20–25%, снижение времени подготовки планов в 5 раз).

Целесообразность создания системы комплексного планирования обустройства месторождений по заказу ПАО “Газпром нефть” связана с многомерностью исходной задачи, значимой неопределенностью и неполнотой данных на начальных этапах, что исключает возможность эффективно делать это вручную. Назначением системы является создание эффективных и устойчивых планов работ по комплексному освоению месторождений нефти и газа (включая создание инженерной инфраструктуры) в условиях неопределенности и вариативности ресурсов на этапах концептуального проектирования и экономического обоснования проекта. В основе системы заложены механизмы мультиагентного планирования на базе гибридных эволюционных алгоритмов ИИ, использующие суррогатные модели производительности операций на графах знаний, отражающих опыт строительства похожих объектов. Усвоение знаний в алгоритмы выполняется на основе автоматизированного анализа имеющейся документации по ранее реализованным проектам. В состав разработки входят: библиотека мультиагентного планирования производственных процессов в условиях неопределенности, вычислительный стенд для реализации различных алгоритмов планирования, ядро и интерфейс интеллектуальной системы планирования комплексного освоения месторождений. В совокупности система позволяет за разумное время обеспечить планирование до 60 тысяч взаимосвязанных операций, реализуемых бригадами общим числом до 500 единиц. Система может применяться как для создания общих планов освоения месторождений, так и их составных элементов, включая строительство кустов скважин, трубопроводов и ЛЭП. На рис. 4 приведены примеры интерфейсов системы, реализующей функции планирования.

Практические результаты работ Центра используются в деятельности его индустриальных партнеров, в первую очередь, ПАО “Газпром нефть” и ПАО “Роснефть”.

3. ЗАКЛЮЧЕНИЕ

Исследовательский Центр “Сильный ИИ в промышленности” на базе Университета ИТМО имеет четкую научно-практическую фокусировку, непосредственно связанную с обеспечением технологического суверенитета России в области ИИ. Он ориентирован на создание математического и программного инструментария, обеспечивающего демократизацию и тиражирование решений ИИ за счет автоматизации процесса разработки при сохранении общего уровня качества. Это позволяет не только получать работоспособные решения ИИ массовым специалистам, не имеющим высокой квалификации и навыков разработки, но и существенно (в разы) сократить время разработки. Как следствие, таким образом может быть преодолен кадровый барьер, затрудняющий массовое внедрение ИИ в отечественные отраслевые решения и инженерное ПО. Практическое подтверждение этой позиции основано на том, что прикладные решения, созданные сотрудниками Центра на основе разработанных им продуктов, системно попадают в победители и призеры различных хакатонов (например, “Цифровой прорыв”, хакатоны МЧС и Россельхозбанка). Таким образом, это демонстрирует, что даже начинающий специалист в области анализа данных, оснащенный необходимым инструментарием ИИ, способен выступать на равных с профессионалами, традиционно опирающимися на “ручной труд”.

Результаты работы Центра в 2022 г. представлены на 14 конференциях уровня A* в области ИИ и 7 журналах Q1 в области ИИ.

СПИСОК ЛИТЕРАТУРЫ

1. Национальная стратегия развития искусственного интеллекта на период до 2030 года / Указ Президента Российской Федерации 10 октября 2019 г. № 490 [http://www.kremlin.ru/acts/bank/44731]
2. Ассоциация “Искусственный интеллект в промышленности” [https://rusindustrial.ai/]
3. *Hvatov A.* Data-Driven Approach for the Floquet Propagator Inverse Problem Solution //ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022. С. 3813–3817.
4. *Sarafanov M., Pokrovskii V., Nikitin N.O.* Evolutionary Automated Machine Learning for Multi-Scale Decomposition and Forecasting of Sensor Time Series // 2022 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2022. С. 01–08.
5. *Borisov I.I. et al.* Reconfigurable Underactuated Adaptive Gripper Designed by Morphological Computation // 2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022. С. 1130–1136.
6. *Velichko A. et al.* Complex Paralinguistic Analysis of Speech: Predicting Gender, Emotions and Deception in a Hierarchical Framework // INTERSPEECH 2022. 2022. С. 4735–4739.

EDN: DRTYBC
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 41–49

РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ИССЛЕДОВАТЕЛЬСКИХ ЦЕНТРОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

УДК 004.8

ПЕРСПЕКТИВЫ ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ПРИКЛАДНЫХ БИЗНЕС-ЗАДАЧАХ

© 2022 г. В. В. Кондратьев[1,*], И. О. Пивоваров[1], Р. А. Горбачев[2], В. В. Матюхин[3], Д. А. Корнев[2], Д. А. Гаврилов[2], Е. А. Татаринова[2], В. Э. Буздин[2], И. М. Михайлов[2], О. А. Поткин[4]

Представлено академиком РАН С.С. Гончаровым
Поступило 28.10.2022 г.
После доработки 31.10.2022 г.
Принято к публикации 03.11.2022 г.

В статье рассмотрены основные научные результаты и достижения Исследовательского центра прикладных систем искусственного интеллекта Московского физико-технического института. Описаны достижения по ключевым научно-исследовательским направлениям – “Анализ естественного языка методами искусственного интеллекта” и “Искусственный интеллект для робототехники и управления беспилотными системами”. В частности, в рамках направления “Анализ естественного языка методами искусственного интеллекта” изучены мультимодальные и рекомендательные модели, показано, что перспективным с точки зрения объединения модальностей, оказывается текст: преобладающее большинство успешных мультимодальных продуктов так или иначе работает с модальностью текста, и часто именно к текстовому векторному пространству сводится векторное пространство иной модальности. В то же время очевидна непроработанность прикладного и продуктового применения мультимодальных моделей: способность сформулировать и решить конкретные бизнес-задачи с их помощью находится в начальном состоянии. В ходе выполнения работ по направлению “Искусственный интеллект для робототехники и управления беспилотными системами” выполняется разработка методико-алгоритмического обеспечения подсистемы управления роботизированного транспортного средства для построения карты и локализации на ней по камерам в реальном времени, позволяющего улучшить качество навигации беспилотного роботизированного транспортного средства при различных погодных условиях и разной окружающей обстановке (город, сельская местность, шоссе и др.). Кроме того, реализация проекта позволит упростить первичное прототипирование систем навигации, технического зрения и позиционирования беспилотных робототехнических комплексов и устройств за счет быстрого получения результатов обработки данных. Также проводятся работы по разработке бипедальных антропоморфных роботов: во всем мире в этом научно-техническом направлении активно ведутся исследования, публикуется большое количество научных работ, проводятся различные соревнования. Для обеспечения необходимой многофункциональности и гибкости для работы в человекоориентированной среде робот должен иметь конструкцию и механику, максимально приближенную к человеческим параметрам, и именно бипедальные антропоморфные роботы наиболее близко соответствуют этим требованиям. Разработана концепция конструкции робота, которая соответствует предъявляемым требованиям, начата работа по ее детальному проектированию для реализации реального прототипа робота. Также в статье описаны ключевые публикации по результатам работ в научных журналах, образовательные активности Центра.

Ключевые слова: vSLAM, диалоговые системы, бипедальные антропоморфные роботы, искусственный интеллект

[1] *Исследовательский центр прикладных систем искусственного интеллекта, Московский физико-технический институт, Москва, Россия*
[2] *Московский физико-технический институт, Москва, Россия*
[3] *Лаборатория продвинутой комбинаторики и сетевых приложений ФПМИ МФТИ, Москва, Россия*
[4] *Sber Automotive Technologies, ООО Сбер Автомотив Технологии, Москва, Россия*
*****E-mail: biggroup1@gmail.com*

DOI: 10.31857/S2686954322070104

1. ВВЕДЕНИЕ

Каждый год приносит нам новые удивительные технологии, которые меняют наш мир. Жизнь становится удобнее, быстрее, экономичнее. Но за каждой технологией стоит целая история кропотливых научных фундаментальных и прикладных исследований, разработки техноло-

гии и ее внедрения, создания продукта и выведения его на рынок. Как правило, прикладные исследования и разработка – это самый сложный участок пути, требующий высокой квалификации. Возможно поэтому в составе приоритетов ближайших лет ректор Московского физико-технического института Дмитрий Ливанов отметил "... практико-ориентированные исследования и разработки, инжиниринг для решения задач национального масштаба...".

В число таких приоритетных исследований, безусловно, входят исследования и разработки в сфере искусственного интеллекта. Для них в МФТИ в 2022 г. создан Исследовательский Центр прикладных систем искусственного интеллекта.

Программа ИЦ предусматривает создание на базе открытых платформенных решений программно-аппаратного обеспечения (отраслевых платформ) для разработки разговорных ассистентов, робототехнических систем и беспилотного автотранспорта с текстовыми, голосовыми, фото- и видеосервисами и их экспериментальных образцов на этой основе с элементами сильного искусственного интеллекта для применения в электронной коммерции и ряде других областей.

Деятельность Центра сфокусирована на исследованиях, разработках и коммерциализации по следующим передовым направлениям:

1. основное направление: "Анализ естественного языка методами искусственного интеллекта";

2. смежное направление: "Искусственный интеллект для робототехники и управления беспилотными системами".

Благодаря партнерству со Сбербанком, исследовательские команды Центра работают вместе с передовыми инженерными и исследовательскими командами, реализующими самые актуальные бизнес задачи, что позволяет поддерживать уровень технологий и коммерциализации Центра на самом передовом уровне.

К факторам, объединяющим оба направления в одной Программе, относятся:

1. новые математические методы и эффективные алгоритмы обучения (глубоких) нейронных сетей и другие методы и алгоритмы, реализация текстовых, голосовых, фото- и видеосервисов в программных и аппаратных решениях (обработка, анализ, интерпретация) являются технологической основой (ядром) всех областей исследования и применения ИИ.

2. разработанные ранее платформенные решения, успешно реализованные совместно со Сбербанком России в рамках проекта "iPavlov" в 2017–2020 гг. и признанные мировым сообществом, включая компанию Amazon, предоставившей грант МФТИ, как одному из победителей открытого международного конкурса, позволяют использовать технологическое ядро для разных применений, одновременно развивая и расширяя платформу.

3. появление потребности в перспективных системах управления робототехникой и беспилотным транспортом, в которых встроены голосовые ассистенты и вместе с ними составляют единое целое, что требует совместной разработки уже на этапе эскизного проектирования.

За год, прошедший с начала активной деятельности центра, было достигнуто многое, и в первую очередь – существенные научные результаты в рамках ключевых научно-исследовательских проектов Центра. В настоящей статье будет дан краткий обзор основных результатов первого года работы.

2. УПРАВЛЕНИЕ ДИАЛОГОМ, ПЕРСОНАЛИЗАЦИЯ, ЭМОЦИОНАЛЬНОСТЬ И МУЛЬТИМОДАЛЬНОСТЬ ДЛЯ РУССКОЯЗЫЧНЫХ ЦИФРОВЫХ АССИСТЕНТОВ

В диалоговых системах важна как корректная обработка пользовательской информации, так и ответ пользователю, что релевантно обоим существующим направлениям построения мультимодальных моделей. Перцептивное, фокусирующееся на том, как единообразно обработать данные разных модальностей, и генеративное – о том, как наоборот породить новое.

Несмотря на обилие научных работ, очевидна непроработанность прикладного и продуктового их применения: часто самыми успешными оказываются стартапы, (Stable Diffusion, Mid Journey, AI Dungeon) фактически оборачивающие в забавный интерфейс сырую модель. А вот способность сформулировать и решить конкретные бизнес-задачи с помощью мультимодальных моделей находится в начальном состоянии.

Вычислительный инференс (исполнение) мультимодальных моделей очень ресурсоемок:

• существующие мультимодальные модели обладают огромным количеством параметров (напр. ~80*109 параметров у модели Flamingo),

• популярная практика позднего смешивания модальностей требует наличия нескольких мономодальных систем, каждая из которых ресурсоемкая.

Кроме того, не существует архитектуры, оптимальной с точки зрения скорости, качества и многодоменности генерации текста. GAN-сети быстрые и порождают качественные примеры, VAE и модели потоков быстрые и легко обобщаются на новые области, а диффузионные модели порождают качественные примеры и легко обобщаются на новые области.

Перспективным с точки зрения объединения модальностей оказывается текст: преобладающее большинство успешных мультимодальных продуктов так или иначе работает с модальностью текста, и часто именно к текстовому векторному пространству сводится векторное пространство иной модальности.

В конце концов, и генерирующие, и наоборот воспринимающие данные модели отличаются огромной сложностью для человеческого анализа, а конкретные сильные и слабые стороны существующих разработок могут указать техники интерпретации мультимодальных моделей, они же могут предложить дальнейшие направления научного развития. Интересным представляется пробинг, техника, позволяющая установить специализацию отдельных участков нейронных сетей.

3. ПОСТРОЕНИЕ И УПРАВЛЕНИЕ ДИАЛОГОМ НА РУССКОМ ЯЗЫКЕ

Диалоговый менеджмент на основе правил имеет высокую сложность разработки в открытом домене. А использование только вероятностного подхода к управлению диалогом снижает интерпретируемость и контролируемость диалогового менеджмента. Поэтому в работе Лаборатории нейронных систем и глубокого обучения мы используем гибридный подход к управлению диалогом, комбинирующий подход на основе целей и вероятностные модели. Это позволяет определять и задавать интерпретируемое направление диалога. Гибридный подход наиболее распространен при создании диалоговых систем открытого домена, например, большинство участников Alexa Prize Challenge 3 и 4 использовали подход, комбинирующий применение вероятностных моделей и правил, регламентирующих поведение системы в целом и в специальных случаях. Это позволило участникам ввести контролируемое управление диалогом при сохранении обобщенности на открытый домен.

Прагматический и дискурсивный анализ в управлении диалогом основывается на теории диалоговых актов и теории риторических структур, однако они недостаточно изучены на русскоязычном материале. Тем не менее диалоговые акты достаточно активно используются иностранными компаниями в разработке чат-ботов. Например, в Alexa используется модифицированная таксономия диалоговых актов для интерпретации действий, совершаемых в каждой реплике как пользователем, так и самой системой. Речевые акты используются и в XiaoIce для классификации намерений пользователя. Команды конкурса Alexa Prize SocialBot Grand Challenge, в рамках которого необходимо создать диалоговую систему открытого домена, также использовали для управления разговором диалоговые акты. Команды-победители “Slugbot”, “Gunrock” и “Alquist”, а также команда “Iris” обучали собственные модели для классификации абстрактных намерений. В результате работы над диалоговыми актами команда “Gunrock” разработала новую таксономию MIDAS, заимствующую принципы предыдущих схем аннотации, но адаптированную под современные задачи в области диалогового менеджмента. Для управления ходом диалога голосового ассистента Google также была разработана схема аннотации диалоговыми актами, которая представляет собой несколько групп абстрактных намерений со спецификацией речевых действий говорящего в определенный момент диалога.

4. ОПРЕДЕЛЕНИЕ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ ДИАЛОГА И ФОРМИРОВАНИЯ ЭМОЦИОНАЛЬНЫХ ОТВЕТОВ НА РУССКОМ ЯЗЫКЕ

Область определения эмоциональной окраски диалога является достаточно развитой. Широко известны подходы, в которых используются лингвистические признаки, однако они теряют свою актуальность. Наиболее высокое качество показывают методы с использованием нейронных сетей. Анализ этих методов показал, что необходимо учитывать особенности диалоговых данных с помощью специального моделирования контекста двух участников. Первые работы также показали важность данных от различных модальностей и то, что модальности имеют разный вклад в точность определения эмоциональной окраски.

Наиболее перспективные направления исследований – 1) моделирование эмоционального поведения на основе психологических черт Big Five, а также 2) задание личности с помощью персональных фактов. Модели, основанные на правилах и сценариях, оказываются слишком ограниченными. Системы, вдохновленные биологическими процессами человека, слишком сложны в реализации, а их эффективность еще не доказана.

Существующие российские исследования предлагают применять правила и заготовленные сценарии или же генеративные модели с использованием векторных представлений диалога. Разработки ведутся в направлении как генерации эмоционального текста, так и синтеза речи с выражением эмоций. Для создания собственного эмоционального генеративного инструмента также следует обратиться к разработкам для английского языка, чтобы почерпнуть современные идеи и подходы. Большая часть недавних зарубежных разработок применяет генеративные модели на основе трансформеров, а также оснащает их дополнительными знаниями о мире и текущем диалоге. Другим интересным и перспективным

направлением для экспериментов в этой области является применение трансфера стиля, где эмоциональная окраска реплик воспринимается как стиль текста.

5. ВОЗМОЖНЫЕ ОБЛАСТИ ПРИМЕНЕНИЯ

Одной из самых растущих областей является электронная коммерция, в которой все общение с клиентом происходит в онлайне, в том числе посредством чатов. Рекомендательные системы электронных коммерций должны уметь справляться с постоянным добавлением новых данных, у которых не обязательно при этом имеется подробное описание. Так, в электронной коммерции используют модели для классификации изображений и текста, чтобы самостоятельно извлекать недостающую информацию о товарах. Также рекомендательные системы должны учитывать информацию о совместимости комплементарных товаров с покупаемым, и то, уместно ли рекомендовать товар, принадлежащий той же категории, что и товар, уже купленный пользователем (чтобы не рекомендовать пользователю купить еще один телефон, но при этом рекомендовать купить еще одну книгу), и множество других нюансов, способных повлиять на качество пользовательского опыта. Некоторые системы используют модели глубокого обучения, которые позволяют также учитывать последовательность действий пользователей, так как это тоже является важной информацией, способной повлиять на качество рекомендаций.

Модели для рекомендательных систем беспрерывно совершенствуются за счет роста данных. Кроме того, исследователи постоянно предлагают новые признаки для анализа предпочтений. Так, это могут быть специфичные для сферы особенности контента, например, стилистические черты фильма или акустические признаки музыки. В случае с фильмами можно также анализировать по отдельности каждую модальность, так как кому-то при выборе фильма важнее визуальная составляющая, кому-то – музыкальное сопровождение и т.д. Но можно дополнять и данные о пользователе — информация о его личности, психотипе, а также текущем настроении значительно влияет на его потребности. Больше всего от эмоционального состояния, настроения и контекста зависит выбор музыки. Музыкальные предпочтения зависят от того, чем занят пользователь, погоды, дня недели, времени дня, местонахождения, его окружения. Поэтому важно уметь определять контекст, в котором находится пользователь. Для этого можно использовать информацию с его устройств, социальных сетей, или же из разговоров с ним (в случае голосовых помощников).

Область электронной коммерции представляется огромным полем для внедрения мультимодальных моделей.

6. РАЗРАБОТКА СИСТЕМЫ УПРАВЛЕНИЯ БЕСПИЛОТНОГО РОБОТИЗИРОВАННОГО ТРАНСПОРТНОГО СРЕДСТВА

Транспорт был и остается одной из самых перспективных отраслей экономики. А применение технологий искусственного интеллекта позволяет создавать системы управления беспилотными транспортными средствами.

Целью работы является разработка методико-алгоритмического обеспечения подсистемы управления роботизированного транспортного средства для построения карты и локализации на ней по камерам в реальном времени, а также создание научно-технического задела в области разработки интеллектуальных систем управления беспилотными робототехническими устройствами.

Объектом исследования является подсистема управления роботизированного транспортного средства для построения карты и локализации на ней по камерам в реальном времени (Visual Simultaneous Localization and Mapping). В результате разрабатывается макет специального программного обеспечения vSLAM (СПО vSLAM), верификация которого будет осуществляться на разрабатываемом макете программно-аппаратного комплекса локального позиционирования (ПАК ЛП), состоящем из макета блока оптико-электронного, макета блока вычислителя алгоритма vSLAM, комплекта вспомогательных инструментов и приспособлений.

В рамках разработки СПО vSLAM на первом этапе проекта выполнен аналитический обзор современного состояния исследований в области визуальной локализации автономной беспилотной системы и построения карты заранее неизвестной местности в режиме реального времени (vSLAM). Современные vSLAM-решения можно разделить на две основные группы: прямые и непрямые. В прямых vSLAM напрямую используются яркости пикселей изображений, а оценки позы камеры получаются путем минимизации фотометрической ошибки между соответствующими пикселями изображений. В непрямых vSLAM сначала извлекаются признаки изображений, а затем признаки описываются и сопоставляются для оценки позы путем минимизации ошибки перепроецирования. Одним из ключевых шагов vSLAM-решений, реализующих стереорежим, является поиск соответствующих точек на изображениях сцены, полученных с различных ракурсов (стереоотождествление). Решение задачи стереоотождествления и получение значений несоответствия для точек изображений позволяет

далее (после вычисления элементов ориентирования) получить функцию дальности до видимого рельефа наблюдаемой сцены. Для решения задачи стереоотождествления вместо традиционных методов могут быть использованы методы стереоотождествления, основанные на использовании глубоких нейронных сетей, например, методы CoEx и HSMNet, которые обеспечивают работу в режиме реального времени. Для решения задачи семантической сегментации изображений в режиме реального времени наиболее целесообразным представляется использование нейронной сети DDRNet-23 или какой-либо из ее модификаций, характеризующихся высокой производительностью (например, модификации DDRNet-23_Bayer). Среди рассмотренных нами нейронных сетей, используемых для семантической сегментации в режиме реального времени, наилучшую производительность (скорость вычислений), демонстрирует сеть STDC1-50, для которой FPS = 250.

Основными задачами второго этапа являлись:

• исследование перспективных открытых vSLAM-решений и программных решений в области визуальной одометрии, направленных на выявление алгоритмических решений, которые будут положены в основу разрабатываемого макета специального программного обеспечения vSLAM;

• анализ ограничений на отобранные алгоритмы применительно к аппаратным платформам, на которых они реализуются;

• разработка промежуточных версий алгоритмов построения карты местности и локализации на ней по камерам.

В результате проведенных исследований выявлены наиболее предпочтительные, которые предполагается заложить в основу разрабатываемого vSLAM-решения. Во-первых, алгоритм Stella-SLAM, основанный на ORB-SLAM. Данная реализация позволяет использовать камеры различных типов, загружать и использовать для локализации ранее созданные карты, а также превосходит по точности определения траектории другие решения, в основе которых лежит ORB-SLAM. Во-вторых, алгоритм DROID-SLAM, основанный на методе DSO. Данный метод целиком основан на глубоком обучении, строит плотные 3D-карты окружающей среды, превосходит по точности определения траектории другие решения, в основе которых лежит DSO, реализован на языке Phyton, на котором реализованы также отобранные нами методы стереоотождествления. Исследованы возможности интеграции в разрабатываемое vSLAM-решение подходов, реализованных в решениях, в основу которых положен ORB-SLAM либо DSO. Получены предварительные результаты по интеграции отобранного real-time-метода стереоотождествления AANet в разрабатываемое vSLAM-решение.

В результате работ разработана промежуточная версия алгоритмов построения карт местности и локализации на ней по камерам, на основе которых будет разработан макет СПО vSLAM.

В рамках разработки макета программно-аппаратного комплекса локального позиционирования на первом этапе разработана концепция программного обеспечения для управления беспилотным роботизированным транспортным средством, выполнен аналитический обзор текущих разработок и современного уровня техники в области создания многофункциональных оптико-электронных систем.

На втором этапе разработан и создан макет многофункциональной оптико-электронной системы. Разрабатываемая многофункциональная оптоэлектронная система предназначена для управления беспилотным транспортным средством с помощью оперативного анализа окружающей обстановки, формирования изображения для построения карты местности, определения параметров курса, локализации и навигации беспилотного транспортного средства в режиме реального времени. Внешний вид многофункциональной оптико-электронной системы представлен на рисунке.

Внедрение СПО vSLAM в перспективе позволит улучшить качество навигации беспилотного роботизированного транспортного средства при различных погодных условиях и разной окружающей обстановке (город, сельская местность, шоссе и др.). Реализация проекта позволит упростить первичное прототипирование систем навигации, технического зрения и позиционирования беспилотных робототехнических комплексов и устройств за счет быстрого получения результатов обработки данных. При этом использование гибридных данных нейросетевого анализа и данных, получаемых от сенсоров, позволит вывести эффективность подобных систем на новый качественный уровень.

Зарегистрированы результаты интеллектуальной деятельности: программы для ЭВМ “Программа обработки визуальной информации для беспилотного транспортного средства”, свидетельство о регистрации № 2022610215 от 27.12.2021 г., “Программное обеспечение блока вычислителя алгоритма vSLAM”, свидетельство о регистрации № 022669709 от 18.10.2022 г., полезная модель “Многофункциональное оптико-электронное устройство кругового обзора для управления движением беспилотного транспортного средства” № 210565 от 27.12.2021 г.

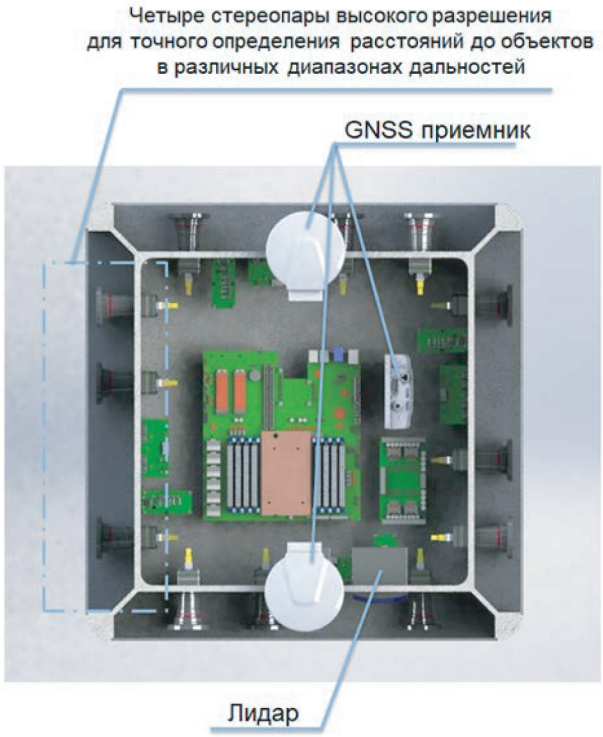

Рис. 1

7. ИССЛЕДОВАНИЯ В ОБЛАСТИ АНТРОПОМОРФНОЙ РОБОТОТЕХНИКИ ДЛЯ СОЗДАНИЯ БИПЕДАЛЬНЫХ АНТРОПОМОРФНЫХ РОБОТОВ, РАЗРАБОТКА И ИСПЫТАНИЯ ИССЛЕДОВАТЕЛЬСКОЙ ПЛАТФОРМЫ ДЛЯ ОТРАБОТКИ ЭКСПЕРИМЕНТОВ ПО ИСПОЛЬЗОВАНИЮ АНТРОПОМОРФНЫХ РОБОТОВ НА РАЗЛИЧНЫХ УЧАСТКАХ ЦЕПОЧКИ СОЗДАНИЯ ЦЕННОСТИ В ЭЛЕКТРОННОЙ КОММЕРЦИИ

В настоящее время разработка бипедальных антропоморфных роботов является приоритетным научно-техническим направлением, в котором активно ведутся исследования, публикуется большое количество научных работ, проводятся различные соревнования (RoboCup, Робофест). Для обеспечения необходимой многофункциональности и гибкости робот должен иметь конструкцию и механику, максимально приближенную к человеческим параметрам. Именно бипедальные антропоморфные роботы наиболее близко соответствуют этим требованиям. Вследствие этого данный вид роботов является практически единственным универсальным типом робототехнических систем, которые хорошо способны выполнять многочисленные задачи.

Основной целью данного проекта являются исследование и разработка физически реализуемой цифровой модели антропоморфного робота и его реального прототипа, его комплектующих, а также программного обеспечения для обеспечения требуемой функциональности робота и реализация симуляторов реального мира для обучения и отработки алгоритмов управления движениями робота, тестирования системы технического зрения. Разрабатываемый прототип робота должен функционировать не только для работы в лабораторных условиях, но в условиях переменчивой окружающей среды, например, передвигаться не только по прямой поверхности, но и по поверхности со средним уровнем неровностей без падений, в том числе на улице.

Для реализации проекта была выработана концепция разработки, заключающаяся в следующем: сначала передовые алгоритмы движения, а только потом способная реализовать эти передовые алгоритмы конструкция. В связи с этим первоначальным этапом проекта было проведение анализа уже существующих разработок и современного уровня техники, в данном случае, в области алгоритмов и систем управления антропоморфными роботами, существующих симуляторов для разработки и исследования алгоритмов управления, математических и реальных моделей роботов, методов построения оптимальных траекторий движений роботов, а также комплектующих и материалов для изготовления реального прототипа.

Было проведено исследование современных трендов к подходам как проектирования, так и управления самыми современными роботами. В первую очередь были рассмотрены существующие решения в области разработки бипедальных антропоморфных роботов, такие как, например, ASIMO, Atlas, Cassie, Digit, LOLA. Для изучения их особенностей, не имея реальных роботов, используются их виртуальные модели, которые в свою очередь могут быть запущены в одном из симуляторов. Были исследованы основные симуляторы Webots, V-REP, Gazebo и MuJoCo, а также движки – ODE, Bullet, DART и MuJoCo. Проведя обширное исследование, были выделены ключевые моменты конструкции и алгоритмов, которые позволяют двуногому роботу быть устойчивым и эффективным:

• Практические успехи в области реализации таких сложных движений, как быстрая ходьба, бег, прыжки, были достигнуты за счет ряда конструктивных особенностей, например, принципа “облегчения ног”;

• Важным с точки зрения повышения энергоэффективности при разработки новых роботов является также использование модели Spring Loaded Inverted Pendulum (модель перевернутого маятника с пружиной), позволяющей строить малозатратные движения;

• Принцип “программной” реализации упругости конструкции робота за счет активной схемы управления приводами, реализующей прокручивание его сочленений при воздействии внешней силы для минимизации рисков поломки робота при значительных внешних воздействиях;

• Применение редукторов с малым передаточным числом для обеспечения гибкости и минимизации вероятности поломок при падениях и выполнении сложных движений (однако данное решение способствует повышению энергозатрат).

Для управления движениями антропоморфного бипедального робота были рассмотрены существующие подходы: традиционные (PID, MPC control, Robust control) и интеллектуальные (Machine Learning, Deep Learning, Fuzzy control). Был реализован ряд алгоритмов с использованием Reinforcement Learning, рекуррентных нейронных сетей, Feed Forward Torque Control и других подходов. Данные алгоритмы успешно прошли тестирование в симуляторе для разных типов роботов, как бипедальных, так и квадропедальных. Для апробации их реализации в реальности была использована модель квадропедального робота-собаки. Данные алгоритмы показали эффективность в обеспечении стабильности ходьбы в различных направлениях и поворотах, сохранении устойчивости при ходьбе по неровной поверхности и воздействии на робота внешних сил. Результатом данных исследований является набор алгоритмов управления для робота-собаки, который позволяет ей стабильно передвигаться, и который работает как в среде обучения и верификационном симуляторе, так и в реальности.

По итогам текущих исследований были выбраны основные направления исследований и разработок алгоритмов управления роботом, подходов к реализации его виртуальной и реальной модели. Был разработан ряд испытательных стендов для отработки реализуемых алгоритмов и исследования конструктивных особенностей составляющих компонентов робота. Были исследованы алгоритмы по управлению различными типами роботов в симуляционной среде, а также апробация некоторых из них на реальной модели робота-собаки. Итогом работ на текущий момент является разработанная концепция конструкции робота, которая соответствует предъявляемым требованиям, вследствие чего начата работа по ее детальному проектированию для реализации реального прототипа робота.

8. ЗНАЧИМЫЕ НАУЧНЫЕ РЕЗУЛЬТАТЫ, ПРЕДСТАВЛЕННЫЕ НА КОНФЕРЕНЦИЯХ И В ВЕДУЩИХ НАУЧНЫХ ЖУРНАЛАХ

В рамках деятельности ИЦ прикладных систем искусственного интеллекта важное место занимает деятельность по публикации полученных в ходе исследований научных результатов. Сотрудники Центра публикуют свои статьи в ведущих мировых журналах уровня Q1 (первый квартиль рецензируемых журналов), таких как Euro Journal on Computational Optimization, Optimization Methods and Software и другие).

Еще более значимым результатом является публикация своих работ на международных конференциях уровня A*, таких как ICML, AISTATS и NeurIPS. В 2022 г. ожидается 5 публикаций на конференциях уровня A* (было запланировано 4).

В 2022 г. сотрудниками центра были опубликованы следующие работы:

1. Decentralized personalized federated learning: Lower bounds and optimal algorithm for all personalization modes (Borodich, Beznosikov) – журнал Q1 Scopus Euro Journal on Computational Optimization

Работа сосредоточена на проблеме персонализации в федеративном обучении – разновидности распределенного машинного обучения, где предполагается, что вычислительные агенты – это пользовательские устройства (например, смартфоны, планшеты, ноутбуки, персональные компьютеры).

В работе исследуется формулировка децентрализованного персонализированного федеративного обучения, а также доказываются нижние границы сложности на число коммуникаций и локальных вычислений, разрабатывается несколько алгоритмов, способных достичь нижних границ.

2. Extragradient Method: O (1/K) Last-Iterate Convergence for Monotone Variational Inequalities and Connections with Cocoercivity (Gorbunov) – конференция A* AISTATS 2022

В данной работе удалось впервые вывести O(1/K) оценку на сходимость экстраградиентного метода для последней точки в терминах квадрата нормы оператора (и, соответственно, O(1/sqrt{K}) оценку для Gap-функции).

Кроме того, ключевой особенностью разработанного в статье анализа является тот факт, что основные части доказательства отличия между экстраградиентным методом и методом Попова получены частично при помощи компьютера, а именно, при помощи техники Performance Estimation Problem (PEP) (Taylor et al., 2017; Ryu et al., 2020). Данная техника получения доказательств является не очень популярной в виду своей нетривиальности. Однако сам по себе подход имеет

огромный потенциал, что было продемонстрировано в данной работе.

3. Stochastic Extragradient: General Analysis and Improved Rates (Gorbunov) – конференция A* AISTATS 2022

В работе был разработан новый теоретический фреймворк для анализа метода SEG (Stochastic Extragradient method).

Сделаны точные оценки на скорость сходимости в известных частных случаях: наш анализ дает либо наилучшие известные оценки для известных частных случаев (например, для EG и I-SEG в частном случае, когда параметр delta в равномерной оценке дисперсии равен нулю). Рассмотрены новые методы с хорошими оценками. Разработан новый способ выбора шагов в SEG.

4. Last-Iterate Convergence of Optimistic Gradient Method for Monotone Variational Inequalities (Gorbunov) – конференция A* NIPS 2022

В этой работе предложен первый (неасимптотический) анализ сходимости PEG (Past Extragradient method) для последней точки, закрывающий тем самым важный открытый вопрос в литературе по экстраградиентным методам. В безусловном случае доказано, что PEG сходится со скоростью O(1/N) для последней точки в терминах квадрата нормы оператора. Для условных задач получен аналогичный результат для квадрата нормы невязки между точками на двух последних итерациях (естественное обобщение критерия на условный случай). В частности при помощи техники PEP (Performance Estimation Problem) найдены потенциальные функции для PEG для вариационных неравенств с ограничениями и без, из которых вытекает упомянутый выше результат. Кроме того, продемонстрирована нетривиальность данного вопроса; показано, что ключевое для анализа неравенство, выполненное для EG, может нарушаться для PEG.

5. Secure Distributed Training at Scale (Gorbunov) – конференция A* ICML 2022

В работе предложен новый протокол для децентрализованного обучения с устойчивостью к Византийским атакам на датасетах, доступных всем участникам. Дополнительные коммуникационные затраты предложенного протокола не зависят от количества параметров модели. Также в работе предложен математически строгий анализ нового протокола и доказываются оценки на скорость сходимости для выпуклых и невыпуклых задач с Византийскими рабочими. Кроме того, получаются ускоренные сходимости для одной и той же задачи при реалистичных предположениях о градиентах модели.

Предожена эвристика для сопротивления Sybil attacks со стороны вычислительно ограниченных злонамеренных рабочих, позволяющие принимать новых ненадежных рабочих по ходу обучения. Проверяется эффективность алгоритма в контролируемых экспериментах и реальных крупномасштабных прогонах обучения.

6. Clipped Stochastic Methods for Variational Inequalities with Heavy-Tailed Noise (Danilova) – конференция A* NIPS 2022

В работе было обнаружено, что шум в стохастических градиентах, возникающих при обучении популярных генеративно-состязательных моделей (GAN), имеет тяжелые (не суб-гауссовские) хвосты распределения. Это послужило основной мотивацией к исследованию сходимости стохастических методов для решения вариационных неравенств с большой вероятностью.

В частности, в работе были предложены два новых метода – clipped-SEG и clipped-SGDA. Оба метода используют популярный трюк в глубинном обучении, а именно, градиентный клиппинг. Были доказаны первые результаты о сходимости с большой вероятностью стохастических методов для решения монотонных вариационных неравенств без предположений о легкости хвостов распределения шума. Более того, дополнительно рассмотрены 5 классов задач, допускающих немонотонные операторы F. Для указанных классов задач полученные результаты не имеют аналогов даже в предположении легких хвостов распределения шума.

7. Accelerated variance-reduced methods for saddle-point problems (Borodich) – журнал Q1 Scopus Euro Journal on Computational Optimization

В данной работе предлагается ускоренный алгоритм с оракулом первого порядка для задач в виде суммы, который использует технику уменьшением дисперсии. В работе доказывается, что сложность данного алгоритма почти оптимальная, т.е. совпадает с нижними оценками с точностью до логарифмических факторов. Важно отметить, что алгоритм гарантирует необходимую точность с высокой вероятностью, а не в среднем. Насколько известно, эти алгоритмы являются первыми оптимальными для данной задачи. Таким образом, алгоритм позволяет понять, что нижние оценки достижимы.

9. ОБРАЗОВАТЕЛЬНЫЕ АКТИВНОСТИ

Центр поддерживает “Всероссийский учебный фестиваль по искусственному интеллекту и программированию “RuCode Festival” (далее – RuCode Festival). RuCode Festival управляется и проводится созданным консорциумом из 16 ведущих вузов страны (ДВФУ, ЗабГУ, НГУ, ННГУ, БФУ им. Канта, ТГУ, ИжГТУ, ПетрГУ, СГУ, СФУ, УрФУ, Университет “Иннополис”, КГУ, ТИУ, ИТ-университет), общественных организаций и лидирующих IT-компаний во главе с МФТИ.

В 2020 г. RuCode Festival получил старт при поддержке Фонда президентских грантов в дистанционном формате. Сейчас RuCode Festival проходит дважды в год в онлайн и оффлан режиме при поддержке Минобрнауки РФ и благодаря спонсорской поддержке передовых российских IT-компаний. Организаторами фестиваля, наряду с МФТИ, выступают ведущие вузы России, общественные организации, технопарки и кванториумы. Индустриальные партнеры фестиваля: Яндекс, Сбер, 1С, Газпромбанк, Роскосмос и др.

В программе – онлайн-курсы, интенсивы, чемпионат по алгоритмическому программированию и искусственному интеллекту. С 2022 г. фестиваль является площадкой для реализации инновационного проекта “Система интенсивной подготовки ИТ-кадров для быстрого и эффективного устранения кадрового дефицита на рынке труда”. Фестиваль находится на стыке науки и искусства, охватывает и просвещает широкий круг как начинающих, так и опытных специалистов, которые заинтересованы развитием IT-технологий, обладает научной визуализацией и потенциалом трансдисциплинарности, предоставляет возможности каждому быть вовлеченным в мир высоких технологий.

В рамках фестиваля разрабатываются дополнительные профессиональные программы повышения квалификации в области искусственного интеллекта. По заданию Центра разработаны программы “Глубокое обучение в NLP” и “NLP: создание вопросно-ответных систем”.

10. ЗАКЛЮЧЕНИЕ

Исследовательский Центр прикладных систем искусственного интеллекта активно ведет научные исследования в передовых областях ИИ, публикуя свои результаты в ведущих мировых журналах и на конференциях. Наши команды участвуют в международных соревнованиях, занимая призовые места, например в октябре 2022 г. команда роботов Starkit выиграла открытый чемпионат Бразилии и Латинской Америки по робофутболу.

Перед нашими исследователями ставятся самые высокие планки по качеству и результативности исследований. Необходимость внедрения этих результатов в реальные бизнес-процессы делает работу на таком высоком уровне сложной, но потрясающе интересной! Здесь, на передовом крае науки, создаются новые технологии, которые могут потом применяться в промышленности и сельском хозяйстве, медицине и образовании, для решения коммерческих или государственных задач. Мы приглашаем партнеров и заказчиков, которым интересно внедрение передовых технологий искусственного интеллекта – давайте создавать будущее вместе!

EDN: TUHJHE
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 50–69

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ДИНАМИКА И ЛАНДШАФТ ФУНКЦИИ ПОТЕРЬ ДЛЯ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ПРИ ОБУЧЕНИИ С КВАДРАТИЧНОЙ ФУНКЦИЕЙ ПОТЕРЬ

© 2022 г. М. С. Находнов[1], М. С. Кодрян[1], Е. М. Лобачева[2], Д. С. Ветров[1,2,*]

Представлено академиком РАН А.А. Шананиным
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

Знание свойств геометрии функции потерь позволяет успешно объяснять поведение нейронных сетей, динамику их обучения, взаимосвязь получаемых решений и гиперпараметров, таких как способ регуляризации, архитектура нейронной сети или расписание темпа обучения. В данной работе изучаются динамика обучения и поверхность стандартной кросс-энтропийной и популярной в последнее время квадратичной функций потерь для масштабно инвариантных сетей с нормализацией. Для устранения симметрий был произведен переход к оптимизации на сфере, который позволил обнаружить три фазы обучения в зависимости от размера шага обучения на сфере, обладающие принципиально разными свойствами, – фазу сходимости, фазу хаотического равновесия и фазу дестабилизированного обучения. Данные фазы наблюдаются для обеих исследованных функций потерь, однако при обучении с квадратичной функцией потерь нужны бóльшие сети и более долгое обучение для перехода в фазу сходимости.

Ключевые слова: масштабная инвариантность, батч-нормализация, обучение нейронных сетей, оптимизация, квадратичная функция потерь

DOI: 10.31857/S2686954322070189

1. ВВЕДЕНИЕ

Одной из основных задач, успешно решаемых с помощью глубоких нейронных сетей, является многоклассовая классификация. Важной составляющей решения задачи является правильный выбор функции потерь. В большинстве случаев, в задаче классификации ограничиваются использованием кросс-энтропийной функции потерь. При этом данный выбор не является единственно возможным и есть свидетельства, что альтернативные варианты функции потерь могут приводить к качеству не хуже на большом классе задач и архитектур [1].

С другой стороны, решения современных задач в машинном обучении широко используют эмпирические приемы для получения наилучших результатов. Например, выбор оптимизатора или расписания темпа обучения долгое время основывался на эмпирических результатах для конкретных архитектур [2]. Исследование ландшафта функции потерь позволило как обосновать такие инженерные техники, как батч-нормализация [3], соединения быстрого доступа (Residual Connections) [4], так и предложить новые способы улучшения генерализации моделей [5].

Известно, что ширина оптимума имеет сильную корреляцию генерализацией модели [6]. Поэтому при анализе ландшафта функции потерь ширина в текущей точке и динамика ее изменения в процессе обучения вызывают основной интерес.

Исследование поверхности функции потерь затруднено, так как оптимизация происходит в многомерном пространстве, а функция, задаваемая глубокой нейронной сетью, является невыпуклой. Наличие в сетях слоев нормализации еще сильнее усложняет анализ за счет появления масштабно инвариантных симметрий. Переход к оптимизации на сфере позволил избавиться от таких симметрий. При варьировании разрешающей способности на сфере было обнаружено три режима обучения нейронной сети, отвечающих различным областям поверхности функции потерь. В данной работе был проведен анализ этих фаз с точки зрения генерализации моделей и ширины

[1] *Институт искусственного интеллекта AIRI, Москва, Россия*
[2] *Национальный исследовательский университет "Высшая школа экономики", Москва, Россия*
*****E-mail: dvetrov@hse.ru*

получаемых решений. Также было произведено сравнение различных функций потерь с точки зрения влияния на обнаруженные фазы и динамику обучения.

2. ДИЗАЙН ЭКСПЕРИМЕНТОВ

2.1. Симметрии в нейронных сетях

Исследование нейронных сетей осложняется значительным уровнем избыточности параметров [7] и наличием внутренних симметрий. Простейшими примерами таких симметрий являются согласованная перестановка нейронов в последовательных слоях и согласованное масштабирование весов в сетях с функцией активации ReLU [8]. Такие преобразования обычно оставляют сеть функционально неизменной, хотя в пространстве весов модель может существенно измениться. Другой важной симметрией является масштабная инвариантность в сетях с батч-нормализацией. Использование батч-нормализации после сверточного слоя приводит к тому, что умножение весов, предшествующих слою нормализации, не меняет сеть, как функцию от своего входа. Рассмотрим нейронную сеть $f(\theta)$ с весами $\theta \in R^d$. Параметры, умножение которых на произвольный положительный коэффициент α не меняет функциональный вид сети, будем называть масштабно инвариантными (Scale-Invariant, SI). В таком случае, для SI параметров верно:

$$f(\alpha\theta) = f(\theta),\ \forall\ \theta, \alpha > 0, \tag{1}$$

$$\nabla f(\alpha\theta) = \frac{1}{\alpha}\nabla f(\theta). \tag{2}$$

Наличие SI параметров в сети приводит к неоднозначности в определении ширины оптимума, так как в зависимости от нормировки весов функционально одинаковые модели будут иметь различные градиенты и вторые производные в соответствии с уравнением (2). Для того, чтобы избавиться от инвариантности, предлагается рассматривать сети, состоящие только из масштабно инвариантных параметров на сфере фиксированного радиуса. Для определенности будем по умолчанию считать, что сеть задана на единичной сфере, т.е. $\theta \in B\_\{1\} = \{\theta \mid \|\theta\| = 1\}$. Темп обучения сети с полностью масштабно инвариантными параметрами (Fully Scale-Invariant, FSI) на сфере единичного радиуса будем называть эффективным темпом обучения (effective learning rate, *elt*). Обучение такой модели градиентными методами может приводить к тому, что после очередного шага норма весов станет отличной от 1. В таком случае предлагается применять нормировку весов на очередном шаге. Отличие данного подхода от Римановой оптимизации на сфере приведено в Приложении 3.

Стоит отметить, что фиксация общей нормы параметров устраняет только часть симметрии в нейронной сети – отдельные фильтры сверхточных слоев остаются инвариантными к перенормировке.

Для исследования эффектов, связанных с динамикой оптимизации, вдоль поверхности функции потерь необходимо выбрать такую постановку эксперимента, где особенности обучения будут изолированы от сторонних эффектов, таких как переобучение, симметрии внутри нейронной сети. Для этого предлагается рассматривать следующие контролируемые, но в тоже время приближенные к реальным, условия для обучения. Во-первых, в качестве обучающей выборки будет рассматриваться набор данных CIFAR10 без использования аугментации. Во-вторых, в качестве архитектуры нейронной сети используется сверточная нейронная сеть ConvNet с батч-нормализацией из полностью масштабно инвариантных параметров. Переход к FSI архитектуре производится путем фиксации аффинных слоев батч-нормализации и фиксации последнего линейного слоя сети. Подробное описание архитектуры находится в Приложении 1. Известно, что такие ограничения на параметры не влияют на итоговое качество модели на тестовой выборке. В-третьих, обучение происходит с помощью стохастического градиентного спуска (Stochastic gradient descent, SGD) без использования инерции и $L\_\{2\}$ регуляризации. Все модели обучаются из одного и того же начального приближения с одинаковым порядком батчей в процессе оптимизации.

В качестве функции потерь рассматриваются две альтернативы. Стандартным выбором для оптимизируемой ошибки для задачи C-классовой классификации является кросс-энтропия:

$$L(\hat{y}, y) = -\log \frac{\exp \widehat{y_y}}{\sum_{i=1}^{C} \exp \widehat{y_i}}, \tag{3}$$

где $y, \hat{y} \in R^C$ – метка правильного класса и выход сети соответственно.

В качестве альтернативы можно рассмотреть квадратичную функцию потерь:

$$L(\hat{y}, y) = \sum_{i=1}^{C} (\widehat{y_i} - 1[y = i])^2. \tag{4}$$

Существующие работы не дают четкого ответа о том, какая из этих функций предпочтительнее. С одной стороны, долгое время считалось, что квадратичная функция потерь медленнее сходится и приводит к худшему качеству на тестовой выборке при использовании стохастического градиентного спуска [9, 10].

Однако более новые работы показывают противоположную ситуацию – подробный анализ [1]

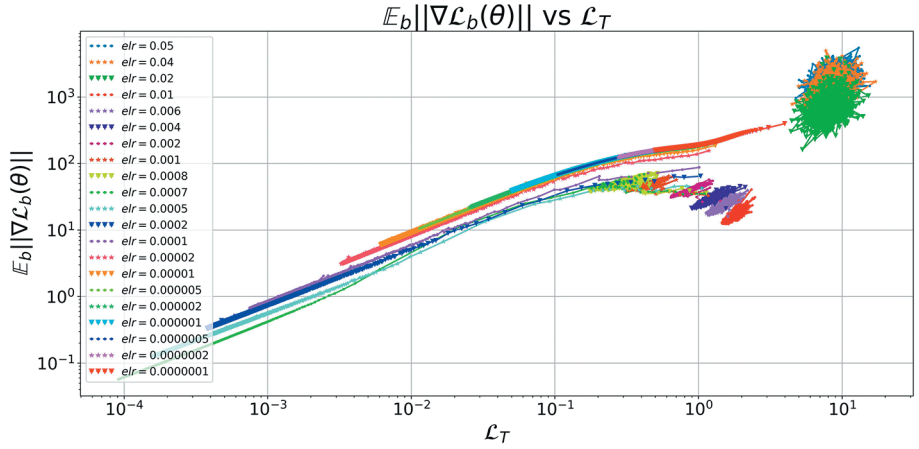

Рис. 1. Фазовая диаграмма для кривизны и кросс-энтропийной функции потерь для различных *elr* для сети ConvNet. Наблюдается три принципиальных режима поведения траекторий.

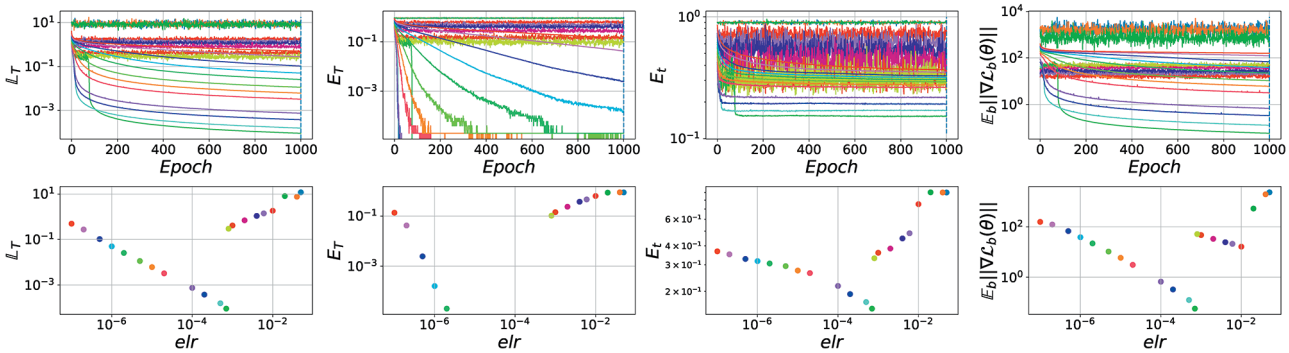

Рис. 2. Основные метрики для различных *elr*. Крайняя правая диаграмма демонстрирует скачкообразные переходы между фазами при изменении *elr*.

на широком классе архитектур и задач показал паритет по качеству при незначительно более медленной скорости сходимости квадратичной функции потерь.

С теоретической точки зрения также нет окончательного ответа. При большой степени перепараметризации, которая свойственна нейронным сетям, и достаточно строгих условиях было показано, что функционально решения с использованием кросс-энтропийной функции потерь и квадратичной функции потерь в точности совпадают [11]. Однако существующие работы не дают ответа на то, можно ли расширить результаты на современные архитектуры глубоких нейронных сетей.

В качестве основных объектов исследования были выбраны среднее значение функции потерь на обучающей выборке L_T, доля неверно классифицированных объектов на обучающей и тестовой выборках E_T, E_t и метрика кривизны $GM = E_b\|\nabla L_b(\theta)\|$.

2.2. Метрики кривизны

В качестве основной метрики ширины предлагается использовать среднюю норму градиентов по отдельным батчам обучающей выборки *GM*. Интуитивно данная метрика показывает, насколько велик разброс градиентов по отдельным объектам в точке пространства весов. В плоских, широких областях данная метрика должна быть мала, в узких – велика.

На практике такая метрика хорошо коррелирует с метриками кривизны второго порядка, такими как след матрицы Фишера или максимальное собственное значение матрицы Гессе. Теоретический анализ также подтверждает высокую корреляцию между данными метриками [12]. При этом вычисление *GM* требует только одного обратного прохода, в отличие от двух обратных проходов при вычислении оценок на статистики Гессиана и матрицы Фишера. Более того, за счет усреднения по батчам, а не по отдельным объектам получается существенно ускорить вычисление оценки кривизны, не теряя в качестве при-

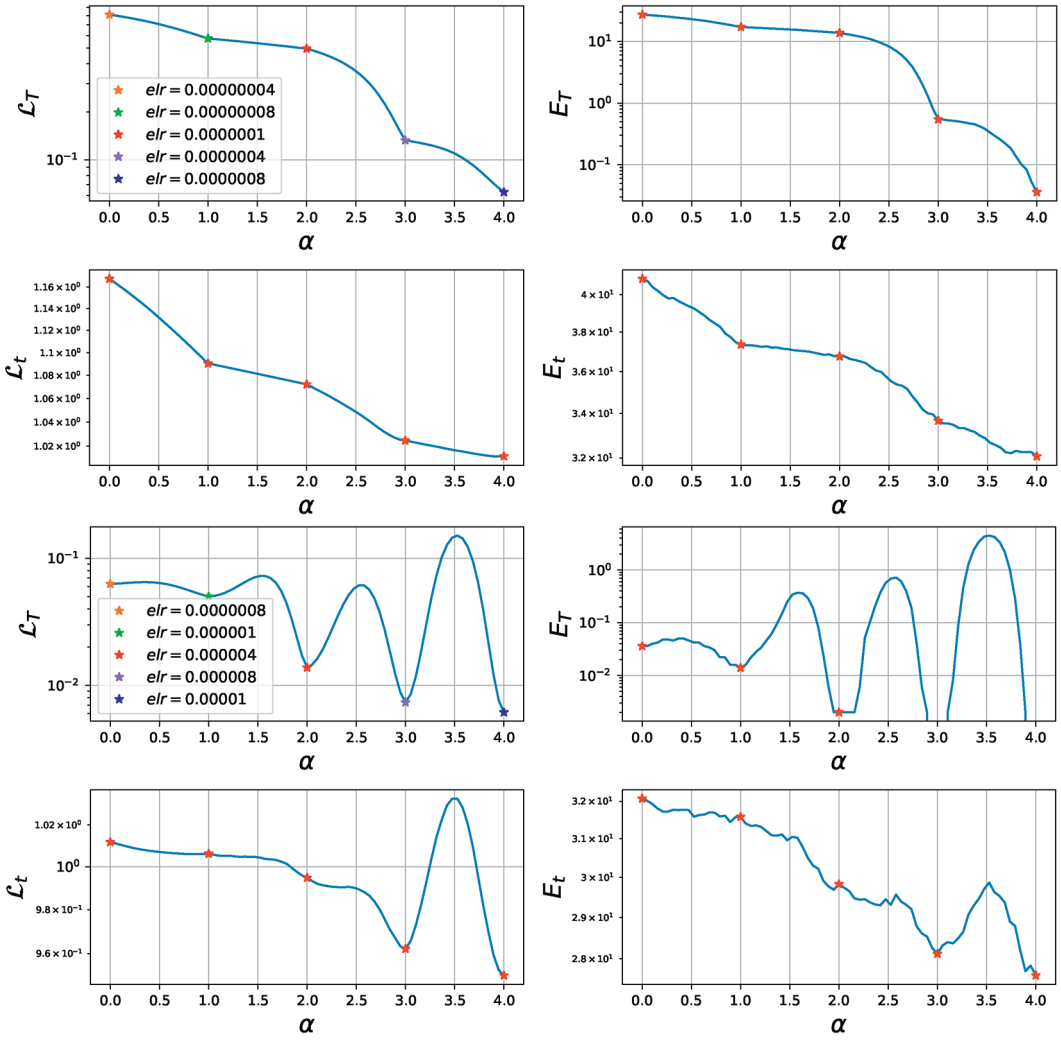

Рис. 3. Mode connectivity для разных моделей из первой фазы. Слева – модели из не сошедшейся первой фазы. Справа – сошедшаяся первая фаза. После достижения сходимости области оптимумов для разных *elr* оказываются линейно несвязными.

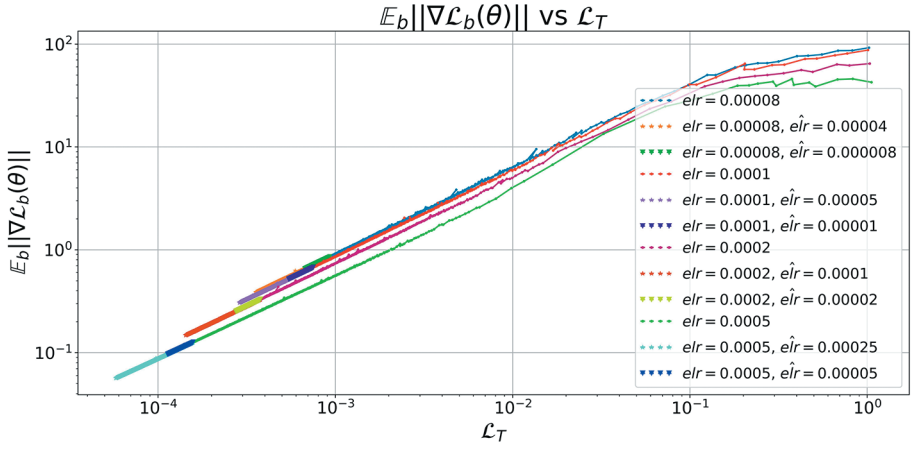

Рис. 4. Фазовая диаграмма при уменьшении *elr* для ConvNet. Итоговый *elr* обозначен как $\widehat{elr}$. Траектории продолжают исходный тренд, что говорит о "застревании" в фиксированной области в окрестности локального минимума.

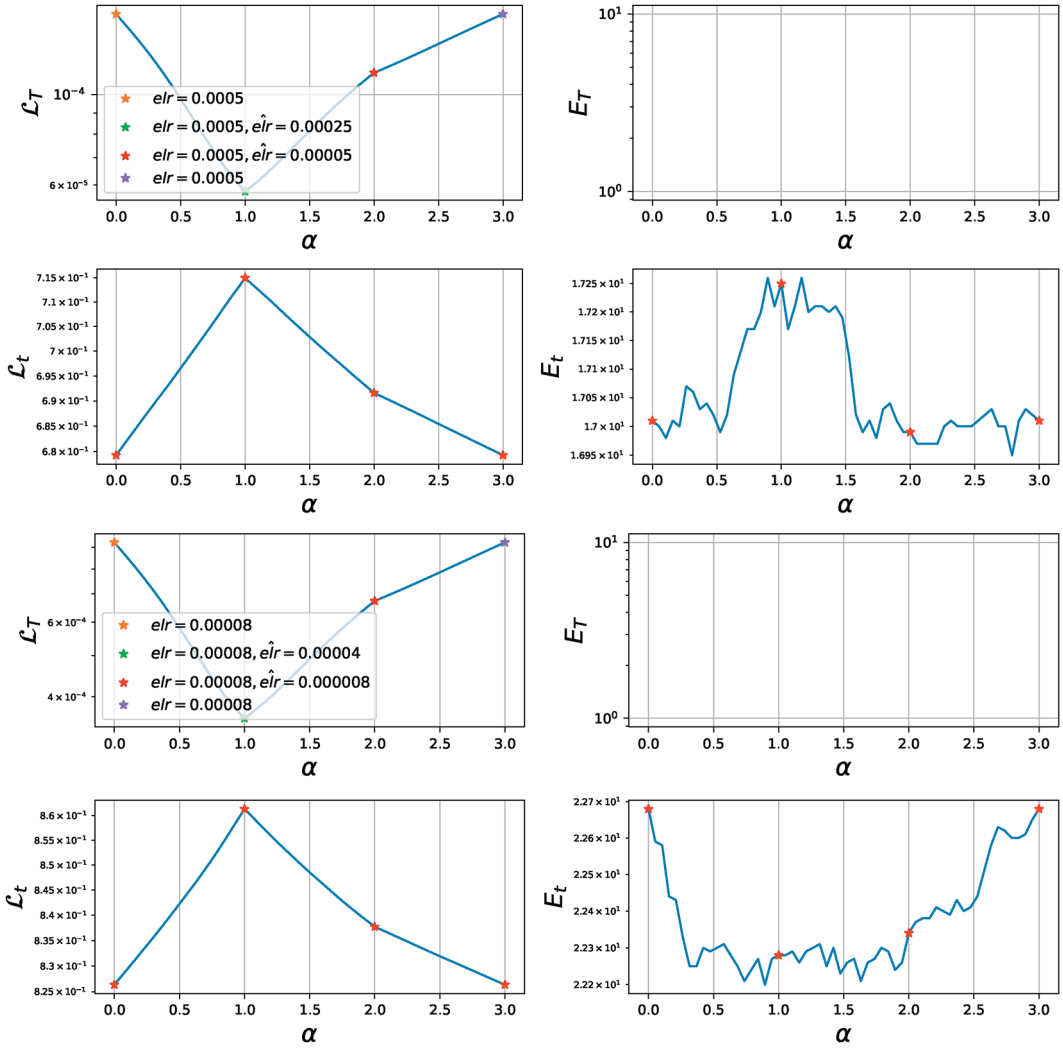

Рис. 5. Mode connectivity для моделей с уменьшенным *elr*. Так как модели достигли нулевой ошибки на обучении, то соответствующие точки на графиках для E_T не отображаются. Модели остаются линейносвязными.

ближения. Сравнение данных метрик приведено в Приложении 2.

3. ОБУЧЕНИЕ С КРОСС-ЭНТРОПИЙНОЙ ФУНКЦИЕЙ ПОТЕРЬ

Для анализа динамики обучения нейронных сетей были обучены FSI модели с различными значениями *elr* для кросс-энтропийной функции потерь.

На рис. 1 видно, что модели разделились на три условных группы. В первой группе модели сходятся в область с широкими минимумами и низким значением функции потерь. Во второй – модели лосс и кривизна колеблются около некоторого значения, как видно на верхних графиках рис. 2. В последней группе модели с наибольшей кривизной.

Проанализируем каждую группу по отдельности.

3.1. Первая фаза

К первой фазе отнесем модели, для которых $elr \le 7 \times 10^{-4}$. Данные модели выделяются тем, что с увеличением *elr* происходит согласованное уменьшение всех метрик в конце обучения. Верхняя граница фазы определяется резким изменением всех метрик (нижние графики рис. 2), что свидетельствует о качественном изменении поведения при переходе через указанную границу. При этом первую фазу можно условно разбить на две подфазы – модели, которые достигают нулевой ошибки на обучающей выборке к концу обучения (сошедшаяся первая фаза) и все остальные модели с меньшими *elr* (не сошедшаяся первая фаза).

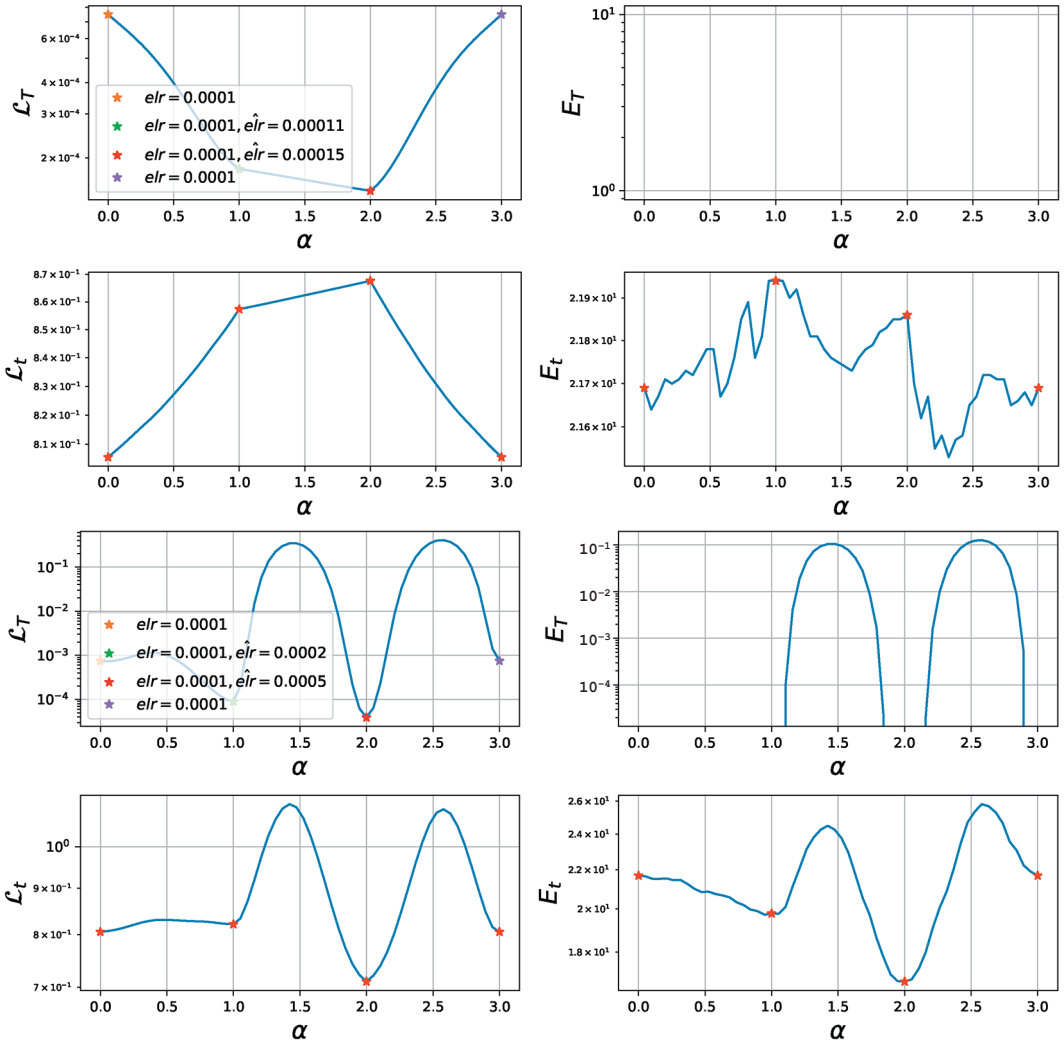

Рис. 6. Mode connectivity для $elr = 0.0001$ при дообучении с большим темпом. При малых увеличениях (левая панель) точки остаются линейно связными. Более сильное увеличение $\widehat{elr}$ приводит к переходу в окрестность другого оптимума (правая панель). Модели, достигшие нулевой ошибки на обучении, на отображаются на графиках для E_T.

Для анализа первой фазы исследуем линейную связность моделей (mode connectivity). Для этого рассмотрим две модели $f(\theta)$, параметризованные весами θ_1, θ_2. Под mode connectivity будем подразумевать значения метрик для моделей на отрезке между парой исходных точек $f(\alpha\theta_1 + (1-\alpha)\theta_2)$, $\alpha \in [0, 1]$. Будем называть модели линейно связными или лежащими в одной области, если на графике mode connectivity отсутствуют ярко выраженные экстремумы в промежуточных точках $\alpha \in (0,1)$. Иначе, будем называть модели линейно несвязными.

Модели, которые не достигают нулевой ошибки ввиду маленького темпа обучения, сходятся в одну линейно связную область. При больших темпах обучения модели успевают разойтись в разные оптимумы, что приводит к наличию пика у значения функции потерь на Графике mode connectivity 3. Стоит отметит, что точная граница между подфазами проходит не по нулевой ошибке на обучении, но по ошибке порядка 10–15 объектов.

Таким образом, при $elr \leq 8 \times 10^{-7}$ модели сходятся в одну и ту же область, но с разной скоростью, что также подтверждается совпадением их траекторий на фазовой диаграмме 1.

По достижении достаточно низкой ошибки на обучении модели начинают сходиться вдоль различных траекторий. При этом увеличение темпа обучения приводит к монотонному росту генерализации. Это хорошо согласуется с тем, что при больших elr разрешающая способность сети уменьшается, что приводит к сходимости во все

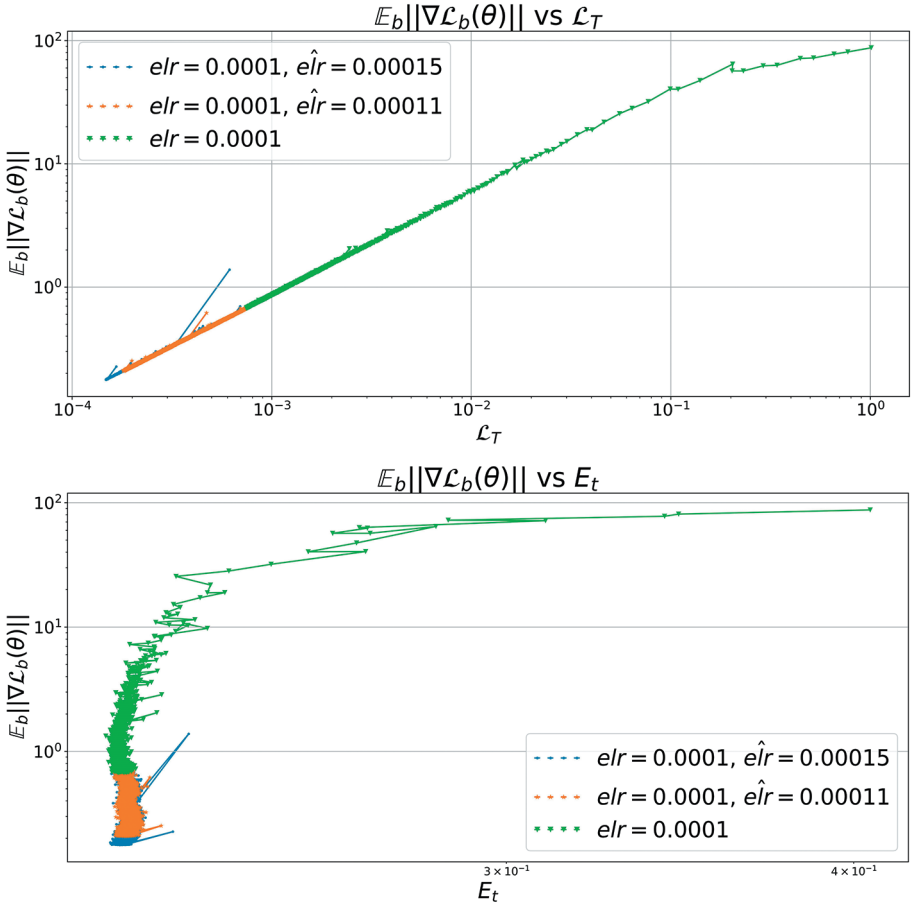

Рис. 7. Фазовые диаграммы для $elr = 0.0001$ при дообучении с бóльшим темпом. Малые увеличения elr не выводят модель из исходного оптимума.

более и более широкие, плоские области, которые в свою очередь и определяют лучшее качество на тестовой выборке. Дальнейший рост elr приводит к тому, что сеть резко перестает сходиться к нулевой ошибке на обучении и качественно меняет свое поведение. Таким образом, данный эксперимент подтверждает утверждение, что лучшая генерализация достигается в самом широком оптимуме, при условии сходимости сети.

Также отсутствие линейной связности в первой фазе при достаточно больших elr показывает, что модели сходятся в разные области пространства весов. Покажем, что внутри каждой из этих областей нет минимумов меньшей ширины. Для этого дообучим модели в течение 5000 итераций, резко уменьшив темп обучения в 2 и в 10 раз.

На диаграмме 4 видно, что уменьшение elr не меняет динамику обучения моделей. Это косвенно подтверждает, что глобальные свойства оптимума остаются стабильными и внутри оптимума заданной ширины нет минимума с меньшей шириной.

Mode connectivity на рис. 5 также подтверждает, что уменьшение elr не приводит к сходимости в линейно несвязные оптимумы.

Теперь рассмотрим поведение сетей при увеличении elr. В зависимости от величины итогового elr возможны несколько принципиальных ситуаций. При небольших коэффициентах увеличения динамика сети меняется слабо. Как видно из рис. 6, 7, модель остается в том же оптимуме с точки зрения генерализации и метрики кривизны.

При дальнейшем увеличении итогового elr сеть начинает обучаться менее стабильно и в какой-то момент наблюдается “скачок” и переход на новую траекторию. Значения E_t показывают, что модель может сойтись в незначительно отличающиеся по качеству минимумы. Результаты демонстрируют, что при наличии скачков во время дообучения предыстория обучения влияет слабо, т.е. модель после выхода из региона нестабильности возвращается на траекторию, которая соответствует изначальному обучению с итоговым $\widehat{elr}$.

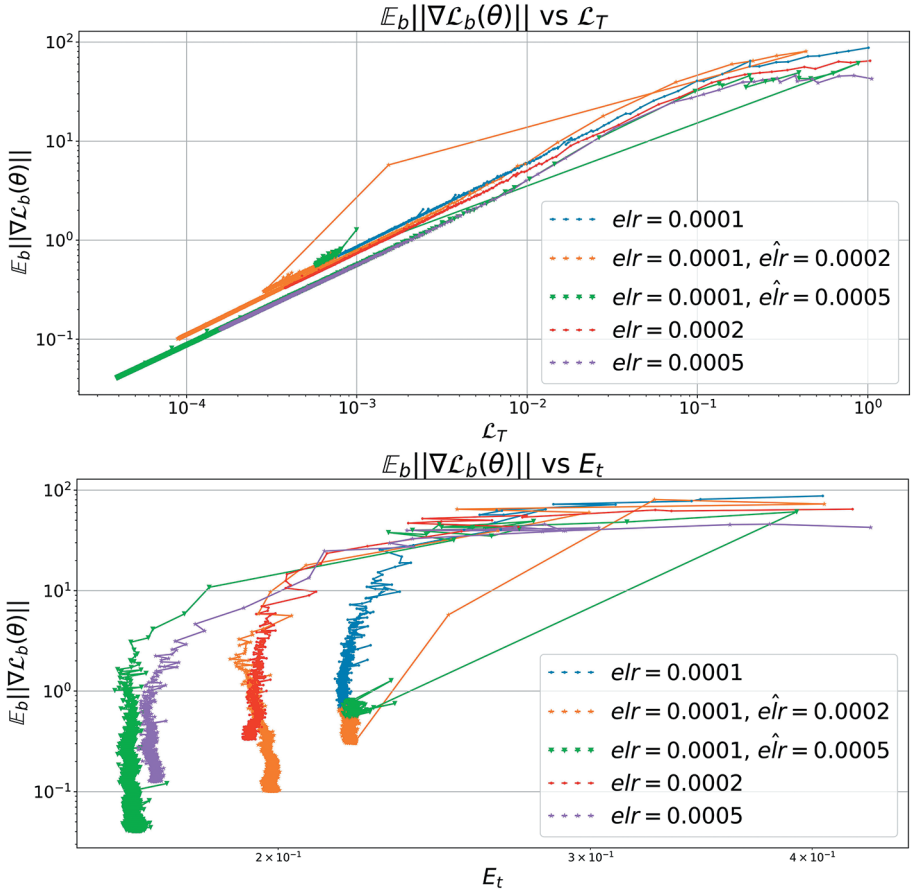

Рис. 8. Фазовые диаграммы для $elr = 0.0001$ при дообучении с большим темпом. На обеих диаграммах виден ступенчатый переход с одной траектории на другую при переходе в другой, более плоский оптимум.

Наконец, если итоговый *elr* превосходит верхнюю границу первой фазы, то модель быстро расходится и переходит в область, которая соответствует обучению с нуля с темпом обучения во второй фазе.

3.2. Вторая фаза

При увеличении темпа обучения до 8×10^{-4} происходит переход в следующую зону: все метрики быстро, в течение первых 5–10 итераций, выходят на фиксированный средний уровень и не меняются в процессе обучения. Переход в новую фазу сопровождается резким скачком всех метрик. При этом во второй фазе модель сходится не к случайным предсказаниям, а к значительно лучшему качеству на обучающей выборке. С увеличением *elr* модель сходится все хуже и хуже до тех пор, пока она не перейдет в третью фазу. Интересно, что в данной фазе увеличение *elr* приводит к уменьшению кривизны.

Можно предположить, что различие между первой и второй фазой вызвано наличием в пространстве весов области с высокой кривизной, которую необходимо преодолеть для достижения низкого значения функции потерь. На фазовой диаграмме можно видеть характерный загиб на траекториях моделей из первой фазы в области $L_T \sim 10^{-1} - 2\times10^{0}$, при прохождении которой наблюдается локальный пик кривизны.

Во второй фазе веса сети не сходятся и на соседних итерациях вектора весов менее коррелированы, чем в первой фазе и с ростом *elr* корреляция падает, что видно из рис. 10.

Дальнейшее увеличение *elr* приводит к тому, что веса на соседних итерациях не коррелируют друг с другом, что соответствует переходу в третью фазу.

Исследуем свойства моделей во второй фазе. В течение всей 1000 итераций не наблюдается значительных изменений средних значений метрик. Модели из-за большого темпа обучения “застревают” на фиксированном уровне и могут уменьшить значения метрик только при условии перехода через “бутылочное горлышко” кривизны. Однако, так как в данной зоне градиенты слишком велики, такой переход возможен только при

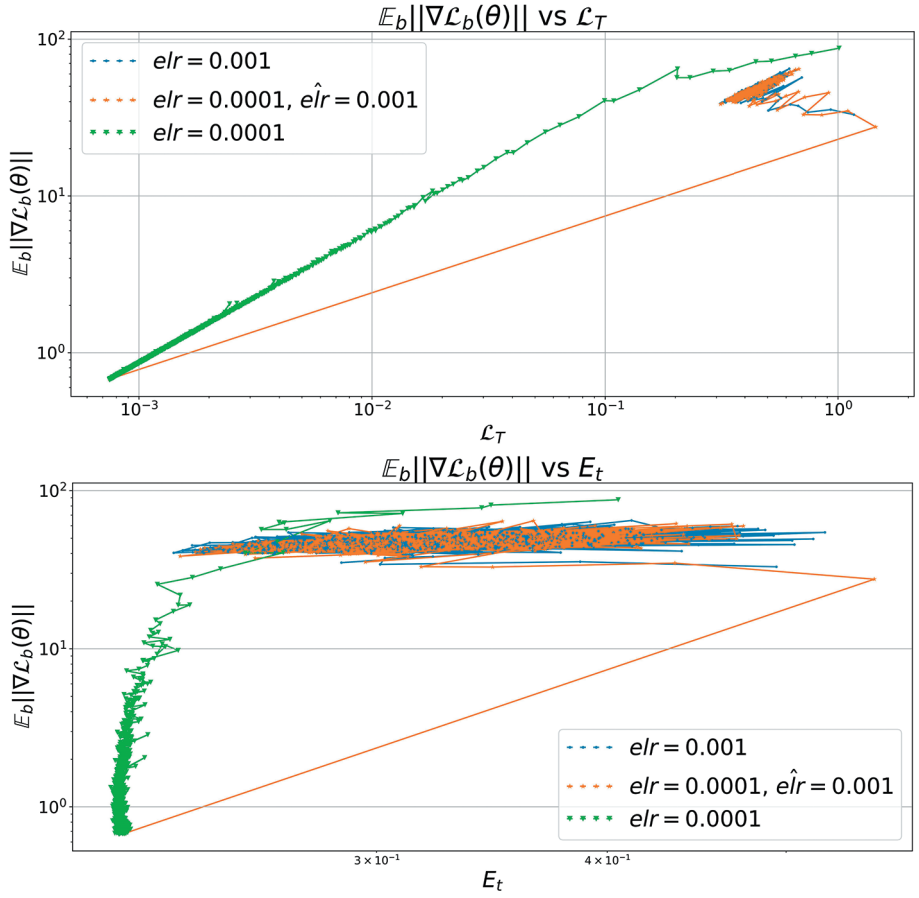

Рис. 9. Фазовые диаграммы для $elr = 0.0001$ при дообучении с большим темпом. При этом происходит мгновенный переход в хаотический режим.

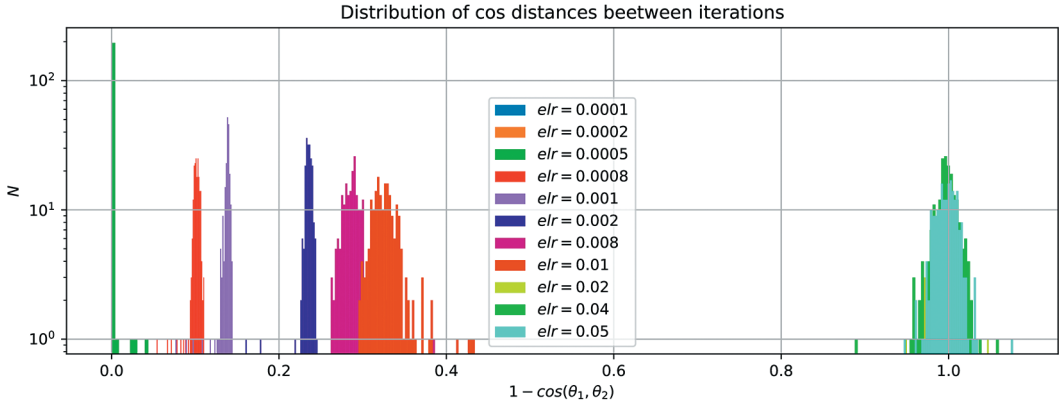

Рис. 10. Распределение косинусных расстояний между всеми соседними эпохами для различных темпов обучения показывает разделение фаз обучения между собой.

уменьшении шага SGD. Поэтому рассмотрим эксперимент с дообучением модели из второй фазы с резко уменьшенным *elr*.

Из рис. 11 видно, что при запуске из небольшого *elr* второй фазы модели ведут себя на фазовой диаграмме подобно тому, как происходил переход из первой фазы в первую на графике 4 — траектория продолжает тренд точек из второй фазы, сходясь в окрестность линии, соответствующей наибольшему темпу обучения первой фазы

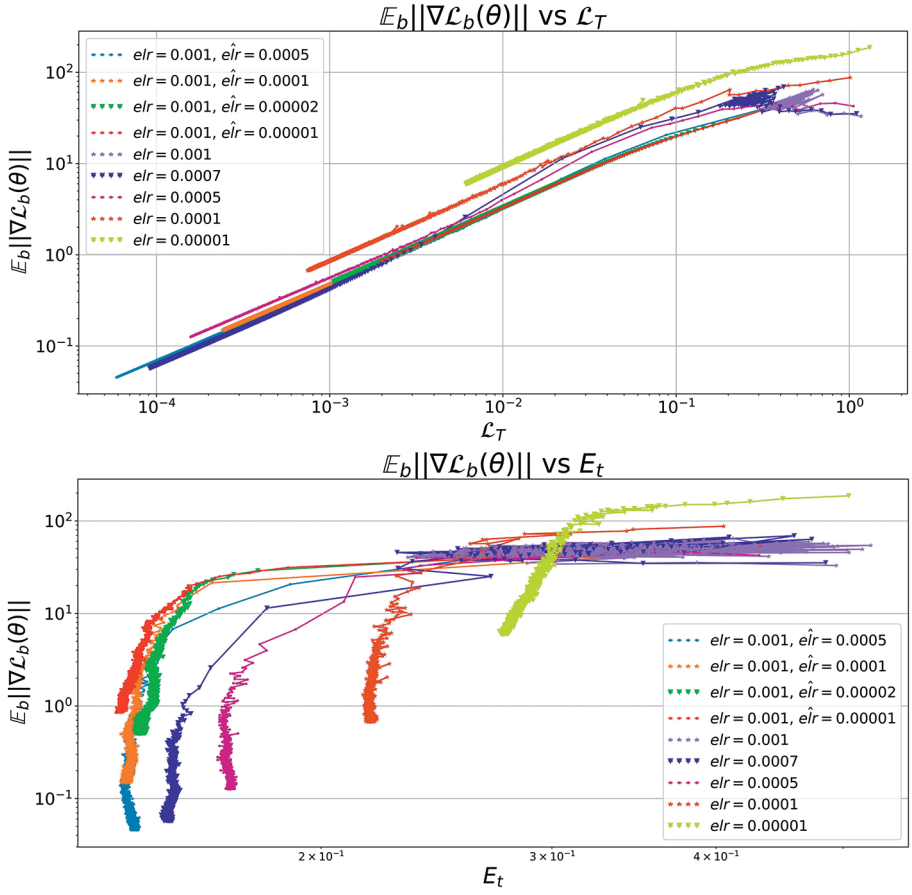

Рис. 11. Фазовые диаграммы для $elr = 10^{-3}$ при дообучении из второй фазы с меньшим темпом. Модели сходятся к самому широкому оптимуму.

($elr = 7\times10^{-4}$), вне зависимости от итогового *elr*. При этом генерализация таких моделей превосходит лучшую генерализацию среди всех сетей, полученных в первой фазе. Некоррелированность *GM* и E_t в данном примере еще раз подчеркивает недостаток локальных метрик кривизны при использовании их как прокси для генерализации.

Запуск дообучения из большего $elr = 10^{-2}$ приводит к противоположным результатам. Во-первых, в соответствии с правым рис. 12 траектории таких моделей сходятся к тем же кривым, которые соответствуют моделям, обучавшимся с заданным *elr* изначально. Левый график показывает, что улучшение итогового качества наблюдается только при маленьких итоговых темпах обучения ($elr \le 10^{-4}$).

Сравнение результатов для большого и маленького *elr* второй фазы показывает, что предобучение во второй фазе тем больше влияет на итоговое тестовое качество E_t, чем меньше стартовый *elr*.

Также независимость итогового качества и траектории при дообучении из маленького *elr* второй фазы позволяет выдвинуть гипотезу, что выбор оптимума в первой фазе осуществляется в момент перехода из второй фазы в первую в самом начале обучения. Для проверки этого утверждения поставим следующий эксперимент: зафиксируем стартовый $elr = 10^{-2}$ и запустим дообучение с кусочно-линейным расписанием темпа обучения 13.

Из рис. 14 видно, что, изменяя скорость уменьшения *elr*, можно получить спектр траекторий, где на одном конце сходимость вдоль линии, соответствующей наибольшему *elrelr* первой фазы, а на другом, при самом быстром изменении *elr* – траектория модели для $\widehat{elr}$ без предобучения. Медленное уменьшение *elr* из второй фазы приводит к тому, что модель будет постепенно переходить на траектории, соответствующие меньшим *elr* второй фазы, до тех пор, пока модель не выскочит в область с меньшей кривизной.

Величина прироста качества на тестовой выборке E_t также плавно зависит от скорости

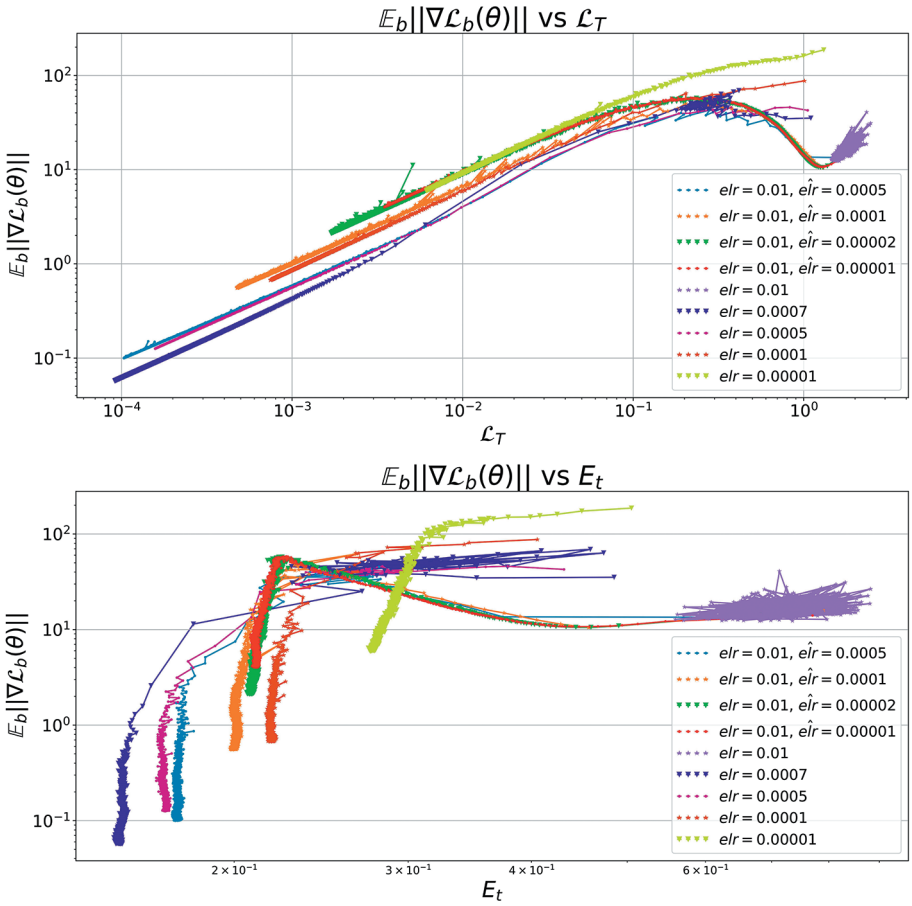

Рис. 12. Фазовые диаграммы для $elr = 10^{-2}$ при дообучении из второй фазы с меньшим темпом. Модели сходятся к оптимумам различной ширины.

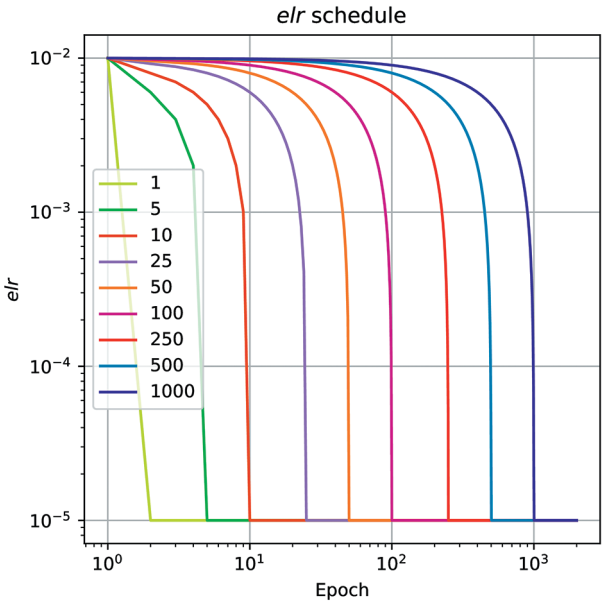

Рис. 13. Линейное расписание *elr* с различной длительностью перехода.

уменьшения *elr*. Таким образом, можно предположить, что оптимальной стратегией для расписания *elr* является как можно дольше находиться в точках второй фазы, прежде чем перейти в первую, так как в момент перехода модель "фиксирует" минимум, в который она будет сходиться. Изменить минимум в первой фазе можно, только существенно увеличив *elr* и добившись нестабильности в обучении, которая позволит "перескочить" в область с другими функциональными характеристиками.

3.3. Третья фаза

Рассмотрим оставшиеся модели, соответствующие $elr \geq 2\times10^{-2}$. При дальнейшем увеличении *elr* во второй фазе снова происходит качественное изменение поведения – градиенты у модели резко увеличиваются, ошибка на обучении становится равной случайному гаданию (90% в случае 10 классов). Переход в эту зону соответствует переходу к случайному предсказанию. Для проверки этого утверждения было выполнено обучение

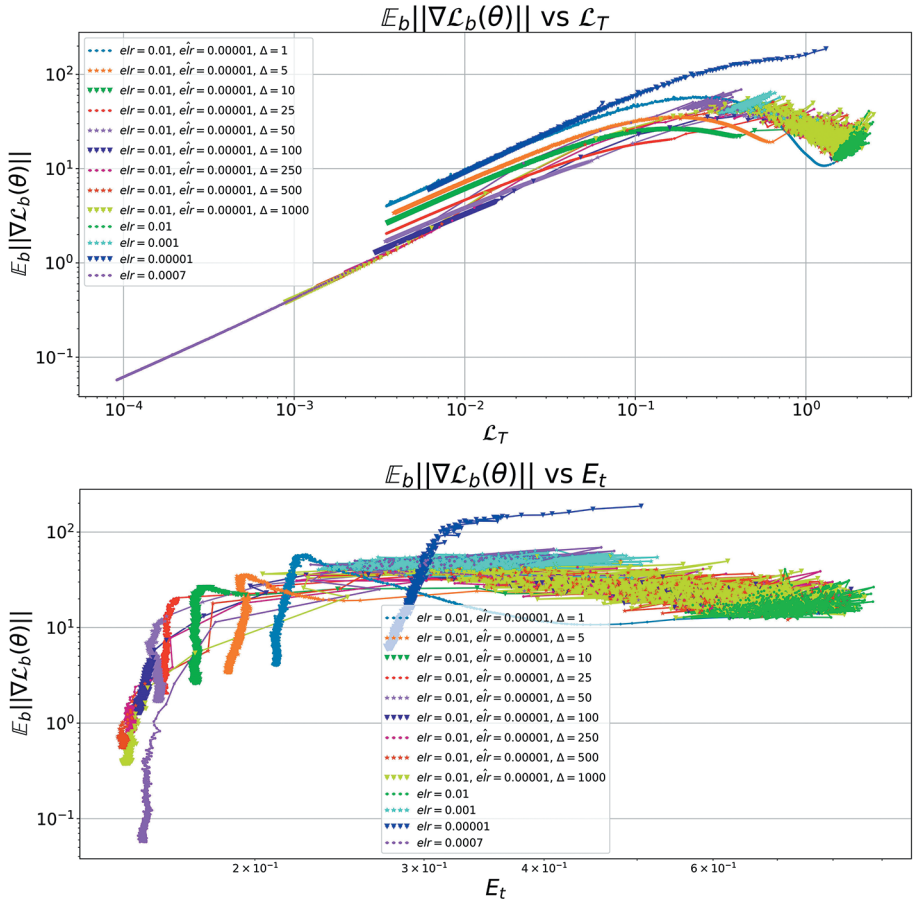

Рис. 14. Фазовые диаграммы для $elr = 10^{-2}$ при дообучении в $\widehat{elr} = 10^{-5}$ с различными расписаниями. Δ – длительность перехода.

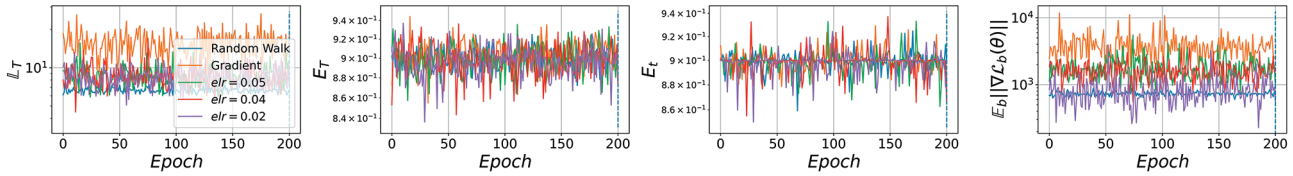

Рис. 15. Основные метрики в третьей фазе. Синяя линия – случайное блуждание, оранжевая линия – движение по градиенту. Обучение в третьей фазе схоже со случайным блужданием.

модели, в которой каждый шаг градиентного спуска заменялся на шаг вдоль случайного направления той же магнитуды. Также, для сравнения, был поставлен эксперимент, где в качестве очередного шага вместо антиградиента используется градиент. Полученные результаты показывают, что случайное блуждание и движение по градиенту позволяют ограничить наблюдаемое поведение кривизны в третьей фазе снизу и сверху.

4. ОБУЧЕНИЕ С КВАДРАТИЧНОЙ ФУНКЦИЕЙ ПОТЕРЬ

Теперь сравним поведение на фазовых диаграммах для квадратичной функции потерь (Mean Squared Error, MSE). Известно, что решение задачи регрессии сложнее, чем решение задачи классификации [11], следовательно, обучение с квадратичной функцией потерь требует большего числа итераций до сходимости. Поэтому в

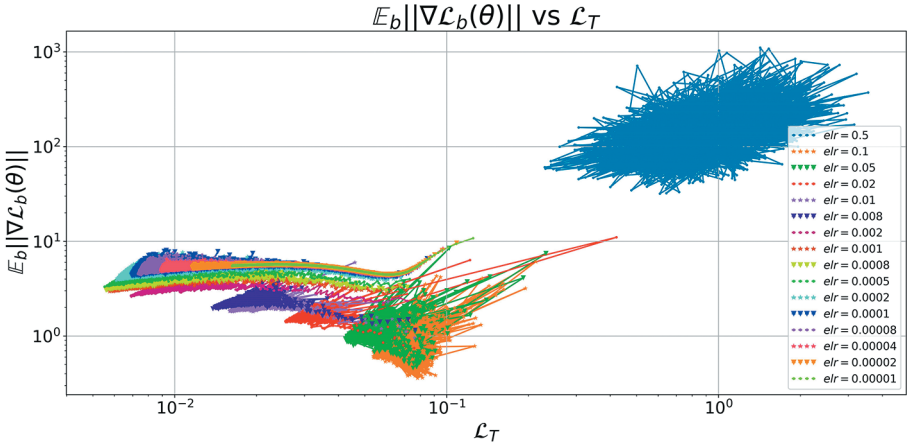

Рис. 16. Фазовая диаграмма для кривизны и квадратичной функции потерь для различных *elr*.

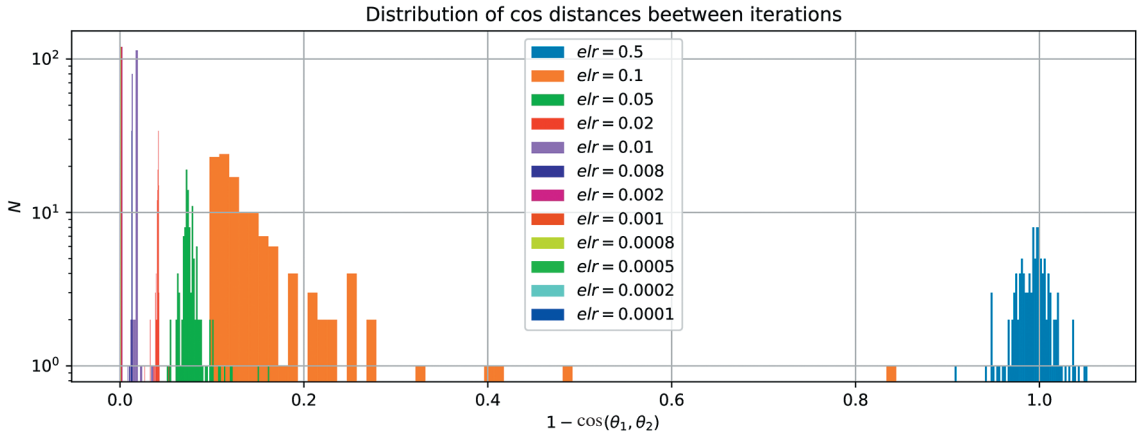

Рис. 17. Распределение косинусных расстояний между соседними эпохами для различных темпов обучения при обучении с квадратичной функцией потерь.

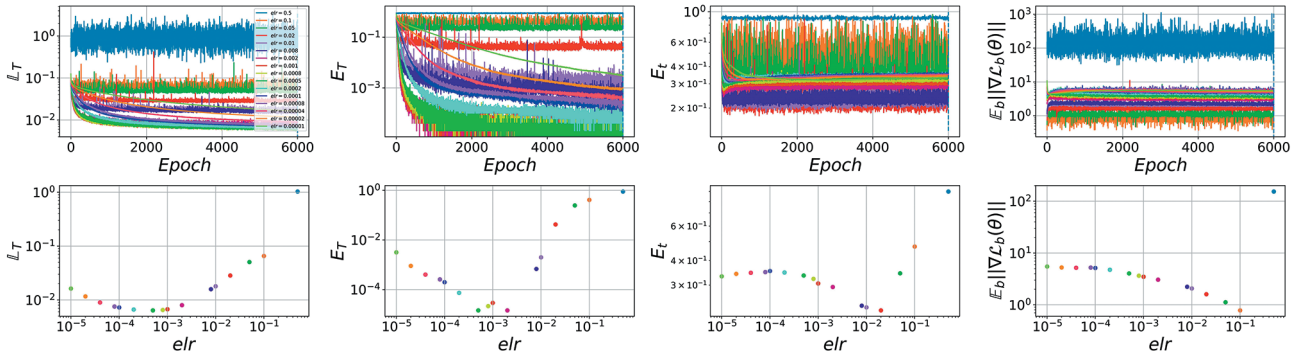

Рис. 18. Основные метрики для различных *elr* для ConvNet с квадратичной функцией потерь. График ошибки на тесте указывает на наличие двух фаз обучения.

экспериментах с MSE все модели будем обучать 6000 итераций.

Анализ фазовой диаграммы 16 позволяет четко отделить третью фазу ($elr=0.5$) и не сошедшуюся часть первой фазы ($elr \lesssim 0.0001$). Остальные модели визуально напоминают как первую, так и вторую фазу для кросс-энтропии. Для более точного анализа рассмотрим распределение коси-

Таблица 1. Базовая архитектура сети ConvNet

№	Слой
1	Conv2d(3, 32, kernel_size=(3, 3), padding=(1, 1), bias=False)
2	BatchNorm2d(32, momentum=0.1, affine=False)
3	ReLU()
4	Conv2d(32, 64, kernel_size=(3, 3), padding=(1, 1), bias=False)
5	BatchNorm2d(64, momentum=0.1, affine=False)
6	ReLU()
7	MaxPool2d(kernel_size=2, stride=2)
8	Conv2d(64, 128, kernel_size=(3, 3), padding=(1, 1), bias=False)
9	BatchNorm2d(128, momentum=0.1, affine=False)
10	ReLU()
11	MaxPool2d(kernel_size=2, stride=2)
12	Conv2d(128, 256, kernel_size=(3, 3), padding=(1, 1), bias=False)
13	BatchNorm2d(256, momentum=0.1, affine=False)
14	ReLU()
15	MaxPool2d(kernel_size=2, stride=2)
16	MaxPool2d(kernel_size=4, stride=4)
17	Linear(in_features=256, out_features=10, bias=True)

нусных расстояний для последовательных моделей вдоль траектории обучения. Из рис. 17 видно, что модели с $elr \gtrsim 0.02$ можно отнести ко второй фазе. Остальные – отнесем к первой фазе.

Теперь соотнесем такое разделение на фазы с поведением метрик на рис. 18. Видно, что ни одна из моделей не сошлась к нулевой ошибке на обучающей выборке, что еще раз подтверждает, что модели с квадратичным лоссом сходятся медленнее моделей с кроссэнтропийной функцией потерь. Другим отличием является тот факт, что граница между первой и второй фазами более размытая. Действительно, модели с $elr \sim 0.008$–0.02 монотонно уменьшают число неправильно классифицированных объектов, при этом оставаясь в области с фиксированой шириной. Это контрастирует с первой фазой кросс-энтропии, где лосс и кривизна убывали согласовано вдоль всей траектории.

Так как на всей выборке из 50 000 объектов модели с MSE лоссом не достигают сходимости, обучим сети на подвыборке объемом 10%.

В таком случае метрики 20 и фазовая диаграмма 19 становятся похожими на случай кроссэнтропии. По фазовой диаграмме можно четко выделить все три фазы обучения, а также две подфа-

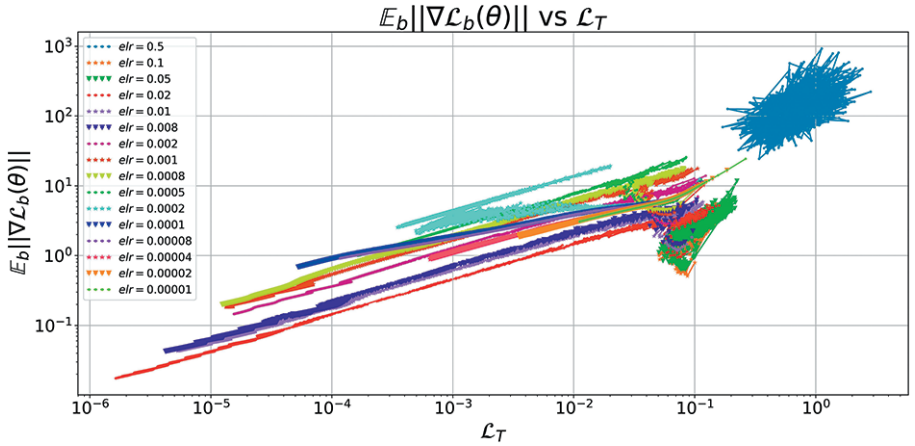

Рис. 19. Фазовая диаграмма для кривизны и квадратичной функции потерь для различных *elr*. ConvNet на подвыборке из CIFAR10.

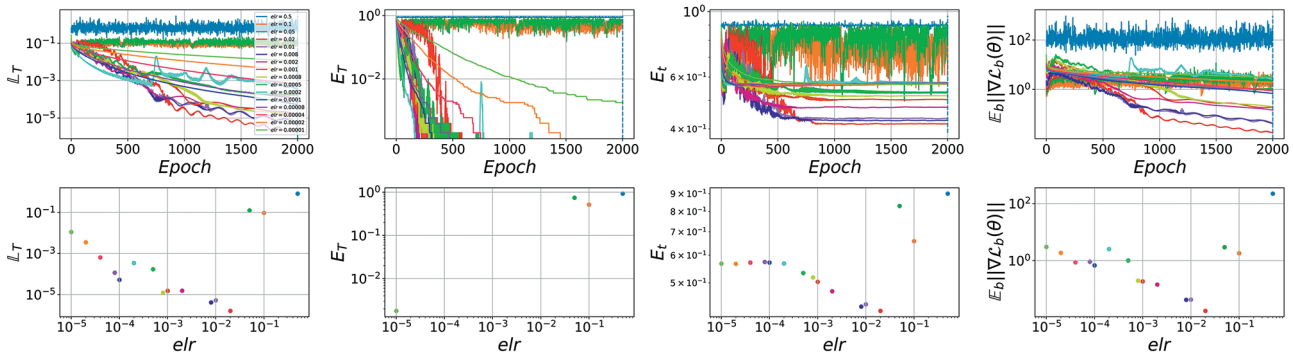

Рис. 20. Основные метрики для различных *elr*. ConvNet с квадратичной функцией потерь на подвыборке из CIFAR10.

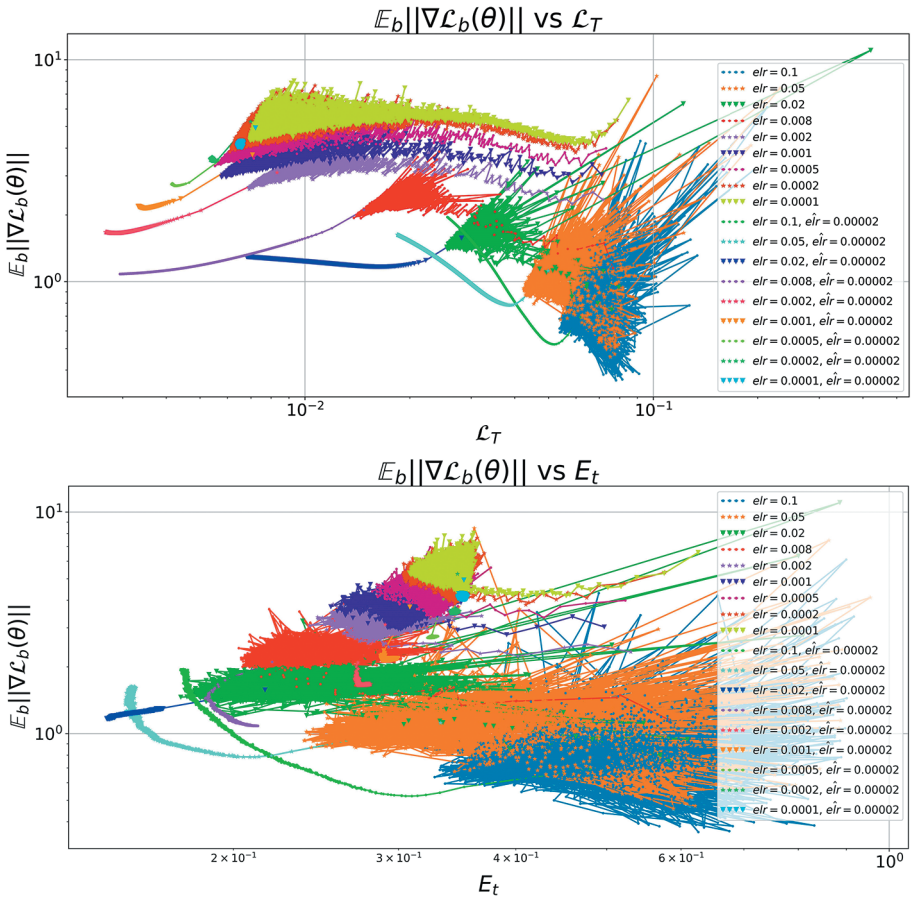

Рис. 21. Фазовые диаграммы при дообучении с меньшим *elr* для MSE.

зы в первой фазе. Оптимальный *elr* с точки зрения ошибки на тестовой выборке соответствует наибольшему темпу обучения в первой фазе. В той же точке достигается минимальная кривизна.

По аналогии с кроссэнтропийной функцией потерь рассмотрим эксперимент с дообучением с различными *elr*.

Анализ результатов показывает, что уменьшение *elr* из траекторий, относящихся к первой фазе, приводит к поведению, с одной стороны, согласованному с кроссэнтропией – траектории продолжают двигаться вдоль того же тренда, что был и до уменьшения темпа. При этом функция потерь убывает. Однако, в отличие от кроссэнтропии, где кривизна монотонно убывала, для квадратичного лосса *GM* или стабилизируется ($elr \approx 10^{-2}$, граница между первой и второй фазами), или даже начинает возрастать (меньшие *elr*).

Такое поведение свидетельствует о начале переобучения, что более четко видно на правом графике 21 – сначала E_t убывает, но затем начинает резко расти.

5. ВЫВОДЫ

Проведенные исследования показали наличие нескольких фаз обучения полностью масштаб инвариантных сетей на сфере в зависимости от темпа обучения.

При этом первая фаза обучения соответствует сходимости весов в оптимум фиксированной ширины.

Вторая фаза определяется хаотическим равновесием, при котором метрики для сети стабилизируются около некоторого уровня. При этом область, в которой стабилизируется нейронная сеть, во второй фазе оказывает решающее влияние на итоговое качество модели.

Третья фаза характеризуется переходом к блужданию по сфере, корреляция между различными моделями на соседних итерациях отсутствует.

Выявленные фазы проявляются для разных функций потерь, как для квадратичного лосса, так и для кросс-энтропии. Исследование в Приложении 4 показывает, что выводы воспроизводятся в том числе и на сетях с не масштабно инвариантными параметрами и с регуляризацией.

Также результаты дополнительно подтверждают гипотезу о том, что для MSE требуется большее число итерации или большее число параметров у сети для сходимости к схожему качеству по сравнению с кросс-энтропией.

Приложение

1. АРХИТЕКТУРА СЕТИ

В качестве основной FSI модели используется сверточная сеть из четырех слоев в соответствии с табл. 1.

Следуя примеру [13], последний линейный слой зафиксирован в случайной инициализации таким образом, чтобы его норма весов равнялась 10 в случае кросс-энтропии и 1.5 в случае квадратичного лосса. Данная норма близка к тем значениям, которые норма последнего слоя принимает при обучении всех параметров сети. Инициализация нормы необходима для достижения низкой ошибки на обучающей выборке.

2. СРАВНЕНИЕ МЕТРИК КРИВИЗНЫ

Для валидации корректности выбранной метрики кривизны $GM = E_b \|\nabla L_b(\theta)\|$ для модели ConvNet было произведено сравнение фазовых диаграмм с использованием других локальных способов оценки кривизны: следа матрицы Фишера $tr\ F$ и максимального собственного значения Гессиана $\max_i \lambda_i$.

Сравнивая рис. 22, 23 с исходной диаграммой 1, можно видеть четкое сходство: во-первых, все три фазы обучения присутствуют для всех метрик кривизны. Во-вторых, еще более заметен эффект “бутылочного горлышка” – области высокой кривизны, которую модели из первой фазы должны пройти, прежде чем спуститься в область низких значений функции потерь.

3. ОПТИМИЗАЦИЯ НА СФЕРЕ

Так как в дизайне эксперимента предполагается, что модель задана на сфере радиуса 1, то кор-

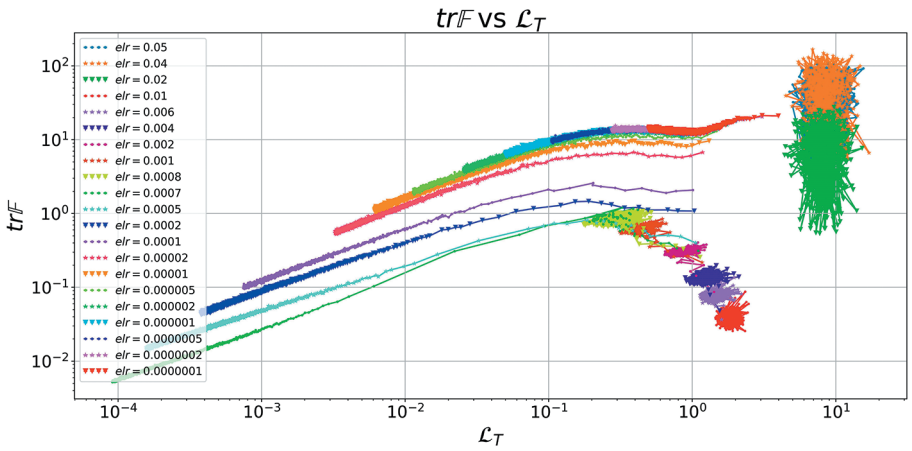

Рис. 22. Фазовая диаграмма для следа матрицы Фишера и кросс-энтропийной функции потерь.

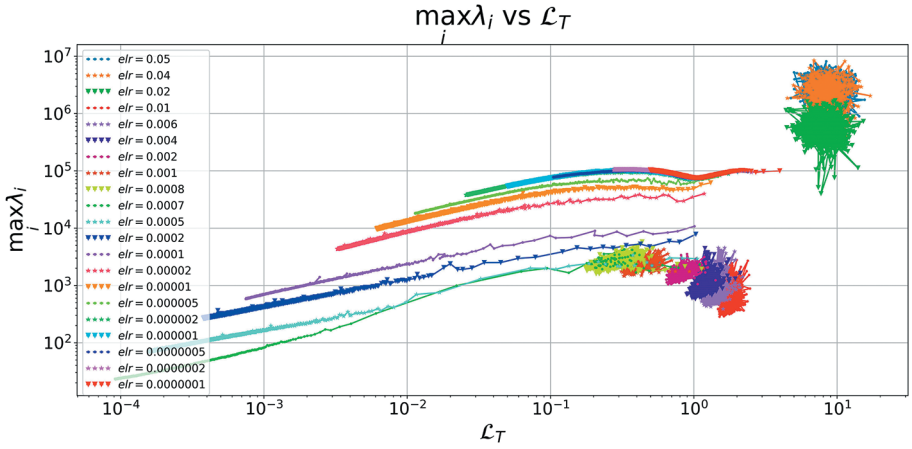

Рис. 23. Фазовая диаграмма для максимального собственного значения Гессиана и кросс-энтропийной функции потерь.

ректным способом оптимизации такой модели являются методы Римановой оптимизации. Однако реализация таких алгоритмов на практике избыточна. Действительно, положим, что сеть $f(\theta)$, заданная в соответствии с уравнением (1), обучается методом градиентного спуска в пространстве весов $\theta \in R^d$ с некоторым темпом обучения *lr*. Так как после выполнения шага оптимизации $\theta' = \theta - lr \times \nabla f(\theta)$ норма весов изменится, спроецируем веса обратно на сферу: $\theta^* = \frac{\theta'}{\|\theta'\|}$. Таким образом, переход от θ к θ^* соответствует движению по сфере вдоль дуги большого круга сферы на расстояние Δl. Истинное значение скорости обучения на сфере $elr = \frac{\Delta l}{\|\nabla f(\theta)\|}$. Определим погрешность между *lr* и *elr*

$$r = \frac{elr}{lr} = \frac{\Delta l}{lr\|\nabla f(\theta)\|} = \frac{\arctan[lr\|\nabla f(\theta)\|]}{lr\|\nabla f(\theta)\|}.$$

Заметим, r зависит от эффективной длины шага $lr\|\nabla f(\theta)\|$, что потенциально может приводить к высокой погрешности при больших нормах градиента.

Рассмотрим, как меняется относительная погрешность $1 - r$ в процессе обучения с кроссэнтропийной функцией потерь.

График 25 показывает, что для всех моделей в первой и второй фазах оптимизация с помощью SGD с перенормировкой весов соответствует “честной” Римановой оптимизации на сфере с *elr*, отличающимся не более чем на 2% от темпа обучения *lr* в исходном евклидовом пространстве. В третьей фазе, из-за роста градиента, данная процедура приводит к существенной переоценке *elr*. Интуитивно, в третьей фазе каждый шаг градиентного спуска приводит к повороту вектора весов на угол $\alpha \approx 90°$. Нужно отметить, что процедура с нормировкой весов не может исследовать поведение сетей при экстремально больших *elr*, которые приводят к повороту вектора весов на угол, больший 90°.

4. РЕЗУЛЬТАТЫ ДЛЯ VGG16BN

Рассмотрим более практический дизайн эксперимента. В качестве нейронной сети выберем VGG16 [14] с батч-нормализацией. При этом не будем фиксировать линейные слои и аффинные параметры батч-нормализации. Также при обучении не будем приводить веса на сферу, а будем оптимизировать параметры в исходном пространстве с помощью SGD без моментума. Так как сеть обучается без ограничений, то понятие

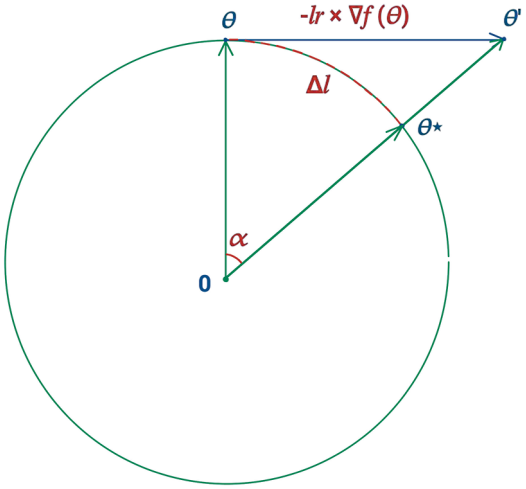

Рис. 24. Оптимизация на сфере.

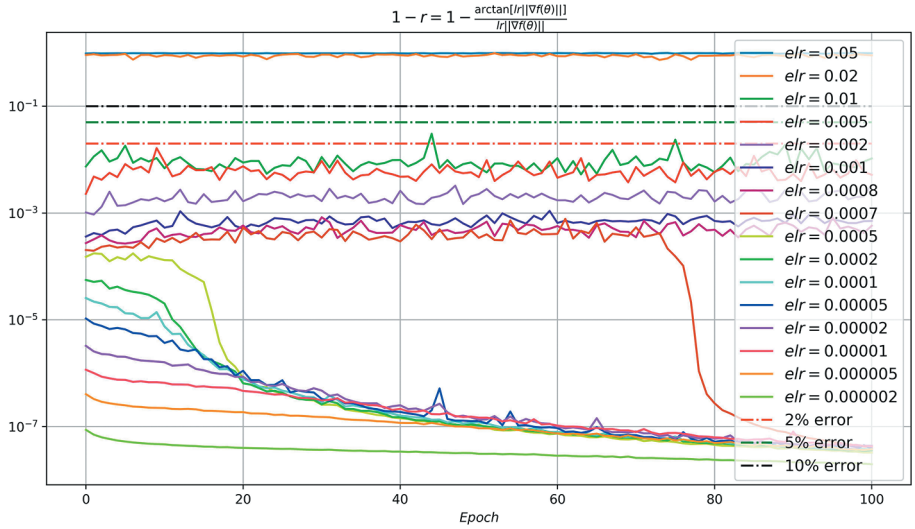

Рис. 25. Погрешность оптимизации в зависимости от *lr*.

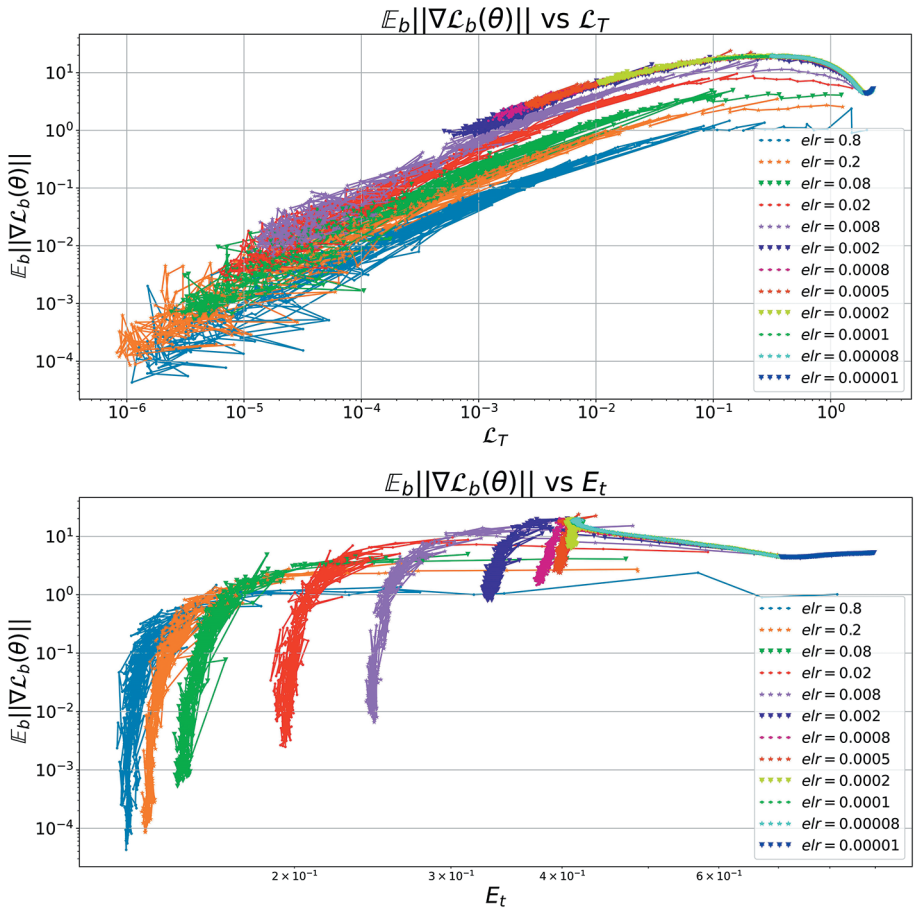

Рис. 26. Фазовые диаграммы для VGG16BN, $wd = 0.0$.

elr в такой постановке теряет смысл. Однако для консистентности в данной секции под *elr* будем понимать обычный темп обучения (*lr*).

Прежде всего рассмотрим обучение без L_2 регуляризации. В таком случае норма весов сети в процессе оптимизации возрастает, что приводит

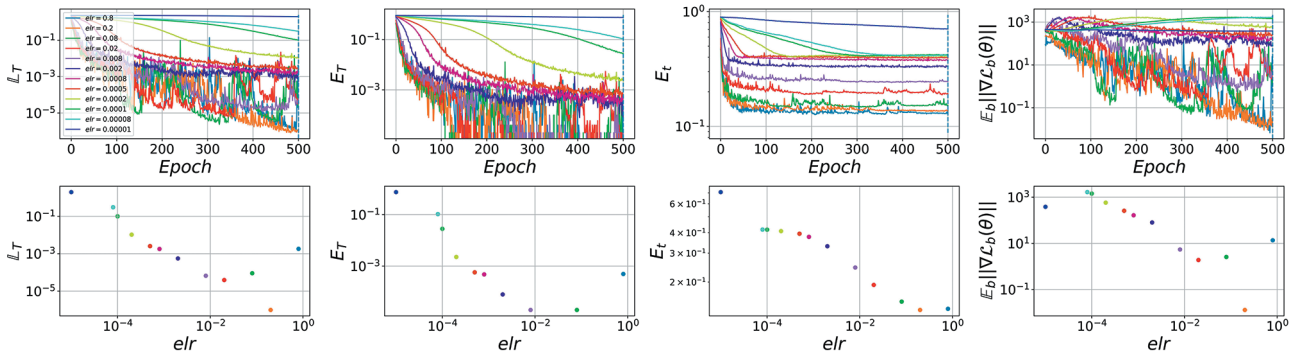

Рис. 27. Основные метрики для различных *elr* для VGG16BN, *wd* = 0.0.

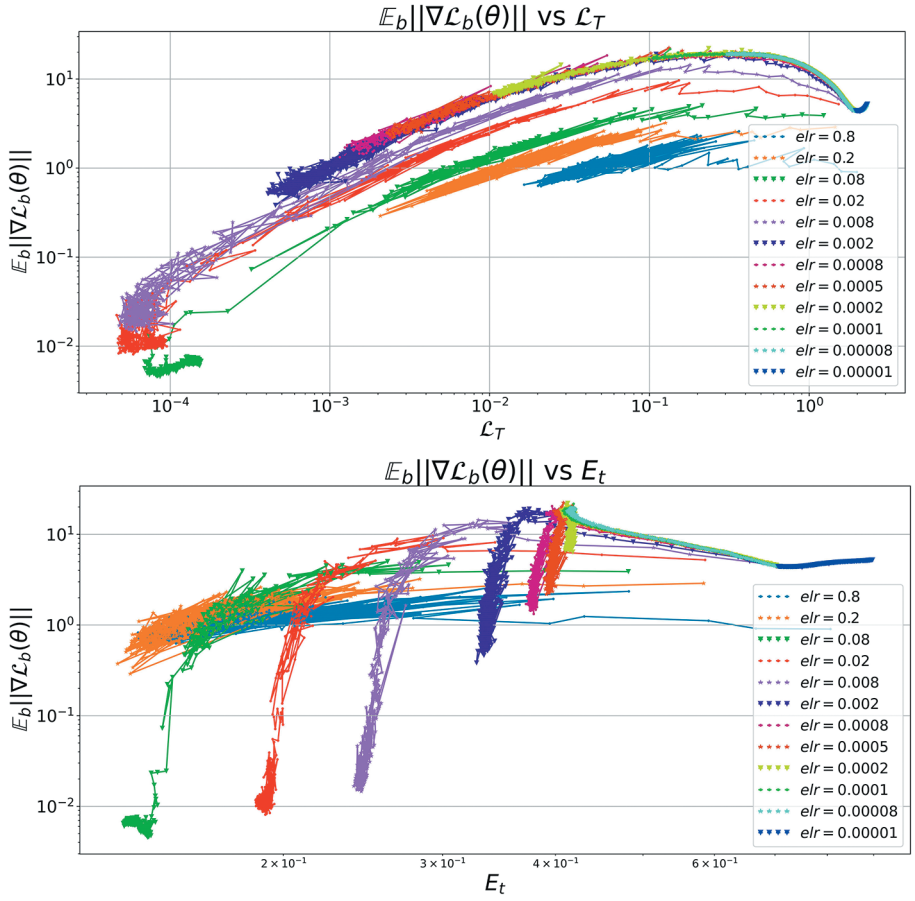

Рис. 28. Фазовые диаграммы для VGG16BN, *wd* = 0.0001.

к нестабильному обучению при больших *elr*. На практике обучение с большими *elr* приводит к мгновенной дестабилизации и расхождению весов сети в *NaN*. Поведение на фазовой диаграмме 26 для меньших темпов обучения показывает картину, схожую с первой фазой, включая две ее подфазы. При этом к концу обучения динамика становится значительно более шумной по сравнению с ConvNet, однако, линейный тренд поведения линий на фазовой диаграмме остается.

Метрики на рис. 27 демонстрируют тренд на снижение с ростом *elr*, однако, из-за большой дисперсии метрики для обучающей выборки нестабильны. Несмотря на это, можно четко видеть, что оптимальная генерализация достигается при

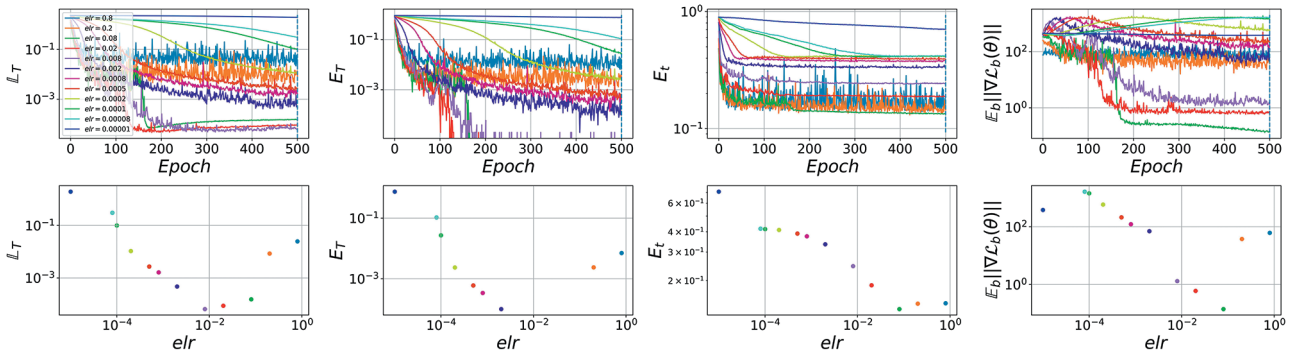

Рис. 29. Основные метрики для различных *elr* для VGG16BN, $wd = 0.0001$.

наибольших *elr* первой фазы, там же, где достигается минимум кривизны.

Добавим L_2 регуляризацию в модель. Даже при малых значениях weight decay ($wd = 10^{-4}$) стабильность сети значительно возрастает и диапазон допустимых *elr* расширяется.

Теперь на правой диаграмме 28 видны первая фаза, а также несколько моделей из второй фазы, которые не переходят в область низких градиентов. При этом для моделей с большими *elr* в первой фазе наблюдается загиб значения функции потерь в конце обучения. Вероятнее всего данный эффект связан с регуляризацией, так как при малой величине лосса, weight decay начинает доминировать в процессе оптимизации. Видно, что такая регуляризация позитивно сказывается и на генерализации – модель сходится в более широкий оптимум с лучшим тестовым качеством E_t. Анализ метрик на нижней панели рис. 29 подтверждает гипотезу, что лучшая генерализация достигается при максимальном *elr* первой фазы. Аналогично результатам для ConvNet все метрики в первой фазе монотонно снижаются с ростом *elr*. При этом модели из второй фазы выходят на константный уровень метрик не с самого начала обучения, а подобно результатам для MSE, сначала демонстрируют снижение.

СПИСОК ЛИТЕРАТУРЫ

1. *Hui L., Belkin M.* Evaluation of Neural Architectures Trained with Square Loss vs Cross-Entropy in Classification Tasks, 2021.
2. *Smith L.N.* Cyclical Learning Rates for Training Neural Networks, arXiv, 2015.
3. *Santurkar S., Tsipras D., Ilyas A., Madry A.* How Does Batch Normalization Help Optimization?, arXiv, 2018.
4. *He F., Liu T., Tao D.* Why ResNet Works? Residuals Generalize, arXiv, 2019.
5. *Foret P., Kleiner A., Mobahi H., Neyshabur B.* Sharpness-Aware Minimization for Efficiently Improving Generalization, arXiv, 2020.
6. *Jiang Y., Neyshabur B., Mobahi H., Krishnan D., Bengio S.* Fantastic Generalization Measures and Where to Find Them, arXiv, 2019.
7. *Allen-Zhu Z., Li Y., Liang Y.* Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers, arXiv, 2018.
8. *Badrinarayanan V., Mishra B., Cipolla R.* Understanding symmetries in deep networks, arXiv, 2015.
9. *Bosman A.S., Engelbrecht A., Helbig M.* Visualising Basins of Attraction for the Cross-Entropy and the Squared Error Neural Network Loss Functions, 2019.
10. *Demirkaya A., Chen J., Oymak S.* Exploring the Role of Loss Functions in Multiclass Classification, 2020 54th Annual Conference on Information Sciences and Systems (CISS), 2020.
11. *Muthukumar V., Narang A., Subramanian V., Belkin M., Hsu D., Sahai A.* Classification vs regression in overparameterized regimes: Does the loss function matter?, 2021.
12. *Thomas V., Pedregosa F., van Merriënboer B., Manzagol P.-A., Bengio Y., Roux N.L.* On the interplay between noise and curvature and its effect on optimization and generalization," ConferenceProceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, 2020.
13. *Lobacheva E., Kodryan M., Chirkova N., Malinin A., Vetrov D.* On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay, 2021.
14. *Simonyan K., Zisserman A.* Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv, 2014.
15. *Nakkiran P., Kaplun G., Bansal Y., Yang T., Barak B., Sutskever I.* Deep Double Descent: Where Bigger Models and More Data Hurt, arXiv, 2019.

EDN: WENTLH
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 70–72

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ГЕОМЕТРИЧЕСКОЕ ГЛУБОКОЕ ОБУЧЕНИЕ ДЛЯ ДИЗАЙНА КАТАЛИЗАТОРОВ И МОЛЕКУЛ

© 2022 г. Р. Ю. Лукин[1,*], Р. А. Григорьев[2]

Представлено академиком РАН Г.И. Савиным
Поступило 28.10.2022 г.
После доработки 31.10.2022 г.
Принято к публикации 03.11.2022 г.

Применение глубокого обучения для поиска катализаторов является важной задачей для решения вызванных глобальным потеплением проблем хранения энергии и преобразования парниковых газов в более ценные продукты. В нашей работе мы представляем несколько графовых нейронных сетей (GNN), включая сверхточные архитектуры и архитектуры передачи сообщений с физически информированными атрибутами узлов и ребер для атомистических систем. Мы демонстрируем улучшение прогнозов энергии адсорбции в наборе данных OC20 с использованием предложенной нами архитектуры в терминах средней абсолютной ошибки прогнозируемой энергии и энергии в пределах пороговых показателей. Предлагаемые архитектуры устойчивы к переобучению и могут быть использованы для прогнозирования экспериментальных и квантово-химических свойств широкого спектра материалов и молекул. Мы предлагаем использовать две архитектуры GNN (EdgeUpdateNet и OFMNet) вместе с расширенным методом описания узлов и ребер. Мы представляем отпечатки ребер как элементы матриц межатомного взаимодействия (матрица Кулона, матрица суммы Эвальда, синусоидальная матрица). Для отпечатков пальцев узлов мы используем элементы матрицы орбитального поля (OFM), однократного представления электронного состояния атомов с окружающими атомными орбиталями. Кроме того, мы предлагаем и реализуем представление каталитически активных атомов в виде подграфа. Предлагаемые методы и архитектуры демонстрируют повышение точности прогнозирования энергии адсорбции. Особенно значительные улучшения наблюдаются в примерах, не относящихся к предметной области, как для адсорбатов, так и для катализаторов. Возможности обобщения и экстраполяции на примеры предлагаемых архитектур вне предметной области также делают предлагаемые GNNs пригодными для использования при скрининге катализаторов в обширном химическом пространстве.

Ключевые слова: графовые нейросети, глубокое обучение, квантовая химия, катализ, хемоинформатика

DOI: 10.31857/S2686954322070153

1. ВВЕДЕНИЕ

Компьютерное химическое моделирование и экспериментальные измерения – это два традиционных метода, которые широко применяются в области материаловедения. Однако использование этих двух методов для ускорения поиска материалов и проектирования затруднено, поскольку они отнимают много времени и неэффективны. В большинстве случаев для расчетов электронной структуры использовалась теория функционала плотности (DFT). Несмотря на широкое применение машинного обучения (ML) ко многим проблемам, возникшим за последние годы при прогнозировании свойств молекул и материалов с использованием машинного обучения, возрастает роль подходов к глубокому обучению. Достижения в области атомистического машинного обучения оказывают большое влияние во многих областях, включая материаловедение, катализ и разработку лекарств. Алгоритмы ML – это обучаемый оценщик между входными представлениями материалов и молекул и выходными данными, представляющими интерес для физических или химических свойств. Эти модели использовались для прогнозирования широкого спектра свойств для различных классов материалов, включая точечные дефекты для нейроморфных вычислений, солнечных фотокатализаторов и, особенно, гете-

[1] *Лаборатория в сфере развития продукта ИИ для новых материалов Исследовательского центра в сфере искусственного интеллекта, Университет Иннополис, Иннополис, Россия*

[2] *Исследовательский центр в сфере искусственного интеллекта, Университет Иннополис, Иннополис, Россия*

*****E-mail: r.lukin@innopolis.ru*

рогенных, и гомогенных, катализаторов. Обычно первым шагом всех основанных на ML подходов к прогнозированию свойств материала является представление материала. Кристаллические структуры могут быть закодированы различными способами, включая объекты (часто называемые дескрипторами), 3D-графики с пространственной информацией и соответствующими атрибутами узлов и ребер. В случае дескрипторов основными требованиями являются: низкая вычислительная стоимость, различимость между подобными структурами и способность кодировать как пространственную, так и информационную, электронную структуру и композицию. Архитектурами нейронных сетей на основе графов установлено, что они эффективны для больших наборов данных о молекулах и материалах из-за субпространственного числа обучаемых параметров, в то время как подходы, основанные на дескрипторах, являются склонными к переобучению.

2. МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Чтобы решить проблему с данными, Ulissi et al. разработали набор данных OC20, состоящий из 1 281 040 плотностей. Релаксации функциональной теории (DFT) (~264 890 000 одиночные точечные оценки) по широкому спектру материалов, поверхностей, и адсорбаты (химия азота, углерода и кислорода). По сравнению с ранее опубликованными наборами данных о катализаторах набор данных OC20 позволяет нам обобщать различные промежуточные продукты, образующиеся в ходе реакций восстановления CO2 и N2. Помимо создания и обмена набором данных, авторы предлагают три связанные проблемы предметной области в качестве открытого соревнования: 1) предсказать энергию и силу для данного состояния (S2EF), 2) предсказать соседнее расслабленное состояние с учетом начального состояния (IS2RS) и 3) предсказывают релаксированную энергию адсорбции при заданном начальном состоянии (IS2RE). Набор данных разделен на обучающую, тестовую и валидационную выборку. Валидационная выборка моделирует химическое пространство, используемое при скрининге катализаторов, включающая в себя как адсорбаты, так и катализаторы, состав и структура которых ранее не была представлена в обучающей выборке. Для практических целей наиболее подходящей задачей является предсказание DFT-энергии связывания на основе эвристически сгенерированных структур катализатор-адсорбат.

Ядерные заряды и положения атомов не являются подходящим входным представлением атомистического, не вращательно или поступательно инвариантного по сравнению с уравнением Шредингера. В настоящее время во многих подходах используются представления объектов, специфичные для атомов или геометрии, а также архитектуры ML на основе ядра или нейронных сетей (NN). Недавние исследования сосредоточены на характеристике атомных систем в абстрактных представлениях, таких как квантово-механические свойства, полученные в результате расчетов электронной структуры с низкими затратами, и использовании новых методов графовых нейронных сетей для повышения эффективности обучения и обобщаемости.

Было разработано несколько методов ML для прогнозирования энергий корреляции высокого уровня (связанных кластеров) на основе квантово-механических характеристик на уровне среднего поля (например, теория ВЧ или ДПФ) расчет электронной структуры. Например, авторы OrbNet использовали архитектуру графовой нейронной сети для прогнозирования высококачественных энергий электронной структуры на основе характеристик, полученных с помощью недорогих/минимальных методов электронной структуры среднего поля.

Различные графовые нейронные сети недавно добились больших успехов в предсказании квантово-механических свойств молекул. В предыдущих работах MPNNS с соответствующими функциями сообщения, обновления и вывода продемонстрировали полезное индуктивное смещение для прогнозирования молекулярных свойств, превосходя несколько сильных базовых показателей и устраняя необходимость в сложном проектировании функций. Важной задачей является разработка MPNN, которые могут эффективно обобщаться на более крупные графы, чем те, которые появляются в обучающем наборе, или, по крайней мере, работать с контрольными показателями, предназначенными для выявления проблем с обобщением по размерам графа. Обобщение на большие размеры молекул кажется особенно сложным при использовании пространственной информации. Межатомные взаимодействия (ковалентные и нековалентные) являются определяющим фактором в каталитических процессах; в частности, их необходимо учитывать при моделировании энергии адсорбции. Поэтому мы считаем важным ввести дескрипторы, описывающие межатомные электростатические взаимодействия, в качестве атрибутов ребер. В то время как орбитально-орбитальные взаимодействия вместе с информацией об электронном состоянии атомов могут быть универсальными атрибутами узла. Также при выборе признаков важны несколько требований, таких как изотропность пространства, изометрия пространства, инвариантность относительно перестановки атомных индексов, непрерывность и вычислительная дешевизна. Из-за вычислительных ограничений мы не используем гибридные подходы DFT/GNN. В качестве матриц межатомных сил мы выбрали куло-

новскую матрицу в качестве простого дескриптора электростатического взаимодействия между ядрами и синусоидальными и Матрицы сумм Эвальда как потенциалы электростатических взаимодействий в периодических (или псевдопериодических) системах. Как представление матрицы орбитального поля (OFM), основанной на распределении электронов валентной оболочки центрального и окружающих атомов.

3. ОСНОВНЫЕ РЕЗУЛЬТАТЫ, ВЫВОДЫ

В нашей работе мы представили несколько графовых нейронных сетей (GNN), включая сверточные архитектуры и архитектуры передачи сообщений с физически информированными атрибутами узлов и ребер для атомистических систем. Мы демонстрируем улучшение прогнозов энергии адсорбции в наборе данных OC20 с использованием предложенной нами архитектуры с точки зрения средней абсолютной ошибки прогнозируемой энергии и энергии в пределах пороговых показателей. Предлагаемые архитектуры устойчивы к переобучению и могут быть использованы для прогнозирования экспериментальных и квантово-химических свойств широкого спектра материалов и молекул.

Мы представили две архитектуры GNN (EdgeUpdateNet и OFMNet) с атрибутами в виде элементов матриц межатомного взаимодействия (матрица Кулона, матрица суммы Эвальда, Синусоидальная матрица). Для узловых отпечатков пальцев мы используем элементы матрицы орбитального поля (OFM), однократного представления электронного состояния атомов с окружающими атомными орбиталями. Разработанные модели, использующие физически информированные отпечатки пальцев, показывают улучшение точности прогнозирования энергии адсорбции. Особенно значительные улучшения наблюдаются в примерах, не относящихся к предметной области, как для адсорбатов, так и для катализаторов. Возможности обобщения и экстраполяции предлагаемых архитектур также делают GNNS пригодными для использования при скрининге катализаторов в обширном химическом пространстве. Код с используемыми сценариями, моделями и конфигурациями доступен по адресу https://github.com/AI4Materials-lab/catgnns.

EDN: XVEAJS
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 73–78

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

СИСТЕМНЫЙ ПОДХОД К ИЗУЧЕНИЮ ГОСУДАРСТВЕННЫХ ПОЛИТИК И ПРОЦЕССОВ ФОРМИРОВАНИЯ ЭТИКИ ПРИМЕНЕНИЯ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ГЛОБАЛЬНЫЙ АТЛАС РЕГУЛИРОВАНИЯ

© 2022 г. Э. Г. Чаче[1,*], Р. И. Дремлюга[2], Н. А. Третьякова[3,4], А. В. Незнамов[1,3,4]

Представлено академиком РАН А.Л. Семеновым
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

В статье представлен краткий обзор научной работы по глубокому сравнительному анализу регуляторного опыта 31 юрисдикции и 16 международных организаций в сфере искусственного интеллекта, которая изложена в книге “Глобальный атлас регулирования искусственного интеллекта. Восточный вектор”. Данная книга является первым в мире системным исследованием, направленным на построение унифицированных критериев анализа правовых систем различных стран с точки зрения влияния регулирования на развитие технологий искусственного интеллекта в таких сферах, как стратегическое планирование, здравоохранение, беспилотный транспорт, персональные данные, государственное управление, этика.

Ключевые слова: искусственный интеллект, регулирование искусственного интеллекта, этика искусственного интеллекта, новые технологии, artificial intelligence, AI, emerging technologies, innovations, AI ethics

DOI: 10.31857/S2686954322070050

1. ВВЕДЕНИЕ

Эффективное и динамичное развитие новых технологий, искусственного интеллекта – одна из первостепенных задач отдельно взятых стран и всего мирового сообщества. Кто сегодня займет лидерские позиции в сфере искусственного интеллекта и новых технологий в целом, тот получит глобальное лидерство на мировой арене.

Вместе с тем безопасное развитие искусственного интеллекта и новых технологий в целом просто невозможно представить без грамотно выстроенной системы регулирования. Современное регулирование как национальное, так и наднациональное значительно отстает от темпов развития искусственного интеллекта. На первом месте стоит выработка подхода, который поможет технологии продолжать развиваться, не сбавляя темпов, а с другой стороны, сможет защитить человека. Практически в каждой национальной стратегии развития искусственного интеллекта отмечаются важность и необходимость построения взвешенной системы регулирования, которая помогла бы ликвидировать или же минимизировать существующие регуляторные барьеры без вреда и с учетом прав человека.

В настоящей статье предлагаем рассмотреть основные и самые интересные подходы к регулированию международных организаций на основании материала, представленного в книге “Глобальный атлас регулирования искусственного интеллекта. Восточный вектор” [1].

В данной работе были рассмотрены опыт 31 юрисдикции и 16 международных организаций в сфере искусственного интеллекта. В процессе подготовки было изучено более 500 нормативных документов разных стран и рассмотрено более 1200 источников в совокупности.

[1] *Центр регулирования Artificial Intelligence Сбербанка, Москва, Россия*
[2] *Институт математики и компьютерных наук (ИМКТ) Дальневосточного федерального университета, Владивосток, Россия*
[3] *Дальневосточный центр искусственного интеллекта Сбербанка, Владивосток, Россия*
[4] *Дальневосточный федеральный университет, Владивосток, Россия*
*****E-mail: EGChache@sberbank.ru*

2. ОСОБЕННОСТИ МЕЖДУНАРОДНОГО РЕГУЛИРОВАНИЯ (ОПЫТ МЕЖДУНАРОДНЫХ ОРГАНИЗАЦИЙ)

При рассмотрении вопроса регулирования искусственного интеллекта следует в первую очередь обратиться к деятельности, проводимой Советом Европы в отношении выработки единого подхода к регулированию. Формирование подходов к регулированию в Совете Европы занимается Специальный комитет Совета Европы по ИИ (CAHAI). Именно в полномочия CAHAI входит создание юридически обязывающего документа. Такой юридически обязывающий документ должен по итогу включать в себя общие принципы и конкретные правовые нормы, которые могли бы далее сопровождаться дополнительными отраслевыми документами [2]. Создаваемый документ должен быть направлен на предотвращение и/или снижение рисков, связанных с применением систем ИИ, которые потенциально могут помешать соблюдению прав человека, функционированию демократии и соблюдению верховенства закона, одновременно продвигая социально полезные приложения ИИ. Аналогичный подход к регулированию искусственного интеллекта отражен в подходе Европейского Союза, который будет рассмотрен далее. В 2022 г. Специальный (т.е. временный) комитет по ИИ (CAHAI) был трансформирован в постоянный комитет по ИИ (CAI). Его дальнейшая работа будет во многом основана на материалах, которые были подготовлены Специальным комитетом.

В 2019 г. 22 мая ОЭСР, объединяющая 36 стран, и шесть стран-партнеров (Аргентина, Бразилия, Колумбия, Коста-Рика, Перу и Румыния) договорились при разработке ИИ придерживаться общих стандартов, гарантирующих надежность, безопасность ИИ-систем. В этом же году страны утвердили первые принципы работы с технологиями ИИ на межправительственном уровне “Руководящие принципы ИИ ОЭСР” (OECD AI Principles) [3]. Принципы ИИ были представлены в составе документа “Рекомендации Совета ОЭСР по правовым инструментам ИИ” (Recommendation of the Council on OECD Legal Instruments Artificial Intelligence) [4]. Главная цель Руководящих принципов – способствовать эффективному развитию технологий ИИ, содействовать правительствам государств, разработчикам и частным лицам в этичной разработке и применении технологий ИИ. Руководящие принципы ИИ впервые закрепили 5 главных рекомендаций в отношении создания этичного и недискриминационного ИИ:

- ИИ должен служить людям и планете, повышать недискриминационный рост, устойчивое социально-экономическое развитие;
- технологии ИИ должны разрабатываться с учетом требований применимого законодательства, обеспечения и защиты прав человека, демократических ценностей и многообразия;
- работа ИИ должна быть прозрачной и понятной для людей;
- технологии ИИ должны работать надежно и безопасно в течение всего жизненного цикла; такие технологии должны постоянно оцениваться на предмет потенциального риска причинения вреда человеку;
- организации и люди, которые разрабатывают, внедряют, управляют технологии ИИ, должны нести ответственность за адекватное и точное функционирование таких технологий.

Данными принципами в своей деятельности руководствуются многие страны и международные организации. Например, Глобальное партнерство по ИИ (GPAI) (The Global Partnership on Artificial Intelligence (GPAI)) [5], которое является инициативой разработанной в рамках G7, свою деятельность основывают на “Руководящих принципах ИИ ОЭСР”.

В 2021 г. в рамках 41-й сессии Генеральной конференции Организации Объединенных Наций по вопросам образования, науки и культуры (ЮНЕСКО) 193 страны пришли к консенсусу относительно общих принципов этики ИИ и приняли исторический документ – первое глобальное соглашение по этике ИИ. Текст Рекомендаций по этике ИИ получился достаточно объемным и включает в себя 141 пункт. Стоит отметить, что в среднем корпоративные этические документы в сфере ИИ сформулированы в виде 5–7 конкретных принципов, а например, один из первых международных документов – Рекомендации Совета ОЭСР по ИИ [6] – имеет чуть больше 20 пунктов. Одна из главных целей Рекомендации – предоставить универсальную базу принципов, которыми государства могли бы руководствоваться при формулировании своего законодательства. Можно выделить следующие принципы: соразмерность и непричинение вреда; безопасность и защищенность; справедливость и недискриминация; устойчивость; неприкосновенность частной жизни и защита данных; подконтрольность и подчиненность человеку; прозрачность и объяснимость; ответственность и подотчетность; осведомленность и грамотность; многостороннее и адаптивное управление и взаимодействие.

Россия была одной из активных участниц процесса разработки Рекомендации по этике ИИ (Recommendation on the Ethics of Artificial Intelligence) [7] и инициировала обсуждения целого ряда вопросов в этой сфере.

Рассмотренные документы Совета Европы, ОЭСР, ЮНЕСКО не являются исчерпывающи-

ми. БРИКС, СНГ, ШОС, АТЭС, G7, G20, B20 сегодня находятся на пути к созданию унифицированного подхода к модельному регулированию ИИ. Так, ШОС и СНГ уже активно обсуждают проекты документов по регулированию ИИ в соответствующих регионах. Отраслевые международные организации, такие как ОБСЕ, ВОИС, ВОЗ, МОТ, также подсвечивают вопросы регулирования ИИ в своих стратегических документах.

3. ОСОБЕННОСТИ РЕГУЛИРОВАНИЯ ОТДЕЛЬНЫХ ЮРИСДИКЦИЙ

В исследовательской работе были предложены унифицированные критерии анализа юрисдикций.

Стратегическое планирование – то, к чему стремятся все страны без исключений. Согласно информации представленной на OECD.AI *(платформа посвящена инновациям в сфере ИИ и опирается на Рекомендации Совета ОЭСР по правовым инструментам ИИ, как на международный стандарт в области ИИ)* более 80% стран мира сегодня разработали/разрабатывают стратегические инициативы в сфере ИИ или цифровизации. Такие документы покрывают стратегический подход к разработке и внедрению ИИ во всех отраслях экономики, определяют план влияния ИИ на повышения ВВП страны и т.д. Например, в России в 2019 г. была принята Национальная стратегия развития ИИ на период до 2030 г., утвержденная Указом Президента РФ от 10 октября 2019 г. № 490 “О развитии ИИ в Российской Федерации” [8], а ориентиры для нормотворчества закреплены в Концепции развития регулирования отношений в сфере технологий ИИ и робототехники на период до 2024 г. [9], утвержденной Распоряжением Правительства РФ от 19 августа 2020 г. № 2129-р “Об утверждении Концепции развития регулирования отношений в сфере технологий ИИ и робототехники на период до 2024 г.”, которая выстраивает определенное видение дальнейшей правовой судьбы технологии. В Китае (КНР) же еще в 2017 г. был опубликован План по развитию ИИ нового поколения (New Generation Artificial Intelligence Development Plan) [10]. С тех пор руководство КНР стало рассматривать ИИ как приоритетную и “стратегическую технологию, которая станет локомотивом новой научно-промышленной революции”. План по развитию ИИ нового поколения излагает руководящие принципы развития ИИ в Китае до 2030 г. Кроме того, такие провинции КНР, как Гонконг, Тайвань также имеют свои стратегические инициативы по развитию ИИ в регионе.

Для реализации стратегических задач в сфере ИИ создаются национальные и наднациональные ассоциации/альянсы в сфере ИИ. Так, например, в России существует Альянс в сфере ИИ, а в Нидерландах Альянс ИИ (Alliance for Artificial Intelligence) и Голландская ассоциация ИИ и правам роботов (Dutch Association for AI and Robot Rights, NVAIR) и т.д. Самые крупные наднациональные партнерства являются участниками Глобального Партнерства по ИИ (Global Partnership on AI) и The Partnership on AI.

С позиции **общего регулирования** наиболее интересным представляется опыт Канады и Европейского Союза. Так, в ЕС 21 апреля 2021 г. Европейская комиссия представила проект Закона об ИИ (AI ACT) [11]. Проект Закона фокусируется на риск-ориентированном подходе к ИИ и на устранении рисков, связанных с конкретным использованием ИИ, с разделением их на 4 различных уровня: неприемлемый риск, высокий риск, ограниченный риск и минимальный риск [12].

Канада, в свою очередь, стала одной из первых стран мира, которая предложила принять универсальный закон, посвященный ИИ: Закон об искусственном интеллекте и данных “the Artificial Intelligence and Data Act, AIDA” [13]. Данный закон является частью сборника законопроектов о реализации цифровой хартии 2022 (the Digital Charter Implementation Act 2022 (Bill C-27) [14]. Помимо AIDA в сборник законопроектов вошли: Закон о защите конфиденциальности потребителей (Consumer Privacy Protection Act), который вносит реформу в сферу защиты персональных данных, Закон о трибунале по защите личной информации и данных (Personal Information and Data Protection Tribunal). AIDA устанавливает общеканадские требования к проектированию, разработке, использованию и интеграции систем ИИ. AIDA может иметь экстерриториальное распространение, в случае если компоненты глобальных систем ИИ используются, разрабатываются, проектируются или управляются в Канаде.

Учитывая тот факт, что максимально эффективное обучение моделей зависит от **доступа к данным** у разработчиков ИИ-решений, то очень важно обеспечить безопасное регулирование доступа к данным.

Каждая страна сегодня стремится создать эффективные, актуальные открытые порталы данных, так называемые единые государственные датасеты для разработчиков ИИ-решений. Некоторые страны даже определяют уровни доступа к государственным датасетам: открытые, с ограниченным доступом (которые включают в себя критичные данные), закрытые. Такой подход, например, определен в ЕС. Регулирование данных ЕС самое обширное и состоит из следующих документов, помимо GDPR: Стратегия по данным 2020 (A European strategy for data) [15], Директива об открытых данных (Directive on open data and the re-use of public sector information) [16], проект За-

кона об управлении данными (Data Governance Act) [17].

Однако данные, которые имеют ценность сегодня для разработчиков, имеет ограниченный уровень доступа. Это персональные данные. Для предоставления доступа к таким данным очень важно выстроить институт анонимизации данных (необратимого обезличивания). Грамотный и эффективный подход выстроен в азиатских странах: Китай, Южная Корея, Япония, Сингапур и т.д. – в каждой стране есть методологии анонимизации данных, которые позволяют обеспечить анонимизацию данных таким образом, что их ценность для разработчиков ИИ-решений не теряется. Анонимизированные данные уже выходят из-под действия законов о персональных данных. Такой же подход и выработан в ЕС: анонимизированные данные перестают быть персональными данными в понимании Единого регламента по защите персональных данных (GDPR).

Если говорить об отраслевом использовании ИИ, то первое что приходит на ум современному человеку – **беспилотный транспорт**, а именно беспилотные автомобили. Лидером в сфере беспилотных автомобилей, согласно международному индексу KPMG 2020 Autonomous Vehicles Readiness Index, является Сингапур.

Сингапур как азиатский технологический лидер одним из первых принял правовые акты, направленные на регулирование беспилотного автотранспорта – Правила дорожного движения (автономные автомобили) стал применяться с августа 2017 г. (Road Traffic (Autonomous Motor Vehicles) Rules 2017) [18]. Еще одним критическим требованием является присутствие в беспилотном автомобиле квалифицированного оператора безопасности. Предполагается, что в случае необходимости он возьмет управление автомобилем на себя. В исключительных случаях автономный автотранспорт может работать без оператора, но только в том случае, если доказана безопасность подобного использования.

В США регулирование беспилотного транспорта осуществляется как на федеральном уровне, так и на уровне штатов. На федеральном уровне действует Федеральная политика в области автономных транспортных средств от 2016 г. (Federal Automated Vehicles Policy), которая ежегодно дополняется новыми положениями [19, 20]. На региональном же уровне действует Типовая модель управления (The Model Policy State), которая регламентирует полномочия в регулировании автономных автомобилей, которыми обладают штаты.

Еще одна отрасль, в которой важно оценить регулирование технологий ИИ – **здравоохранение**. По мнению многих экспертов, именно данная отрасль является одной из самых перспективных с точки зрения интеграции ИИ-технологий. Повсеместно разрабатываются модели ИИ для компьютерной томографии легких, головного мозга, а также сиптомчекеры и многое другое, что способно вывести современную медицину на новый уровень.

Когда мы говорим о регулировании ИИ-технологий в здравоохранении, принимаем во внимание следующие аспекты: а) наличие правил ввода систем ИИ для использования в медицинской практике по правилам регистрации медицинских изделий, если ИИ прямо выделено как медицинское изделие, наличие специальных правил регистрации систем ИИ, в т.ч. стандартов безопасности, правил проведения клинических испытаний, правил проведения технических испытаний, правил регистрации; б) нормативная скорость регистрации систем ИИ для ввода в оборот (независимо от того, специальные или общие правила); в) наличие регуляторных условий для создания медицинских дата сетов (включая правила и методики обезличивания и анонимизации).

Хотелось бы рассмотреть уникальный пример Королевства Саудовской Аравии. Именно эта страна первая в мире выпустила Руководство по медицинским устройствам на основе ИИ и больших данных (Guidance on AI and Big Data-based medical devices) [21]. В Руководстве изложены требования для получения разрешения на медицинские изделия на основе ИИ. Кроме того, в Руководстве отмечено, что такие медицинские изделия относятся к автономному программному обеспечению медицинских изделий, которые используют технологию ИИ на основе машинного обучения для диагностики, управления или прогнозировать заболевания, анализируя медицинские большие данные [22]. Кроме того, Министерство здравоохранения Саудовской Аравии активно работает в направлении создания медицинских датасетов. Сегодня создано более 800 датасетов с медицинскими данными для исследований. Получить такие данные может любой разработчик ИИ [23].

ИИ в здравоохранении, беспилотный транспорт – самые крупные и известные сферы, но не единственные. Сегодня ИИ-технологии внедряются повсеместно: государственное управление, сельское хозяйство, нефтяная промышленность, электронная коммерция, смарт-контракты, умный дом и многое другое. Поскольку ИИ-технологии вокруг нас и интегрировались во все сферы нашей с вами жизни, просто необходимо, чтобы разработка и применение таких технологий были максимально этичны.

Этика – самый обсуждаемый и важнейший элемент развития ИИ-технологий. Этические аспекты появляются на всех стадиях жизненного

Рис. 1. Иллюстрация “Город будущего & новые технологии” выполнена нейронной сетью Сбербанка Kandinsky.”

цикла модели ИИ: от идеи создания модели до ее применения и использования человеком. Сегодня вопросы этики отнесены к “мягкому регулированию”: добровольное принятие на себя этических принципов применения ИИ.

В каждой стране, в каждой компании принципы этики ИИ формируются абсолютно поразному, но везде соблюдаются базовые аспекты: контролируемость и управляемость систем ИИ; прозрачность и предсказуемость функционирования; стабильность и надежность систем ИИ; ответственное применение ИИ; непредвзятый и недискриминирующий ИИ. Опыт России в сфере этики ИИ можно назвать одним из самых удачных. В России разработан и принят Кодекс этики в сфере ИИ [24]. Кодекс служит ориентиром для развития технологий ИИ в стране и призван обеспечивать доверие к ИИ со стороны пользователей, общества и государства. По состоянию на ноябрь 2022 г. к Кодексу присоединилось более 110 участников рынка. В России также есть компании, кто принял свои корпоративные принципы этики ИИ: Сбер [25], Яндекс [26] и ABBYY [27].

Этика и грамотно выстроенное правовое регулирование помогут динамично развиваться технологиям ИИ, не ограничивая и не блокируя прогресс и пользу, которую технология может принести человечеству. Важно, чтобы правовое регулирование такой узкой сферы было своевременным, эффективным и стало триггером развития ИИ и обеспечения прав и свобод человека. Поиском такого баланса сегодня и занимаются эксперты разных отраслей со всего мира. Самый инновационный и интересный опыт стран детально рассмотрен в научном исследовании “Глобальный атлас регулирования ИИ. Восточный вектор”, краткий обзор которого был приведен выше.

СПИСОК ЛИТЕРАТУРЫ

1. Глобальный атлас регулирования искусственного интеллекта. Восточный вектор / под. ред. А.В. Незнамова. М.: Альпина ПРО, 2022. 288 с.
2. Подробнее о деятельности документа см. на сайте спецкомитета: CAHAI - Ad hoc Committee on Artificial Intelligence [Электронный ресурс]. – Режим доступа: https://www.coe.int/en/web/artificial-intelligence/cahai (Дата обращения: 19.10.2022). Основные положения о проекте будущего регулирования доступны в ряде документов, например, Ad hoc Committee on Artificial Intelligence (CAHAI) [Электронный ресурс]. – Режим доступа: https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=09000016809ed062 (Дата обращения: 19.10.2022), Ad Hoc Committee On Artificial Intelligence (CAHAI) [Электронный ресурс]. – Режим доступа: https://rm.coe.int/cahai-lfg-2021-pv5-en-5th-meeting-report/1680a46f5f (Дата обращения:19.10.2022), Ad Hoc Committee On Artificial Intelligence (CAHAI) Policy Development Group (CAHAI-PDG) [Электронный ресурс]. – Режим доступа: https://rm.coe.int/cahai-pdg-2021-pv4-meeting-report-6th-meeting/1680a45412 (Дата обращения: 19.10.2022), Ad Hoc Committee On Artificial Intelligence (CAHAI) [Электронный ресурс]. – Режим доступа: https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da (Дата обращения: 19.10.2022), Ad Hoc Committee On Artificial Intelligence (CAHAI) [Электронный ресурс]. – Режим доступа: https://rm.coe.int/cahai-lfg-2021-pv4-en-4th-meeting-report/1680a449bc (Дата обращения: 19.10.2022), https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680a4e8a5 и других.
3. OECD AI Principles overview [Электронный ресурс]. – Режим доступа: https://oecd.ai/en/ai-principles (дата обращения: 19.10.2022)
4. Recommendation of the Council on Artificial Intelligence [Электронный ресурс]. – Режим доступа: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (дата обращения: 19.10.2022)
5. The Global Partnership on Artificial Intelligence [Электронный ресурс]. – Режим доступа: Global Partnership on Artificial Intelligence - GPAI (дата обращения: 19.10.2022)
6. Forty-two countries adopt new OECD Principles on Artificial Intelligence [Электронный ресурс]. – Режим доступа: https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm (Дата обращения: 19.10.2022).

7. Recommendation on the ethics of artificial intelligence [Электронный ресурс]. – Режим доступа: https://en.unesco.org/artificial-intelligence/ethics (Дата обращения: 19.10.2022).

8. Указ Президента РФ от 10 октября 2019 г. № 490 "О развитии искусственного интеллекта в Российской Федерации" [Электронный ресурс]. – Режим доступа: https://www.garant.ru/products/ipo/prime/doc/72738946/ (дата обращения: 19.10.2022).

9. Распоряжение Правительства РФ от 19 августа 2020 г. № 2129-р "Об утверждении Концепции развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники на период до 2024 г" [Электронный ресурс]. – Режим доступа: https://www.garant.ru/products/ipo/prime/doc/74460628/ (дата обращения: 19.10.2022).

10. New Generation Artificial Intelligence Development Plan [Электронный ресурс]. – Режим доступа: http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm (дата обращения: 19.10.2022)

11. AI ACT [Электронный ресурс]. – Режим доступа: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206 (дата обращения: 19.10.2022)

12. Proposal for a Regulation laying down harmonised rules on artificial intelligence [Электронный ресурс]. – Режим доступа: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (дата обращения: 19.10.2022)

13. Canada's artificial intelligence legislation is here [Электронный ресурс]. – Режим доступа: https://www.dataprotectionreport.com/2022/06/canadas-artificial-intelligence-legislation-is-here/(дата обращения: 19.10.2022).

14. the Digital Charter Implementation Act 2022 (Bill C-27 [Электронный ресурс]. – Режим доступа: https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading (дата обращения: 19.10.2022).

15. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: a European strategy for data [Электронный ресурс]. – Режим доступа: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066 (дата обращения: 16.07.2022)

16. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information [Электронный ресурс]. – Режим доступа: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1561563110433&uri=CELEX:32019L1024 (дата обращения: 16.07.2022)

17. Proposal for Data Governance Act [Электронный ресурс]. – Режим доступа: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PC0767 (дата обращения: 16.07.2022)

18. Road Traffic Act (Chapter 276), Road Traffic (Autonomous Motor Vehicles) rules 2017 [Электронный ресурс]. – Режим доступа: https://sso.agc.gov.sg/SL/RTA1961-S464-2017 (дата обращения: 19.10.2022)

19. Federal Automated Vehicles Policy - September 2016 [Электронный ресурс]. – Режим доступа: https://www.transportation.gov/AV/federal-automated-vehicles-policy-september-2016 (дата обращения: 19.10.2022)

20. Ensuring American Leadership in Automated Vehicle Technologies [Электронный ресурс]. – Режим доступа: https://www.transportation.gov/sites/dot.gov/files/2020-02/EnsuringAmericanLeadershipAVTech4.pdf (дата обращения: 19.10.2022)

21. Guidance on Review and Approval of Artificial Intelligence (AI) and Big Data based Medical Devices [Электронный ресурс]. – Режим доступа: https://beta.sfda.gov.sa/sites/default/files/2021-04/SFDAArtificial%20IntelligenceEn.pdf (дата обращения: 19.10.2022).

22. Руководство по медицинским устройствам на основе ИИ и больших данных [Электронный ресурс]. – Режим доступа: https://beta.sfda.gov.sa/sites/default/files/2021-04/SFDAArtificial%20IntelligenceEn.pdf (дата обращения: 19.10.2022).

23. Open Data: Ministry of Health [Электронный ресурс]. – Режим доступа: https://data.gov.sa/Data/en/organization/ministry_of_health (дата обращения: 19.10.2022).

24. Кодекс этики в сфере искусственного интеллекта [Электронный ресурс]. – Режим доступа: https://a-ai.ru/wp-content/uploads/2021/10/Кодекс_этики_в_сфере_ИИ_финальный.pdf (дата обращения: 19.09.2022)

25. Принципы этики искусственного интеллекта Сбера [Электронный ресурс]. – Режим доступа: https://www.sberbank.com/ru/sustainability/principles-of-artificial-intelligence-ethics (дата обращения: 19.10.2022)

26. Принципы Яндекса [Электронный ресурс]. – Режим доступа: https://yandex.ru/company/values (дата обращения: 19.10.2022)

27. Подход ABBYY к принципам создания надежного искусственного интеллекта [Электронный ресурс]. – Режим доступа: https://www.abbyy.com/ru/company/trustworthy-ai/ (дата обращения: 19.10.2022)

EDN: EFXSUY
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 79–87

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ПЛАНИРОВАНИЕ РАСПИСАНИЙ В МУЛЬТИАГЕНТНЫХ СИСТЕМАХ НА БАЗЕ МЕТОДА ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

© 2022 г. И. К. Минашина[1], Р. А. Горбачев[1], Е. М. Захарова[1,*]

Представлено академиком РАН А.Л. Семеновым
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

Статья посвящена решению задачи планирования расписаний в мультиагентных системах в рамках конкурса Flatland 3. Основная цель конкурса – разработать алгоритм эффективного управления плотным движением на сложных железнодорожных сетях в соответствии с заданным графиком движения. Предложенное решение основано на использовании метода обучения с подкреплением (Reinforcement Learning). Для его адаптации к специфике задачи был разработан новый подход, основанный на методике структурирования вознаграждения, стимулирующий агента следовать своему расписанию. Архитектура предлагаемой модели основана на многоагентной вариации централизованного критика с обучением по типу Proximal Policy Optimization (PPO). Кроме того, была разработана и реализована стратегия обучения по расписанию. Это позволило агенту вовремя справляться с каждым уровнем сложности и тренировать модель в более сложных условиях. Данное решение заняло первое место в конкурсе Flatland 3 в треке Reinforcement Learning.

Ключевые слова: обучение с подкреплением, мультиагентные системы, железные дороги, Flatland, структурирование функции вознаграждений, обучение по расписанию, централизованный критик

DOI: 10.31857/S2686954322070177

1. ВВЕДЕНИЕ

Высокие темпы индустриализации в современном мире способствуют повышению объемов перевозок. Данная проблема особенно остро стоит в области железнодорожных перевозок, т.к. изменение ее инфраструктуры достаточно трудозатратно и возникает задача оптимального использования уже имеющихся ресурсов. Увеличивается плотность перевозок как в грузовом, так и в пассажирском движениях. Вследствие этого повышаются требования к исполнению запланированного плана движения, и любое отклонение от него может привести к значительным неустойкам, например, увеличению задержки поездов, их отмене [1]. Поэтому разработка систем управления для данной области является актуальным и перспективным направлением. В данной сфере существуют высокие требования к безопасности движения, которые приводят к дополнительным трудностям и ограничениям. Вследствие этого задача эффективного управления железнодорожным трафиком становится крайне сложной. С данной проблемой сталкиваются все транспортные и логистические компании по всему миру.

Одними из тех, кто задался целью создать в этой области наиболее эффективное решение, которое будет основано на применении новейших подходов, в том числе и искусственного интеллекта, стали Швейцарские федеральные железные дороги и Deutsche Bahn AG. Для этой цели они организовали соревнования Flatland на платформе AICrowd. В данном соревновании участникам предоставляется симулятор, моделирующий процессы движения поездов и работу железнодорожной инфраструктуры, для проведения экспериментов по апробации реализованных алгоритмов. В данном соревновании есть два трека – один для решений, основанных на использовании классических алгоритмов, другой – основанных на использовании мультиагентного обучения с подкреплением [1].

В третьей версии симулятора, предназначенного для Flatland версии 3 (конкурс 2019 г.) был добавлен функционал, который также проверяет пунктуальность и точность следования заданному графику [2]. Для каждого поезда существует заданное расписание с конкретными временами прибытия и отправления на станциях, а также окно времени, в течение которого они должны стартовать и добраться до места назначения. Расписа-

[1] *Московский физико-технический институт (национальный исследовательский университет), Москва, Россия*

*****E-mail: zakharova.em@mipt.ru*

ние составлено таким образом, чтобы у каждого поезда было больше времени, чем теоретически необходимо, для прибытия в пункт назначения. Поэтому необходимо использовать этот запас времени таким образом, чтобы все поезда прибывали с минимальной задержкой, например, пропуская другие поезда или уступая дорогу.

Цель соревнования состоит в том, чтобы реализовать алгоритм построения расписания движения поездов, в котором все поезда прибудут в пункт назначения с минимальной задержкой по отношению к требуемому времени прибытия.

Проблема построения эффективных расписаний является одной из самых сложных проблем в области планирования и управления железнодорожным транспортом. При ее решении необходимо учитывать различные аспекты организации перевозочного процесса [3]. Эта задача может быть сформулирована как классическая задача исследования операций (operation research; OR) и как задача обучения с подкреплением. Решение, представленное в данной статье, было предложено для участия в конкурсе Flatland 3 в треке Reinforcement Learning, где заняло первое место [4].

Поиск решения в данной области связан с проблемой принятия решений при появлении других агентов на пути, стохастического характера поломок и реакцией на разреженную функцию вознаграждения, заданную в симуляторе. Данная проблема связана с особенностями мультиагентного обучения и основным инструментом ее решения является использование методов обучения с подкреплением с централизованным критиком [5]. Однако введенная новая концепция расписаний создает особые трудности при обучении и придает чувствительность конструированию функции вознаграждения. В связи с этим реализованный подход основан на технике структурирования вознаграждения, а также специально разработанной стратегии обучения по расписанию. Распространенный способ адаптации к новым особенностям задачи заключается в преобразовании новых знаний предметной области в дополнительные вознаграждения и обучении агентов с помощью комбинации оригинальных и новых вознаграждений [6, 7]. В данном случае была разработана дополнительная компонента, отражавшая степень отставания поезда от его расписания.

Результаты показали, что разработанная стратегия обучения по расписанию способна вовремя справляться с каждым уровнем сложности, что дало возможность обучить модель в более сложных средах. Предложенная гибридная структура вознаграждения отвечала новым аспектам задачи и позволила эффективно преодолеть трудности нового издания конкурса Flatland 3.

2. ОСОБЕННОСТИ СИМУЛЯТОРА

Рассмотрим особенности среды Flatland 3 для управления в сфере транспортных коммуникаций [8].

Среда

Flatland – это симулятор, моделирующий динамику движения поездов, а также железнодорожную инфраструктуру. Данный симулятор генерирует уровни, которые представляют собой двумерную сетку, где каждая клетка имеет свой тип: поворот, дорога, развилка или недоступная местность. Каждый поезд занимает одну клетку на сетке и имеет цель и направление. Как и в реальных железнодорожных сетях поезда не движутся с одинаковой скоростью, а имеют разную заданную скорость передвижения в соответствии с типом поезда. Например, грузовой поезд будет двигаться медленнее, чем пассажирский поезд и в связи с этим необходимо избегать планирования быстрого поезда за медленным. Также каждый поезд имеет свое временное окно, в течение которого он должен стартовать и прибыть в пункт назначения. В случае столкновения поездов возникает пробка.

Наблюдения

Со стороны агента Flatland предоставляет доступ практически ко всей информации о текущем состоянии среды, на основе которой можно построить свой вид наблюдений. В симуляторе уже реализованы 3 вида наблюдений [8]:

– глобальное наблюдение – самое простое. В этом случае каждому агенту предоставляется глобальный обзор всей среды;

– наблюдение по локальной сетке – похоже на глобальное, однако размеры наблюдаемого окружения ограничены;

– “Tree observation” – основано на том факте, что сеть железных дорог является графом, и поэтому наблюдение строится только вдоль разрешенных переходов в графе. Наблюдение создается путем охвата 4-х разветвленного дерева от текущей позиции агента. Каждая ветвь следует разрешенным переходам (обратная ветвь разрешена только в тупиках) до тех пор, пока не будет достигнута ячейка с несколькими разрешенными переходами. Здесь информация, собранная по ветке, сохраняется в виде узла в дереве.

Действия

Симуляция происходит пошагово. Каждый агент поезда на каждом шаге должен определить, какое действие ему совершить. Поезда во Flatland имеют весьма ограниченный набор действий, как

и следовало ожидать от симулятора железной дороги. Это означает, что допустимы только несколько действий, а именно: повернуть налево, повернуть направо, двигаться вперед, продолжать предыдущее действие или остановиться.

К этому стоит добавить то, что с заданной частотой симулятор имитирует поломки. Непредвиденные события часто встречаются на железнодорожных сетях. Первоначальный план необходимо перепланировать в реальном времени, поскольку незначительные события, такие как задержка отправления с железнодорожных станций, поломки поездов или инфраструктуры, или даже проблемы с погодой, приводят к опозданию поездов. Механизм возникновения поломок реализован с использованием процесса Пуассона, а именно задержка имитируется путем остановки агентов в случайное время на случайные промежутки времени. Поезд с такой неисправностью не может двигаться в течение некоторого количества шагов симуляции, в результате чего он блокирует следующие за ними поезда, что часто приводит к пробкам.

Метрика поведения агента

Во Flatland 3 отклик вознаграждения предусматривается только в конце эпизода, что делает заданную по умолчанию функцию вознаграждения разреженной. Решения оцениваются по суммарному отклонению поездов от графика.

Эпизод заканчивается, когда все поезда достигают своей цели или при достижении максимального количества временных шагов. В конце эпизода для каждого поезда могут быть следующие варианты:

1. Поезд приехал в пункт назначения:

o 0 – если приехал вовремя или раньше;

o $-\min(t_l - t_a, 0)$ – если опоздал, где t_l указывает шаг моделирования до или на котором ожидается, что агент достигнет пункта назначения, а t_a – реальное время прибытия агента.

2. Поезд не достиг своей цели. Если поезд вообще не прибывает к месту назначения, то в соответствии с общими принципами работы железных дорог учитываются дополнительные штрафы. Тогда вознаграждение отрицательное и пропорционально опозданию поезда, а также расчетному количеству времени, необходимому для достижения цели агента с его текущей позиции по кратчайшему пути (t_{sp}):

$$t_l - t_a - t_{sp}.$$

3. Поезд так и не тронулся. В данном случае поезд считается отмененным, и предоставляется следующее вознаграждение:

(-1)* *cancellation_factor* * (t_{sp} + *cancellation_time_buffer*),

где cancellation_factor и cancellation_time_buffer – специальные параметры, настраиваемые организаторами конкурса.

Данная структура вознаграждения создает особые трудности при обучении. Рассмотрим их и методы их преодоления подробнее.

3. ЗАДАЧА И ЕЕ ОСОБЕННОСТИ

Задача соревнования состоит в наборе максимального количества очков по двум критериям. Во-первых, основываясь на суммарной функции вознаграждения, набранной всеми агентами. Во-вторых, показав наилучший процент агентов, благополучно доехавших до станции назначения. При этом оценка решения устроена так, что с каждым успешно завешенным уровнем структура следующего уровня железнодорожной сети усложняется и число поездов увеличивается, поэтому агентам приходится решать задачи все большего масштаба. С каждым раундом сложность первоначальной среды также увеличивался, а кроме этого, добавляется новое дополнительное требование, например, скоростной режим. И главная задача третьего издания конкурса состоит в том, чтобы составить наилучшее расписание, при котором все поезда прибывают в пункт назначения с минимальной задержкой.

Задачу также усложняют непредсказуемые задержки при сбоях в работе поездов. Поломки заставляют агентов оперативно изменять свои планы, что может привести к негативным последствиям, что в свою очередь отражается на сложности обучения модели RL.

Рассмотрим последствия, к которым приводят упомянутые трудности задачи.

Основной проблемой являются пробки – плохо обученная модель не справляется с многоагентной маршрутизацией поездов. Однако и более успешные модели могут не решить проблему пробок из-за сопутствующих факторов, например, поломки или большого количества одновременно стартующих поездов.

Другой проблемой является тенденция скорых поездов следовать за более медленными по удобной им дороге, ухудшая таким образом оптимальность конечного решения.

Наконец, самой глобальной и сложной проблемой оказалась уже упомянутая структура функции вознаграждения, которая во многих аспектах мешала обучению. Она могла привести, например, к потере ориентира, проезда мимо цели или намеренному попаданию в пробку.

Рассмотрим предложенную модель решения поставленной задачи и ее сопутствующих.

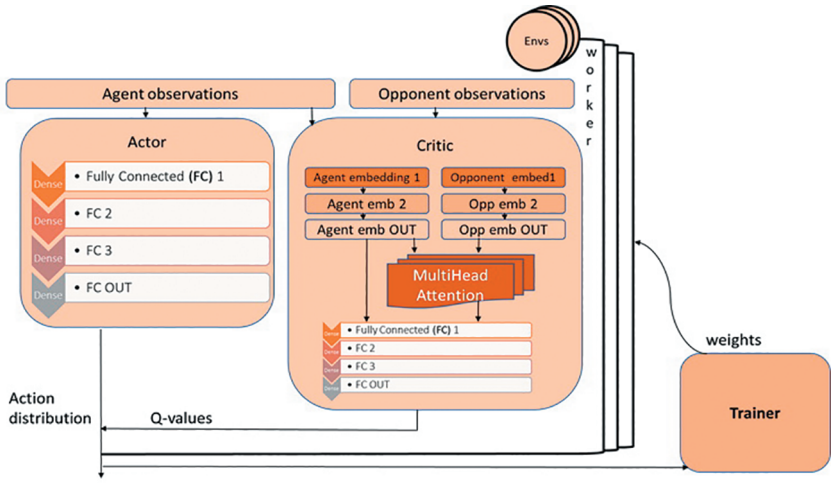

Рис. 1. Архитектура модели.

4. АРХИТЕКТУРА МОДЕЛИ

Базовая архитектура предложенной модели представляет собой полносвязную сеть с обучением по типу Proximal Policy Optimization (PPO) [9]. В режиме выполнения это обеспечивает быстроту принятия решения агентом. При этом во время обучения в дополнение к первой используется другая, более сложная сеть, которая дает свой вклад в градиенты для обучения сети агента. Этому агенту, называемому “центральным критиком”, предоставляется более целостное представление о состоянии среды, с помощью которого он аппроксимирует адекватный отклик на то или иное действие агента в текущей ситуации.

Существует множество вариаций метода централизованного критика [5, 10, 11]. Реализованный вариант не использует глобальные наблюдения, но объединяет наблюдения обучаемого агента с наблюдениями других агентов. Такой подход был выбран для того, чтобы создать решение, легко расширяемое для масштабных сред. Еще одним дополнением в мультиагентный алгоритм PPO с централизованным критиком является то, что в архитектуру критика добавлен трансформер, который позволяет критику эффективно объединять представления своих и чужих наблюдений [12]. Это позволяет работать с разреженными наблюдениями и неоднородностью данных, а также с возможной инвариантностью перестановок в случае изменения подмножества агентов, включенных в наблюдение [8].

На рис. 1 представлена архитектура модели, состоящей из Актора, который прогнозирует политику агента, т.е. вероятностное распределение действий в данном состоянии, и Критика, который выдает аппроксимацию вознаграждения среды на то или иное действие агента (Q-value). Далее модель обучается, чтобы улучшить выгоду (Advantage) от данного действия. Для обеспечения устойчивости модели использовалось сразу несколько параллельных потоков, которые одновременно собирали опыт для объекта Trainer’а, при этом каждый взаимодействовал сразу с несколькими средами.

В ходе экспериментов была также реализована альтернативная модель с трансформатором, перемещенным в Актора. Здесь основное наполнение было помещено в Актора. Он получал информацию о наблюдениях ближайших соседей агента и, преобразуя их в векторные представления (embeddings) вместе с собственными представлениями агента передавал в слой Multi Head Attention. Слой представлений агента был сконструирован таким образом, что его можно было сделать как разделенным для Критика и Актора, так и общим.

К сожалению, для настройки данной модели не было достаточно времени, поэтому она дала более слабые результаты и не была использована в окончательном решении. К тому же такой нагруженный Актор в режиме выполнения может достаточно долго проводить вычисления. В дальнейшем планируется исследование потенциала данной альтернативной архитектуры модели.

5. ОБЩАЯ СТРАТЕГИЯ РЕШЕНИЯ

Разработанные наблюдения

Для более результативного использования информации, получаемой от окружающей среды, предложенная симулятором система наблюдений “Tree observation” [8] была модифицирована. Основным нововведением стал учет особенностей работы с расписаниями движения поездов. Информация, хранящаяся в узле, была расширена – помимо основной информации о маршруте поез-

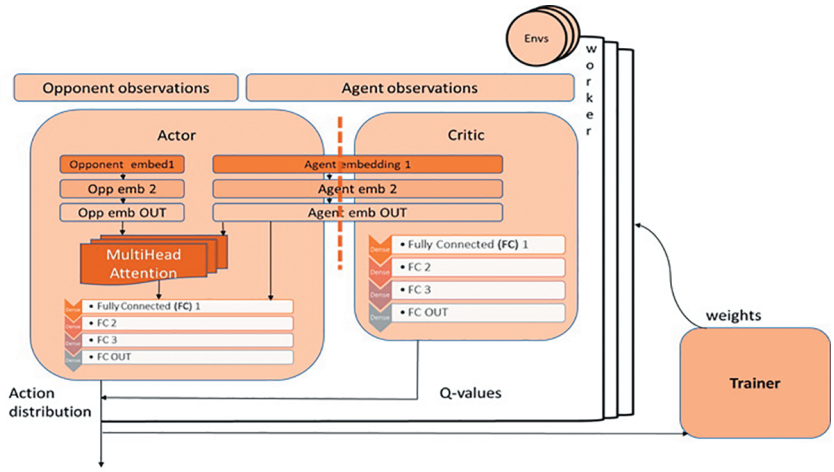

Рис. 2. Альтернативная модель.

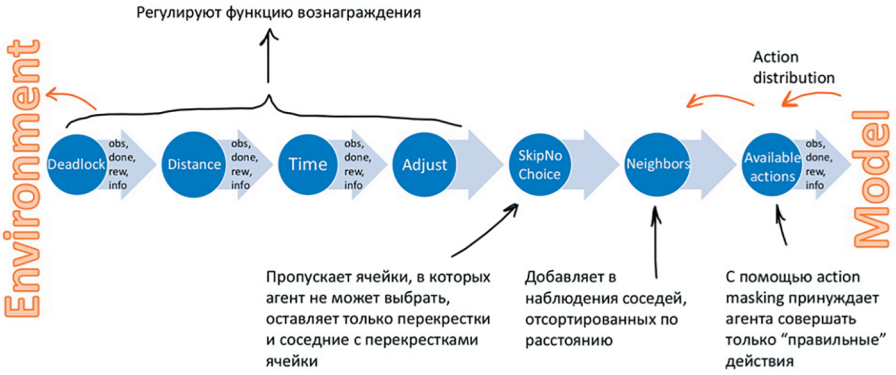

Рис. 3. Обертки над средой.

да, его скорости, инфраструктуре сети и других поездах, встречающихся на пути, была добавлена информация о временном отклонении от графика, сравнение индекса с другими поездами на пути, а также о потенциальных конфликтах на рассматриваемой ветке движения.

Структурирование вознаграждений

Для повышения эффективности разработанного алгоритма, а также для борьбы с описанными выше трудностями был изменен процесс взаимодействия агента с симулятором. Большинство оберток над средой (рис. 3) были разработаны для настройки функции вознаграждения, но некоторые из них также были призваны облегчить процесс обучения. Так, одним из таких значимых изменений была фильтрация обучаемых данных с акцентом на положения агента на или вблизи перекрестков. Также была применена техника "action masking", помогающая избавляться от недопустимых действий при обучении [8].

Рассмотрим последние 4 обертки, разработанные для регулировки функции вознаграждения. Она претерпела значительные изменения. Были разработаны дополнительные компоненты, которые позволяли справиться с последствиями ее разреженной структуры, стимулировали агента, достигшего цели, и наказывали в обратной ситуации.

В предыдущих версиях соревнования функция вознаграждения не была разреженной, т.е. на каждом шаге агент получал пенальти за то, что он еще не доехал цели, кроме этого, на концах также давалось награда за общее положение дел. В этот раз организаторы усложнили задание главным образом для того, чтобы ввести концепцию расписания в общую постановку задачи. Поэтому модель не обучалась в исходном виде (WITHOUT-ALL). Тогда

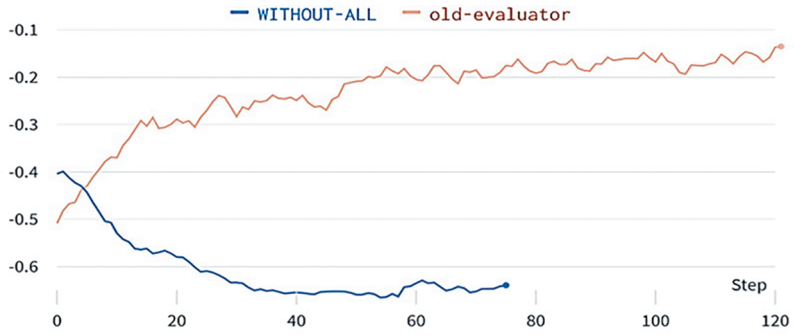

Рис. 4. Различия в обучении модели для разных версий оценки решений во Flatland.

как для предыдущей структуры вознаграждения (old-evaluator) та же самая модель давала неплохие результаты (рис. 4).

Поэтому структура функции вознаграждения была сильно модифицирована. В общем виде формулу вознаграждения можно представить в виде

$$r' = r + F,$$

где r – исходная функция вознаграждения, F – формирующая функция вознаграждения, r' – модифицированная функция вознаграждения.

В базовых решениях к конкурсу предлагался терминальный вариант формирующей функции F_{DR}, нацеленный на обнаружение и борьбу с пробками. Предлагаемая компонента добавляла в функцию награды "штраф" за поведение, вызвавшее блокировку пути (рис. 3 Deadlock wrapper). Эта обертка идентифицирует агента в пробке, дает ему за это штраф и заканчивает для него эпизод. Но в сочетании с новой структурой вознаграждений она приводила к тому, что агенты вместо того, чтобы сторониться пробок, наоборот, стремились попасть в них. Это поведение агентов можно объяснить тем, что при больших наказаниях за опоздания легче всего было закончить эпизод пораньше и получить прибыль из-за того, что эпизод закончился раньше времени прибытия.

Поэтому данная идея была преобразована в более общий вид, а именно использовалась F_{AR} (рис. 3 Adjust wrapper), которая может быть представлена следующим образом:

$$F_{AR} = \begin{cases} r_{fin}, & \text{если агент закончил эпизод} \\ & \text{и достиг пункта назначения} \\ r_{notfin}, & \text{если агент закончил эпизод} \\ & \text{и не достиг пункта назначения} \\ 0, & \text{если агент не закончил эпизод} \\ & \text{(только нормализует вознаграждение r)}, \end{cases}$$

где F_{AR} – терминальный корректирующий компонент функции формирования вознаграждения F, r_{fin} – настраиваемое вознаграждение за завершение эпизода в нужном пункте назначения, r_{notfin} – настраиваемый штраф за завершение эпизода не на станции назначения агента.

Данная компонента в любом случае дожидалась конца эпизода и давала пенальти за пробку или неприбытие и дополнительное поощрение за достижение цели. Также она нормализовала пошаговое вознаграждение, на котором стоит остановиться более подробно.

Чтобы уменьшить последствия разреженной функции вознаграждения, заданной в симуляторе, дополнительно была введена пошаговая награда, отражающая степень отставания поезда от графика движения (рис. 3 Time wrapper). В общем виде она может быть представлена следующими формулами:

$$F_{TR} = \min(P_T, P_{max}),$$

$$P_T = \begin{cases} e^{d_{max}/(d_{max}+t_r)}, & \text{если} \quad t_r \succ -d_{max}, \\ P_{max}, & \text{если} \quad t_r \le -d_{max} \end{cases}$$

где F_{TR} – временная составляющая функции формирования вознаграждения, d_{max} – это заданная максимально возможная задержка агента по его расписанию; P_{max} – настраиваемый максимальный штраф; t_r – оставшееся время до прибытия по расписанию.

С одной стороны, этот компонент функции формирования вознаграждения соответствует новым требованиям учета расписания поезда. С другой – это напоминает пошаговое наказание из старой структуры вознаграждения Flatland за каждый временной шаг, сделанный в среде, пока есть достаточно времени до последнего прибытия агента. Когда время до последнего заезда истека-

ет, штраф сильно возрастает. Смысл ее отражает график на рис. 5.

Такая модификация структуры функции вознаграждения радикально преобразовала характер обучения модели в лучшую сторону. Она стимулировала агента находиться в пути не слишком долго и в то же время следовать заданному графику движения.

Для лучшего ориентира к цели была разработана компонента F_{DR} (рис. 3 Distance wrapper), которая добавляла награду за приближение агента к пункту назначения и наказывала в случае простоя:

$$F_{DR} = \sigma\Delta d - N_i * P_i,$$

где F_{DR} – путевая компонента формирующей функции вознаграждения, σ – коэффициент для нормализации путевой компоненты вознаграждения, Δd – дельта расстояния, пройденного агентом на текущем шаге, N_i – количество последних шагов простоя агента (в случае простоя), P_i – штраф за простой. Штраф за простой увеличивался с увеличением времени простоя. Этот вроде бы полезный Wrapper имеет один неприятный нюанс. Его нужно дозировать в минимальном количестве, чтобы у агентов не появилось нездоровое желание совсем не заканчивать эпизод, а как можно дольше “двигаться к цели”.

Таким образом, суммарная функция награды была сформирована следующим образом:

$$R = r + aF_{AR} + bF_{TR} + cF_{DR},$$

где a, b, c – коэффициенты для настройки совместного поведения, R – общая функция вознаграждения, r – исходная функция вознаграждения.

Стратегия обучения

Для постепенного обучения на более сложных уровнях симуляции была разработана стратегия обучения по расписанию (curriculum learning). Был сконструирован ряд сред с различным уровнем сложности. Каждый уровень добавлял в среду новую особенность задачи, будь то увеличение количества поломок, скоростные режимы или более сложное соотношение городов и агентов. Обучение начиналось с очень маленькой среды из 2 агентов и 2 городов, затем постепенно усложнялось в соответствии с правилом: если текущая среда не доросла до следующего уровня сложности, увеличивалось только соотношение агентов и городов, в противном случае производился переход на следующую по сложности среду.

Событие перехода на новую среду было основано на достижении следующих условий в отношении обучаемого процесса:

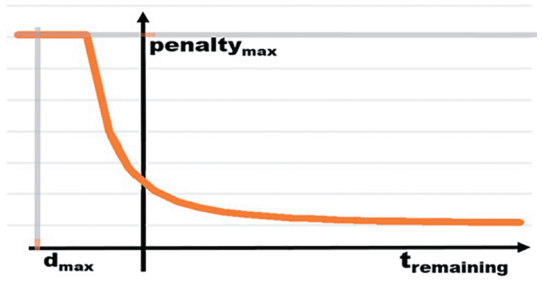

Рис. 5. Дополнительная временная компонента функции вознаграждения.

– с момента последнего переключения среды было произведено достаточное количество шагов обучения;

– производная линии регрессии вознаграждений не превышает заданного значения;

– текущая оценка решения и процент выполнения превышают указанные пороговые значения.

Данная стратегия вместе с предложенными условиями перехода позволили процессу обучения вовремя справляться с каждым уровнем сложности и постепенно приспосабливаться к новым особенностям среды.

Также для регуляризации и предотвращения преждевременной сходимости был установлен график плавного спада энтропии. График энтропии оказался очень значимым фактором в обучении по расписанию. Настроив ее плавный спад, мы смогли обучить модель на более сложных уровнях среды.

6. РЕЗУЛЬТАТЫ

Для оценки результатов работы предложенных методов был проведен анализ графиков обучения по ряду характеристик, включая средний счет в эпизоде (episode score mean) и процент успешных агентов (percentage done).

На рис. 6 проиллюстрирован эффект применения техники структурирования вознаграждений, а именно обучение без и с дополнительной временной компонентой (WITHOUTtimeRew и timeRewWrap0.7). Как видно из графика, добавление к структуре функции награды разработанной добавочной функции F_{TR} дает существенный вклад в успешность всего процесса обучения. В дополнение к этому, применение “action masking” и формирующих функций F_{DR} и F_{AR} дало небольшое улучшение в счете и скорости сходимости (skipNC-avAct-adj-dist01).

На рис. 7 представлен episode score mean для обучения модели по разработанному расписанию.

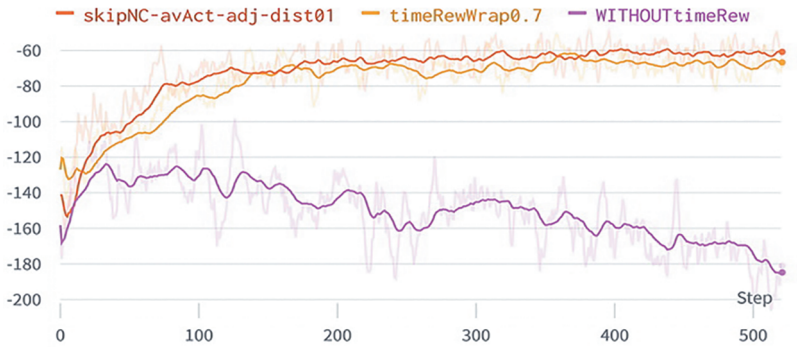

Рис. 6. График метрики episode score mean для обучения без и с применением техники структурирования вознаграждений.

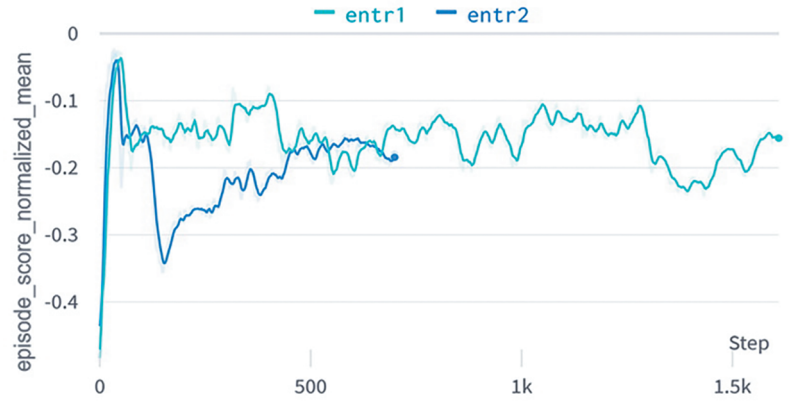

Рис. 7. Обучение по расписанию.

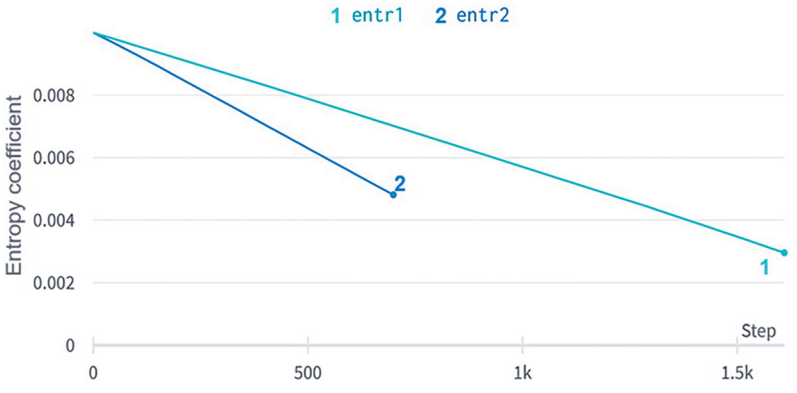

Рис. 8. График спада энтропии.

На рис. 8 показано обучение по двум различным режимам, на темно-синем энтропия спадает более резко, а на голубом – медленно. Можно заметить, что темно синий, хоть и быстро достиг вначале приемлемого значения показателей, но при дальнейшем обучении на новой среде не смог перейти на новый уровень сложности, т.к. начал преждевременно сходиться к локальному максимуму.

Настройка параметров данной модели проиллюстрирована на рис. 9. Больше информации об этом можно найти в [13].

Расширенный механизм наблюдений агента показал немного лучшие результаты, чем базовое наблюдение “Tree observation”, иллюстрируя особую чувствительность агентов к компонентам наблюдений.

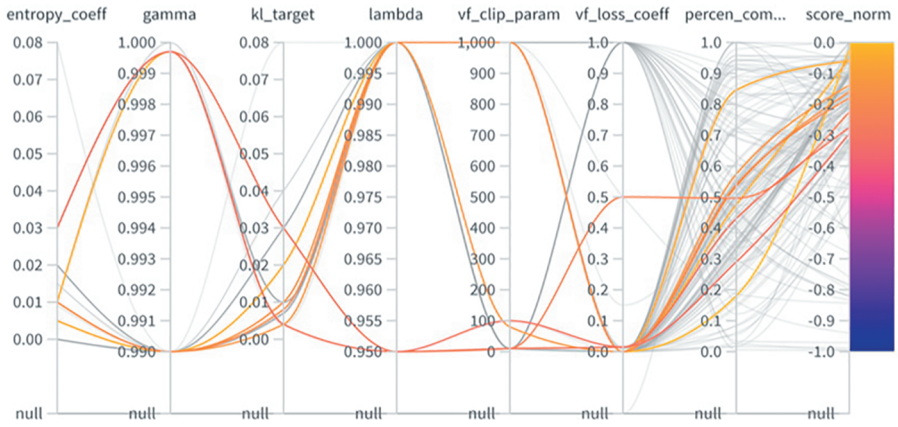

Рис. 9. Настройка гиперпараметров.

7. ВЫВОДЫ

В данной работе предлагается стратегия решения задачи мультиагентного планирования железнодорожных перевозок. Описанное решение, основанное на обучении с подкреплением, заняло первое место в треке RL соревнования Flatland 3. В статье раскрывается задача, поставленная перед участниками соревнования, описываются ее основные сложности и предлагаются методы по их решению. Для решения поставленной задачи был разработан новый подход к проблеме планирования поездов на основе специальной техники структурирования вознаграждения, которая стимулировала агента следовать заданному графику движения. Данная методика вместе с дальнейшей настройкой параметров внесли наибольший вклад в производительность модели.

Кроме того, была разработана стратегия обучения по расписанию, которая позволила процессу обучения вовремя справляться с каждым уровнем сложности и дала возможность обучить модель в более сложных условиях с повышенным уровнем поломок. Полученные результаты показали эффективность разработанных методов. Тем не менее остается большая область для дальнейшего совершенствования и проведения научных исследований в области повышения быстродействия и эффективности алгоритма.

СПИСОК ЛИТЕРАТУРЫ

1. Flatland Intro, https://flatland.aicrowd.com/intro.html. Last accessed 6 June 2022
2. Flatland-3 Homepage. https://www.aicrowd.com/challenges/flatland-3. Last accessed 6 June 2022
3. *Paschchenko F.F., Kuznetsov N.A., Zakharova E.M., Minashina I.K., Takmazian A.K.* Intelligent Control Systems for the Rolling Equipment Maintenance of Rail Transport. 2017 IEEE 11th International Conference on Application of Information and Communication Technologies, IEEE 11th International Conference on Application of Information and Communication Technologies (AICT), IEEE, pp. 1–3, 2017.
4. Flatland-3 Winners, https://www.aicrowd.com/challenges/flatland-3/winners. Last accessed 6 June 2022
5. *Iqbal S., Sha F.* Actor-attention-critic for multi-agent reinforcement learning. International Conference on Machine Learning, pp. 2961–2970, PMLR, 2019.
6. *Ng A.Y., Harada D., Russell S.* Policy invariance under reward transformations: Theory and application to reward shaping. Proceedings of the Sixteenth International Conference on Machine Learning, Icml, vol. 99, pp. 278–287. 1999.
7. *Hu Y., Wang W., Jia H., et al.* Learning to utilize shaping rewards: A new approach of reward shaping, 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada, 2020.
8. *Mohanty S. et al.* Flatland-rl: Multi-agent reinforcement learning on trains. arXiv:2012.05893. 2020. https://doi.org/10.48550/arXiv.2012.05893
9. *Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O.* Proximal Policy Optimization Algorithms. arXiv: 1707.06347 [cs.LG]. 2017. https://doi.org/10.48550/arXiv.1707.06347
10. *Lowe R., Wu Y.I., Tamar A., Harb J., Pieter Abbeel O., Mordatch I.* Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in neural information processing systems. Advances in Neural Information Processing Systems (NIPS 2017), vol. 30. 2017.
11. *Foerster J., Farquhar G., Afouras T., Nardelli N., Whiteson S.* Counterfactual multi-agent policy gradients. AAAI Conference on Artificial Intelligence, vol. 28, n. 1, 2018.
12. Emilio Parisotto et al. Stabilizing transformers for reinforcement learning. International Conference on Machine Learning, PMLR, pp. 7487–7498, 2020.
13. Weights & Biases, https://wandb.ai/innasviri/flatland-sub/reports/Shared-panel22-02-03-12-02-19–VmlldzoxNTE1OTgx. Last accessed 7 June 2022.

EDN: WZJFFH
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 88–93

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

МЕТОДЫ ПЛАНИРОВАНИЯ И ОБУЧЕНИЯ В ЗАДАЧАХ МНОГОАГЕНТНОЙ НАВИГАЦИИ

© 2022 г. К. С. Яковлев[1,2,*], А. А. Андрейчук[1], А. А. Скрынник[1], А. И. Панов[1,2]

Представлено академиком РАН В.Б. Бетелиным
Поступило 28.10.2022 г.
После доработки 31.10.2022 г.
Принято к публикации 03.11.2022 г.

Задача многоагентной навигации возникает, с одной стороны, во множестве прикладных областей. Классический пример – автоматизированные склады, на которых одновременно функционирует большое число мобильных роботов-сортировщиков товаров. С другой стороны, эта задача характеризуется отсутствием универсальных методов решения, удовлетворяющих одновременно многим (зачастую – противоречивым) требованиям. Примером таких критериев могут служить гарантия отыскания оптимальных решений, высокое быстродействие, возможность работы в частично-наблюдаемым средах и т.д. В настоящей работе приведен обзор современных методов решения задачи многоагентной навигации. Особое внимание уделяется различным постановкам задачи. Рассматриваются различия и вопросы применимости обучаемых и необучаемых методов решения. Отдельно приводится анализ экспериментальных программных сред, необходимых для реализации обучаемых подходов.

Ключевые слова: планирование пути, эвристический поиск, обучение с подкреплением, многоагентные системы

DOI: 10.31857/S2686954322070220

1. ВВЕДЕНИЕ

Задача многоагентной навигации в общем виде формулируется следующим образом. Группа мобильных агентов (например, мобильных роботов или персонажей в виртуальной среде) функционирует в общем пространстве, при этом каждому агенту необходимо переместиться в известное ему целевое положение, избегая столкновений как с другими агентами, так и со статическими и стохастическими препятствиями. В последнее время интерес к методам решения этой задачи существенно возрос, в основном в связи с их востребованностью в области складской и сервисной робототехники [1] и в интеллектуальных транспортных системах [2].

Различные допущения, принимаемые на этапе формализации задачи, оказывают существенное влияние на выбор методов решения. Так, одной из наиболее распространенных и активно изучаемых формализаций является так называемая классическая задача многоагентного планирования (Classical MAPF). В этой задаче подразумевается существование централизованного контроллера, который обладает полной информацией о состоянии среды и всех агентов (полная наблюдаемость). При этом время считается дискретным, т.е. за один такт каждый агент может совершить действие перемещения либо ожидания. Пространство также дискретизируется в виде графа, т.е. считается, что агенты могут перемещаться лишь вдоль ребер априори заданного графа, а совершать действия ожидания только в его узлах. На практике обычно используются четырехсвязные графы – решетки (grids) [3]. Известно множество вариаций такой графовой, централизованной постановки задачи. Например, в [4] рассматривается вариант, когда цели агентов не фиксированы, т.е. распределение агентов по целевым положения является частью решения задачи. В работе [5] предполагается, что целей у каждого агента может быть несколько и агенты должны последовательно их посетить. В [6] рассматривается непрерывная задача, когда после достижения одной цели агентов ему сразу же назначается другая (заранее не известная). В целом, несмотря на различия в деталях постановок, централизованные варианты задачи многоагентной навигации обычно решаются классически-

[1] *Институт искусственного интеллекта AIRI, Москва, Россия*
[2] *Федеральный исследовательский центр "Информатика и управление" Российской академии наук, Москва, Россия*
*****E-mail: Yakovlev@airi.net*

ми, необучаемыми алгоритмами, основанными либо на эвристическом поиске в пространстве состояний (в том или ином виде) [8–10], либо на сведении исходной задачи к классическим задачам компьютерных наук, например к задаче о выполнимости булевых формул (SAT) [11], либо задаче о потоках [12].

Помимо задач многоагентной навигации, в которых предполагается полная наблюдаемость и централизованное управление, интерес, в том числе с практической точки зрения, вызывает альтернативная постановка, когда централизованный планировщик отсутствует, а агенты могут наблюдать среду (в том числе других агентов) лишь в определенном радиусе вокруг себя, т.н. частичная наблюдаемость. Такая задача логично формализуется в виде задачи последовательного принятия решений, когда в каждый такт времени каждый агент выбирает для исполнения одно действие, опираясь на текущее наблюдение (и, возможно, на историю наблюдений и взаимодействий со средой). Так же логичным является тот факт, что для решения задачи в такой постановке активно используются методы обучения с подкреплением [13].

Рассмотрим далее методы решения обоих классов задач многоагентной навигации более подробно.

2. НЕОБУЧАЕМЫЕ (КЛАССИЧЕСКИЕ) МЕТОДЫ

Необучаемые методы решения задачи многоагентной навигации обычно применяются, когда рассматривается вариант с полной наблюдаемостью, централизованным контроллером и графовой дискретизацией рабочей области агентов. Задача состоит в том, чтобы построить совокупность неконфликтных траекторий, а именно путей на графе, включающих возможные действия-ожидания в вершинах. Известно, что, с одной стороны, решить подобную задачу, в случае неориентированного графа, можно за полиномиальное время [14], с другой, получение оптимальных решений относится к классу NP-Hard [15]. Если же граф направленный, то и получение неоптимального решения – это NP-трудная задача [16].

Известны способы решения этой задачи, основывающиеся на ее сведении к другим известным задачам компьютерных наук. Так, в [11] задача многоагентного планирования (ЗМАП) сводится к SAT-задаче, в [17] к задаче целочисленного программирования, в [12] к задаче о потоках. Среди подобных методов, более других распространены те, что сводят ЗМАП к SAT. Вероятная причина состоит в том, что для SAT-задачи известно множество эффективных решателей, в результате чего скорость решения исходной задачи также достаточно высока. Также стоит упомянуть о следующей аналогии. ЗМАП при определенных допущениях можно рассматривать как задачу об игре в пятнашки (15 puzzle game). Такой подход используется современными алгоритмами, например, Push and Rotate [18], нацеленными на быстрое получение неоптимальных решений.

Другим способом решения ЗМАП является применение алгоритмов, осуществляющих непосредственно поиск на графе. Очевидно, что для повышения эффективности используются эвристические версии поиска. Классическим алгоритмом эвристического поиска является алгоритм A* [7], с определенными модификациями он может применяться для нахождения оптимальных решений ЗМАП [8], однако в целом этот подход не слишком эффективен, т.к. он по сути рассматривает всех агентов как единого мета агента и осуществляет поиск в комбинированном пространстве с коэффициентом ветвления, который экспоненциально зависит от числа агентов. Во избежание комбинаторного взрыва, применяются различные техники разъединения (decoupled search). К примерам подобных алгоритмов можно отнести CBS [9] и M* [10]. Оба алгоритма гарантируют оптимальность разыскиваемых решений и имеют множество модификаций, среди которых можно выделить модификации, направленные на повышение вычислительной эффективности при сохранении оптимальности решения [19], [20], модификации, позволяющие пренебречь (trade-off) оптимальностью в пользу вычислительной эффективности [21], а также модификации, позволяющие решать ЗМАП в менее строгих ограничениях, например, в непрерывном времени [22].

Еще одним подходом, к решению ЗМАП, базирующемся на эвристическом поиске, является так называемое приоритизированное планирование [23]. Здесь каждому агенту назначается приоритет, а затем ищутся лишь индивидуальные пути, при этом все ранее спланированные траектории считаются неизменяемыми (другими словами – динамическими препятствиями для очередного агента). С теоретической точки зрения такой подход не только не гарантирует оптимальности, но даже не дает гарантий, что решение задачи будет найдено, если оно существует. Тем не менее для определенного класса задач такую гарантию можно дать [24]. Более того, на практике приоритизированные алгоритмы находят решения, очень близкие к оптимальным в очень большом числе случаев, расходуя при этом кратно меньше вычислительных ресурсов. Именно поэтому алгоритмы этого класса очень часто применяются в робототехнике [25].

3. ОБУЧАЕМЫЕ МЕТОДЫ

Известно несколько вариантов применения методов машинного обучения в контексте ЗМАП с полной наблюдаемостью и централизованным контроллером. Во-первых, эти методы могут использоваться для выбора алгоритма многоагентного планирования, наиболее подходящего под конкретную задачу (карту, расположение агентов) [26, 27]. Во-вторых, методы машинного обучения могут использоваться для выучивания различных эвристических правил выбора, присутствующих в классических алгоритмах решения ЗМАП [28, 29]. Также в последние несколько лет широкое распространение получили методы обучения с подкреплением, которые позволяют решать задачу многоагентного планирования в децентрализованной и в частично-наблюдаемых постановках. В одной из первых работ на эту тему [30] была представлена обучаемая стратегия, называемая PRIMAL, которая позже была улучшена и обобщена на случай непрерывного поиска [31], когда после достижения цели агент не завершает эпизод, а получает новую задачу. Оба алгоритма использовали демонстрационные траектории, сгенерированные поисковым алгоритмом ODrM* [32]. Алгоритмы семейства PRIMAL используют сложную функцию вознаграждения и значительное число частных допущений, касающихся конкретных условий и карт (domain knowledge). Например, дополнительный штраф за конфликты или предположение о том, что в локальное наблюдение входят не только позиции других агентов, но и их цели. Похожие допущения были использованы в работе [33], которая предлагает еще один обучаемый алгоритм для решения ЗМАП, но уже для более сложных динамических моделей агентов (например, таких как квадрокоптеры). Обучаемые методы, которые используют полную информацию о статических элементах среды (глобальная информация о положениях других агентов им не доступна) были предложены в работах [34, 35].

Помимо алгоритмов, которые были разработаны специально для задач многоагентного планирования существует ряд универсальных подходов многоагентного обучения с подкреплением, которые могут быть использованы при решении ЗМАП. Из большого набора классических алгоритмов одноагентного обучения (такое обучение еще называют независимым), хорошо себя зарекомендовал для частично-наблюдаемых и многоагентных задач подход градиента стратегии, например, одна из популярных реализаций – алгоритм оптимизации ближайшей стратегии (PPO) [36–38]. Вторым направлением является использование централизованного обучения при обучении кооперативных стратегий. Алгоритмы из этого направления обычно обучаются централизованно, используя глобальную информацию о среде, а тестируются децентрализованно. Так, QMIX [39] использует гиперсети для обучения отдельных стратегий через смешивающую сеть полезности. Каждая отдельная сеть при обучении получает только локальное наблюдение, и оптимизируется гиперсетью, которая использует доступ к глобальному состоянию. Алгоритмы обучения MADDPG [40] (обучение по отложенному опыту) и MAPPO [41] (обучение по актуальному опыту) используют централизованную сеть критика. Это общий подход при обучении, когда критик использует глобальное состояние среды для более качественного обучения аппроксиматора функции полезности. Актор, который определяет стратегию выбора действий, получает на вход только частичное наблюдение, но неявно использует общее наблюдение, пользуясь оценками критика. Алгоритм FACMAC [42] является комбинацией алгоритмов MADDPG и QMIX, что позволяет применять его как для дискретного пространства действий, используя преобразования Гюмбеля, так и для непрерывного.

Алгоритмы направления MARL достаточно сильно оптимизированы под ряд сред, которые уже стали классикой для их тестирования, например SMAC [43], использующий игру Starcraft 2. В отличие от одноагентного обучения с подкреплением в открытом доступе намного меньше готовых реализаций, а те, что существуют, подходят лишь для решения довольно простых задач. Основными причинами этого являются медленные реализации, не использующие распараллеливание и рассчитанные лишь на несколько миллионов шагов в среде, использование простых полносвязных архитектур в качестве аппроксиматоров. Из-за этого большинство исследователей отдают предпочтение использованию известных децентрализованных подходов. Но это приводит к другой крайности – предложенные алгоритмы эксплуатируют знания предметной области, что ограничивает их применимость для широкого класса задач навигации.

Одним из перспективных направлений в создании более продвинутых методов решения ЗАМП могут служить методы обучения с подкреплением на основе модели [44]. Предсказание поведения других агентов, учет этой модели динамики в построении собственной стратегии агента могут оказаться особенно полезными в гетерогенной группе агентов, где у каждого из них может быть разная собственная стратегия [45]. Еще одной незакрытой нишей в MARL является использование демонстраций при обучении. Действительно, использование демонстраций может существенно ускорить обучение, а также позволить использовать современный трансформенные и диффузионные модели. Это особенно

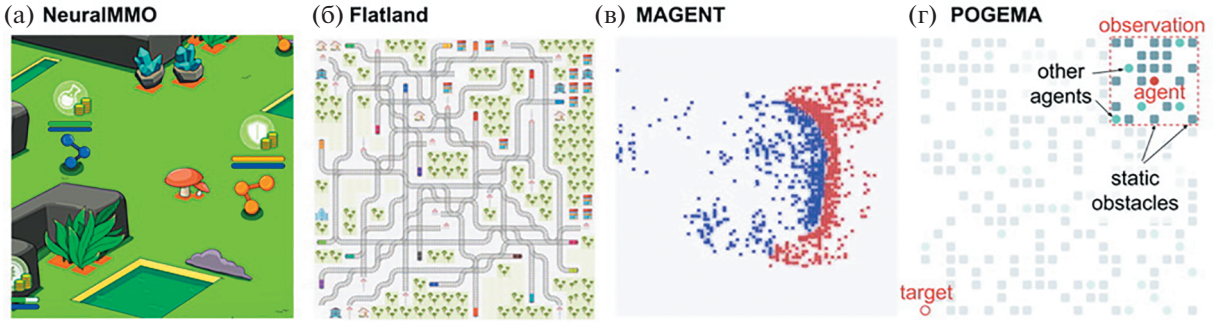

Рис. 1. Примеры сред, используемых для работы с обучаемыми методами решения мультиагентных задач.

актуально для задач навигации, для которых существуют сильные планировочные алгоритмы.

4. ЭКСПЕРИМЕНТАЛЬНЫЕ СРЕДЫ ДЛЯ ТЕСТИРОВАНИЯ АЛГОРИТМОВ

Экспериментальные онлайн среды почти не используются исследователями необучаемых подходов решения ЗМАП, т.к. этим подходы не предполагают обучение путем взаимодействия со средой. Обратная картина наблюдается в сообществе обучения с подкреплением, где существует большое количество различных сред, однако большинство из них являются игровыми и предполагают большое количество дополнительных особенностей, не связанных с задачами многоагентной навигации (например, инвентарь, наличие противодействующих оппонентов и т.п.) (см. рис. 1).

Примером игровой среды может служить NeuralMMO [46], которая является упрощенной версией многопользовательской сетевой игры, в которой группа агентов решает задачу выживания и накопление ресурсов. Команда из 8 агентов соревнуется с другими 15 командами на процедурно генерируемой карте 128 на 128 клеток. Среда является частично наблюдаемой, но агенты могут коммуницировать между собой. Несмотря на то что указанная задача является достаточно сложной, она далека от практического применения, и требует преимущественно реактивного выбора действий на основе системы правил, нежели планирования и навигации.

Одной из наиболее известных сред именно для задачи MAPF является среда Flatland [48] – упрощенная, однако, реалистичная среда для решения задач составления расписания сети железных дорог. Здесь агенты – поезда – должны двигаться от одной станции к другой по путям с односторонним движением, избегая конфликтов друг с другом. В рамках данной задачи было проведено несколько соревнований, целями которых было исследование алгоритмов обучения с подкреплением. Однако оказалось, что доступ к полному состоянию среды дает существенное преимущество подходам планирования и перепланирования [49]. Еще одним недостатком данной среды является то, что она работает очень медленно (около 200 шагов в секунду для небольших карт) при использовании режима наблюдений, предназначенного для обучаемых алгоритмов.

Среда MAGENT – одна из наборов сред библиотеки PettingZoo [47]. Среда предназначена для моделирования роевого поведения агентов, которые могут не только перемещаться из одного места в другое, но и взаимодействовать друг с другом различными способами. Реализация на C++ существенно ускоряет процесс взаимодействия, однако данная среда имеет ограниченный набор сценариев (типов карт) и не имеет интерфейса для тестирования решений, не основанных на обучении с подкреплением.

Наиболее подходящая для задач MAPF среда POGEMA [50] специально создана для задач в частично-наблюдаемой постановке для клеточных карт. Авторы делают особый упор на то, что агенты получают информацию только из ограниченного пространства вокруг себя и не могут передавать какую-либо информацию друг другу, что существенно усложняет задачу как для планировочных алгоритмов, так и для обучаемых. Главными достоинствами данной среды являются ее гибкость и скорость работы. POGEMA позволяет использовать любые созданные пользователем карты препятствий, поддерживает три режима, которые определяются тем, что происходит после достижения агентом цели: агент получает новую цель (непрерывный поиск пути), исчезающие (после достижения цели) агенты и неисчезающие до конца эпизода агенты.

5. ЗАКЛЮЧЕНИЕ

Методы решения задачи многоагентной навигации активно развиваются в последнее время в связи с их востребованностью в различных прак-

тических областях (складская робототехника, транспортные системы и пр.). В условиях централизованного управления и полной наблюдаемости обычно применяются необучаемые методы, основанные либо на эвристическом поиске, либо на сведении задачи многоагентного планирования к другим классическим задачам в области компьютерных наук (SAT-задача, задача о потоках и пр.). В случае, когда централизованный контроллер отсутствует и/или агентам недоступна полная информация об окружающей среде, то зачастую применяются методы обучения с подкреплением, основанные либо на адаптации известных методов поиска стратегии для индивидуальных агентов, либо на парадигме "централизованное обучение – децентрализованное исполнение". Наиболее перспективным (и наименее исследованным) по мнению авторов может являться комбинированный подход, использующий как методы обучения с подкреплением, так и классические методы планирования (эвристический поиск и др.).

СПИСОК ЛИТЕРАТУРЫ

1. *Ma H., Koenig S.* AI buzzwords explained: Multi-agent path finding (MAPF). *AI Matters*. 2017. V. 3 (3). P. 15–19.
2. *Morris R., Pasareanu C.S., Luckow K., Malik W., Ma H., Kumar T.S., Koenig S.* Planning, scheduling and monitoring for airport surface operations. Workshops at the 30th AAAI Conference on Artificial Intelligence, 2016.
3. *Yap P.* Grid-based path-finding. Conference of the canadian society for computational studies of intelligence, Springer, Berlin, Heidelberg, 2002. P. 44–55.
4. *Ma H., Koenig S.* Optimal Target Assignment and Path Finding for Teams of Agents. Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, 2016. P. 1144–1152.
5. *Liu M., Ma H., Li J., Koenig S.* Task and Path Planning for Multi-Agent Pickup and Delivery. Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, 2019. P. 1152–1160.
6. *Li J., Tinka A., Kiesel S., Durham J.W., Kumar T.S., Koenig S.* Lifelong multi-agent path finding in large-scale warehouses. Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2021. P. 11272–11281.
7. *Hart P.E., Nilsson N.J., Raphael B.* A formal basis for the heuristic determination of minimum cost paths. IEEE transactions on Systems Science and Cybernetics, 1968. V. 4(2). P. 100–107.
8. Finding optimal solutions to cooperative pathfinding problems. Proceedings of the 24th AAAI Conference on Artificial Intelligence, 2010. P. 173–178.
9. *Sharon G., Stern R., Felner A., Sturtevant N.R.* Conflict-based search for optimal multi-agent pathfinding. Artificial Intelligence, 2015, 219. P. 40–66.
10. *Wagner G., Choset H.* M*: A complete multirobot path planning algorithm with performance bounds. Proceedings of the 2011 IEEE/RSJ international conference on intelligent robots and systems, 2011. P. 3260–3267.
11. *Surynek P., Felner A., Stern R., Boyarski E.* Efficient SAT approach to multi-agent path finding under the sum of costs objective. Proceedings of the 22nd European conference on artificial intelligence, 2016. P. 810–818.
12. *Yu J., LaValle S.M.* Multi-agent path planning and network flow. Algorithmic foundations of robotics X, 2013. P. 157–173.
13. *Sutton R.S., Barto A.G.* Reinforcement learning: An introduction. 2nd edn. Bradford Books, 2018.
14. *Kornhauser D., Miller G., Spirakis P.* Coordinating Pebble Motion on Graphs, The Diameter of Permutation Groups, And Applications. The 25th Annual Symposium on Foundations of Computer Science, 1984. P. 241–250.
15. *Ratner D., Warmuth M.* The (n2−1)-puzzle and related relocation problems. Journal of Symbolic Computation, 1990. V. 10 (2). P. 111–137.
16. *Nebel B.* On the computational complexity of multi-agent pathfinding on directed graphs. Proceedings of the 20th International Conference on Automated Planning and Scheduling, 2020. P. 212–216.
17. *Yu J., LaValle S.M.* Optimal multirobot path planning on graphs: Complete algorithms and effective heuristics. IEEE Transactions on Robotics, 2016. V. 32 (5). P. 1163–1177.
18. *De Wilde B., Ter Mors A.W., Witteveen C.* Push and rotate: a complete multi-agent pathfinding algorithm. Journal of Artificial Intelligence Research, 2014. V. 51. P. 443–492.
19. *Boyarski E., Felner A., Stern R., Sharon G., Tolpin D., Betzalel O., Shimony E.* ICBS: improved conflict-based search algorithm for multi-agent pathfinding. Proceedings of the 24th International Conference on Artificial Intelligence, 2015. P. 740–746.
20. *Li J., Harabor D., Stuckey P. J., Ma H., Koenig S.* Symmetry-breaking constraints for grid-based multi-agent path finding. Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019. P. 6087–6095.
21. *Barer M., Sharon G., Stern R., Felner A.* Suboptimal variants of the conflict-based search algorithm for the multi-agent pathfinding problem. Proceedings of the 7th Annual Symposium on Combinatorial Search, 2014.
22. *Andreychuk A., Yakovlev K., Surynek P., Atzmon D., Stern R.* Multi-agent pathfinding with continuous time. Artificial Intelligence, 2022. V. 305. P. 103662.
23. *Erdmann M., Lozano-Perez T.* On multiple moving objects. Algorithmica, 1987. V. 2 (1). P. 477–521.
24. *Cap M., Vokrinek J., Kleiner A.* Complete decentralized method for on-line multi-robot trajectory planning in well-formed infrastructures. Proceedings of the 25th International conference on automated planning and scheduling, 2015. P. 324–332.
25. *Yakovlev K., Andreychuk A.* Any-Angle Pathfinding for Multiple Agents Based on SIPP Algorithm. Proceedings of the 17th International Conference on Automated Planning and Scheduling, 2017. P. 586–594.
26. *Kaduri O., Boyarski E. Stern R.* Algorithm selection for optimal multi-agent pathfinding. *Proceedings of the In-*

ternational Conference on Automated Planning and Scheduling, 2020, June, 30. P. 161–165.

27. *Ren J., Sathiyanarayanan V., Ewing E., Senbaslar B., Ayanian N.* MAPFAST: A Deep Algorithm Selector for Multi Agent Path Finding using Shortest Path Embeddings. *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, 2021, May. P.* 1055–1063.

28. *Li J., Chen Z., Harabor D., Stuckey P.J. Koenig S.* MAPF-LNS2: Fast Repairing for Multi-Agent Path Finding via Large Neighborhood Search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

29. *Huang T., Koenig S., Dilkina B.* Learning to resolve conflicts for multi-agent path finding with conflict-based search. *Proceedings of the AAAI Conference on Artificial Intelligence,* 2021*, May,* V. 35, № 13. P. 11246–11253.

30. *Sartoretti G., Kerr J., Shi Y., Wagner G., Kumar T.S., Koenig S., Choset H.* Primal: Pathfinding via reinforcement and imitation multi-agent learning. IEEE Robotics and Automation Letters, 2019. V. 4 (3). P. 2378–2385.

31. *Damani M., Luo Z., Wenzel E., Sartoretti G.* PRIMAL $ _2 $: Pathfinding via reinforcement and imitation multi-agent learning-lifelong. IEEE Robotics and Automation Letters, 2021. V. 6 (2). P. 2666–2673.

32. *Ferner C., Wagner G., Choset H.* ODrM* optimal multirobot path planning in low dimensional search spaces. 2013 IEEE International Conference on Robotics and Automation, 2013, May. P. 3854–3859.

33. *Riviere B., Hönig W., Yue Y., Chung S.J.* Glas: Global-to-local safe autonomy synthesis for multi-robot motion planning with end-to-end learning. IEEE Robotics and Automation Letters, 2020. V. 5 (3). P. 4249–4256.

34. *Liu Z., Chen B., Zhou H., Koushik G., Hebert M., Zhao D.* Mapper: Multi-agent path planning with evolutionary reinforcement learning in mixed dynamic environments. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2020, October. P. 11748–11754.

35. *Wang B., Liu Z., Li Q., Prorok A.* Mobile robot path planning in dynamic environments through globally guided reinforcement learning. IEEE Robotics and Automation Letters, 2020. V. 5 (4). P. 6932–6939.

36. *Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O.* Proximal policy optimization algorithms, 2017, arXiv preprint arXiv:1707.06347.

37. *Skrynnik A., Andreychuk A., Yakovlev K., Panov A.* Pathfinding in stochastic environments: learning vs planning. PeerJ Computer Science, 2022. V. 8. P. e1056.

38. *Berner C., Brockman G., Chan B., Cheung V., Dębiak P., Dennison C., Farhi D., Fischer Q., Hashme S., Hesse C, Józefowicz R.* Dota 2 with large scale deep reinforcement learning, 2019, arXiv preprint arXiv:1912.06680.

39. *Rashid T., Samvelyan M., Schroeder C., Farquhar G., Foerster J., Whiteson S.* Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. International conference on machine learning, 2018, July. P. 4295–4304. PMLR.

40. *Lowe R., Wu Y.I., Tamar A., Harb J., Pieter Abbeel O., Mordatch I.* Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in neural information processing systems, 2017, 30.

41. *Yu C., Velu A., Vinitsky E., Wang Y., Bayen A., Wu Y.* The surprising effectiveness of ppo in cooperative, multi-agent games, 2021, arXiv preprint arXiv:2103.01955.

42. *Peng B., Rashid T., Schroeder de Witt C., Kamienny P.A., Torr P., Böhmer W., Whiteson S.* Facmac: Factored multi-agent centralised policy gradients. Advances in Neural Information Processing Systems, 2021. V. 34. P. 12208–12221.

43. *Samvelyan M., Rashid T., De Witt C.S., Farquhar G., Nardelli N., Rudner T.G., Hung C.M., Torr P.H., Foerster J., Whiteson S.* The starcraft multi-agent challenge, 2019, arXiv preprint arXiv:1902.04043.

44. *Moerland T.M., Broekens J., Jonker C.M.* Model-based Reinforcement Learning: A Survey," pp. 421–429, Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.16712.

45. *Skrynnik A., Yakovleva Y., Davydov D., Yakovlev K., Panov A.I.* Hybrid Policy Learning for Multi-Agent Pathfinding, IEEE Access, 2021. V. 9. P. 126034–126047, https://doi.org/10.1109/ACCESS.2021.3111321

46. *Suarez J., Du Y., Isola P., Mordatch I.* Neural MMO: A massively multiagent game environment for training and evaluating intelligent agents, 2019, arXiv preprint arXiv:1903.00784.

47. *Terry J., Black B., Grammel N., Jayakumar M., Hari A., Sulliva, R., Santos L.S., Dieffendahl C., Horsch C., Perez-Vicente R., Williams N.* Pettingzoo: Gym for multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 2021. V. 34. P. 15032–15043.

48. *Laurent F., Schneider M., Scheller C., Watson J., Li J., Chen Z., Zheng Y., Chan S.H., Makhnev K., Svidchenko O., Egorov V.* Flatland competition 2020: MAPF and MARL for efficient train coordination on a grid world. NeurIPS 2020 Competition and Demonstration Track, 2021, August. P. 275–301. PMLR.

49. *Li J., Chen Z., Zheng Y., Chan S.H., Harabor D., Stuckey P.J., Ma H., Koenig S.* Scalable rail planning and replanning: Winning the 2020 flatland challenge. Proceedings of the International Conference on Automated Planning and Scheduling, 2021, May. V. 31. P. 477–485.

50. *Skrynnik A., Andreychuk A., Yakovlev K., Panov A.I.* POGEMA: Partially Observable Grid Environment for Multiple Agents, 2022, arXiv preprint arXiv:2206.10944.

EDN: IKOLKF
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 94–99

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ПРИМЕНЕНИЕ ПРЕДОБУЧЕННЫХ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ В ЗАДАЧАХ ВОПЛОЩЕННОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

© 2022 г. А. К. Ковалёв[1], А. И. Панов[1,*]

Представлено академиком РАН А.А. Шананиным
Поступило 28.10.2022 г.
После доработки 31.10.2022 г.
Принято к публикации 03.11.2022 г.

Особенностью задач воплощенного искусственного интеллекта является формирование запроса к интеллектуальному агенту на естественном языке. Это приводит к необходимости использования методов обработки естественного языка для перевода этого запроса в формат, удобный для составления корректного плана поведения. Существует два основных подхода к решению этой задачи. Первый подход заключается в использовании специализированных моделей, обученных на конкретных примерах перевода инструкций в исполнимый агентом формат. Второй подход использует способность больших языковых моделей, обученных на большом объеме неразмеченных данных, хранить знания общего назначения (common sense). Это позволяет использовать такие модели для построения плана поведения агента по запросу на естественном языке без предварительного дообучения. В данной обзорной статье подробно рассматриваются модели, использующие второй подход в задачах воплощенного искусственного интеллекта.

Ключевые слова: воплощенные искусственный интеллект, большие языковые модели, знания общего назначения, построение плана поведения

DOI: 10.31857/S268695432207013X

1. ВВЕДЕНИЕ

В последние годы возрос интерес к задачам воплощенного искусственного интеллекта (ВИИ/embodied artificial intelligence), которые, в основном, представлены либо задачами оперирования объектами в человеко-ориентированных средах (household tasks), либо перемещением объектов и навигацией в помещениях или на открытой местности. Отличительной чертой задач ВИИ является формулирование инструкций, описывающих выполняемую задачу или достигаемую цель и передаваемых воплощенному интеллектуальному агенту (ВИА), на естественном языке. Такая постановка приводит к необходимости использовать техники обработки естественного языка для перевода инструкций в вид, удобный для использования ВИИ. Существуют два подхода к решению этой задачи.

Первый подход использует специализированные модели, обученные для получения плана действия агента на основе инструкции. Примерами могут служить техника использования шаблонов возможных действий и определения аргументов этих действий [1] или модели генерации последовательности токенов (Seq2seq) [2].

Второй подход основан на том, что современные большие языковые модели (БЯМ) [3, 4], предобученные на больших корпусах неразмеченных текстов, демонстрируют хорошие результаты на задачах, для решения которых они изначально не были спроектированы, после небольшого дообучения (few-shot learning) [5] или совсем без дообучения [6]. Это достигается за счет того, что такие модели хранят знания общего назначения (common sense). Современные работы используют это свойство БЯМ в задачах воплощенного искусственного интеллекта, например [7].

В данной статье мы подробно рассматриваем модели, относящиеся ко второму подходу к задачам ВИИ, которые используют предобученные большие языковые модели для генерации плана поведения ВИА в среде.

2. ПОДХОДЫ НА ОСНОВЕ ПРЕДОБУЧЕННЫХ БЯМ

В работе Zero-Shot Planners [7] предлагается использовать БЯМ для “заземления” высокоуровневой задачи, выраженной на естественном языке, на множество элементарных действий, до-

[1] *Институт искусственного интеллекта AIRI, Москва, Россия*
*****E-mail: panov@airi.net*

ступных интеллектуальному агенту. В результате по описанию задачи БЯМ должна построить план действий, приводящий к выполнению поставленной задачи, при этом подразумевается, что БЯМ не дообучается. В качестве среды, в которой действует интеллектуальный агент, используется симулятор Virtualhome [8]. Построение плана происходит итеративно: сначала на вход модели подается специально сформулированный запрос с описанием задачи на естественном языке, после чего модель генерирует в свободной форме описание действия, которое необходимо выполнить на первом шаге. Для полученного описания действия считается специальное векторное представление [9], для которого ищется действие из множества элементарных действий с наиболее близким векторным представлением. После этого описание полученного действия добавляется к тексту запроса, и процедура повторяется. Запрос БЯМ формулируется в виде подсказки, где в начале запроса идут пример задачи и план действий по ее выполнению, а в конце добавляется описание текущей задачи. Основное внимание авторы уделяют составлению плана, при этом предполагается, что агент уже умеет выполнять элементарные действия.

В G-PlanET [10] также рассматривается использование БЯМ для генерации плана действий, однако, в отличие от Zero-Shot Planners [7], акцент делается на привязке к конкретной среде, а не только к действиям агента. Используется модификация задачи ALFRED [11], в которой сцена представляется как таблица с перечислением всех доступных на сцене объектов с их типом, положением, ориентацией и родительским объектом (на чем лежит/ в чем находится данный объект). На этапе планирования таблица сцены построчно разворачивается и объединяется с описанием задачи. Полученный текст в виде запроса подается на вход БЯМ. Таким образом, в качестве запроса в G-PlanET [10] используется сгенерированное табличное представление сцены, в которой функционирует интеллектуальный агент, и описание задачи на естественном языке. План генерируется итеративно, результат текущего шага конкатенируется с запросом для этого шага и подается на вход модели для генерации следующего шага. Похожий подход используется в подходе EA-APG (environmentally-aware action plan generation) [12], однако в нем описание сцены состоит только из перечисления объектов.

В архитектуре SayCan [13] процесс отдельного обучения агента выполнению элементарных действий является одним из ключевых этапов. Действие состоит из трех частей: стратегии, т.е. последовательности команд по перемещению интеллектуального агента или его подвижных частей; описания на естественном языке и функции соответствия цели (affordance function), возвращающей вероятность успешного выполнения действия в текущем состоянии среды. Процесс генерации плана похож на процесс, реализованный в Zero-Shot Planners [7], с тем различием, что БЯМ не генерирует описание следующего действия, а дает оценку вероятности того, что элементарное действие полезно для выполнения поставленной задачи. Такая оценка выдается для всех возможных элементарных действий и умножается на вероятность успешного выполнения действия, полученную по функции соответствия (полезности). В качестве следующего действия выбирается действие, с максимальным значением оценки. Отличительной особенностью этой работы является то, что эксперименты проводились как в виртуальной среде, так и на робототехнических платформах в реальной среде. Так же стоит отметить, что запрос состоит не из одного примера как в Zero-Shot Planners [7], а содержит несколько примеров. Еще одним отличием в формировании запроса является то, что он формируется не как описание задачи с планом действий, а как диалог между пользователем и интеллектуальным агентом, в котором пользователь задает вопрос, например, “Как бы ты мог принести мне перекус?”, а агент перечисляет действия, необходимые для решения поставленной задачи. Также авторы применили подход к формированию запроса, предложенный в [14]. Данный подход заключается в добавлении к запросу описания процесса решения задачи помимо самих задачи и решения. Использование такой подсказки в SayCan [13] позволяет улучшить работу модели для задач, где используется отрицание или процесс рассуждения. Ограничение такого подхода заключается в том, что, как показано в [14], улучшения демонстрируются только для БЯМ с более чем 100 миллиардами параметров.

В архитектуре ProgPromt [15] основная идея заключается в представление запроса в виде Python-подобного кода. Запрос состоит из описания доступных действий в виде импортирования соответствующих программных модулей, списком с перечислением доступных объектов на сцене, примерами задачи и их выполнения в виде программных модулей. Использование такого вида запроса обосновывается тем, что БЯМ, такие как GPT-3 с 175 млрд параметров [5], обучаются в том числе и на большом количестве данных из открытых репозиториев программного кода. Похожий подход используется в модели CaP [16], которая генерирует стратегию агента. Отличительной чертой является то, что получаемые программы являются полноценным исполняемым программным кодом и позволяют организовать иерархичность выполнения самих программ.

Некоторые работы, например Socratic Model [17] и Inner Monologue [18], предлагают не конкретную модель, а подход к построению и объ-

единению множества моделей. Так, в Socratic Model [17] предлагается использовать предобученные модели, использующие разные модальности (звук, текст, изображение и др.). Предлагается объединять такие модели в системы, которые способны решать задачи, выходящие за рамки задач каждой отдельной модели. Это достигается за счет создания интерфейса обмена данными между моделями. В качестве примера робототехнической задачи предлагается планирование перемещения объектов на столе. Используется симулятор Pybullet [19] для детектирования объектов на сцене. Для составления описания по изображению используется подход ViLD [20], после чего описание сцены в виде запроса передается в БЯМ для генерации плана по аналогии с Zero-Shot Planners [7] и SayCan, далее план выполняется CLIPort-подобной стратегией [21, 22]. Запрос состоит из описания сцены в формате перечисления объектов python-подобным списком, примеров формулировки задач на естественном языке и их выполнения в виде псевдокода.

В архитектуре Inner Monologue [18] предлагается использовать обратную связь в виде текста, полученную либо от среды (описание сцены, успешность выполнения действия), либо от пользователя (уточнение необходимого действия агента). При этом, сохраняя общий подход, в зависимости от задачи для реализации используются разные модели. Основная идея заключается в итеративном добавлении обратной связи от среды в виде текста во входной запрос, используемый для планирования БЯМ. При манипулировании с объектами (в среде с виртуальным или реальным столом) используется запрос с описанием сцены и примерами задач. Для выполнения задачи на реальной кухне запрос форматируется в виде диалога пользователя и воплощенного агента.

В работе LM-Nav [23] используется подход, попадающий под определение Socratic Models [17], когда для задачи навигации по тексту и изображению используются предобученные отдельно модели, соединенные в общую систему. БЯМ GPT-3 [5] используется для генерации последовательности текстовых ориентиров на основе инструкций на естественном языке, визуально-языковая модель сопоставляет текстовые ориентиры с изображениями, получаемыми агентом [24], а модель навигации по изображению [25] строит и выполняет план перемещения агента. В качестве запроса для БЯМ используются три примера извлечения текстовых ориентиров.

3. КЛАССИФИКАЦИЯ БЯМ ДЛЯ ЗАДАЧИ ПЛАНИРОВАНИЯ

Классификация рассмотренных подходов по типам используемых запросов, средам тестирования и применяемых БЯМ представлена в табл. 1. Обобщая данные, представленные в табл. 1, необходимо заметить, что в общем виде запрос для языковой модели может состоять из следующих частей:

1. Описания сцены, которое заключается или в простом перечислении доступных объектов, или дополняется приведением свойств этих объектов;

2. Перечисления доступных для воплощенного агента действий;

3. Примеров задач, поставленных перед воплощенным агентом;

4. Примеров выполнения поставленных задач.

При этом сам запрос может выражаться или простым текстом, или в виде диалога. Также помимо запроса на естественном языке возможно использовать запросы, представляющие собой или псевдокод, или выполняемый программный код, например на языке Python.

Необходимо отметить, что задача подбора правильного запроса является непростой и на результат могут влиять такие, казалось бы, незначительные изменения, как нумерация списка действий плана и добавления символов переноса строки (см. примеры в SayCan [13]).

Все рассмотренные подходы могут быть описаны общей архитектурой использования БЯМ для задачи планирования действий воплощенного агента, представленной на рис. 1.

На первом этапе, на основе инструкции на естественном языке, описывающей задачу для воплощенного агента, формируется запрос для БЯМ. В простейшем случае запрос состоит из примеров задач с планами выполнения в терминах, доступных для воплощенного агента инструкций [7, 13, 23]. Более сложная структура запроса может включать дополнительную информацию о среде, например, перечисление присутствующих объектов и их свойств [10, 12, 15–18] или доступных действий агента [15, 16]. Далее запрос поступает в БЯМ, которая итеративно генерирует план поведения. Стоит заметить, что в основном запрос формулируется на естественном языке [7, 10, 12, 13, 17, 18, 23], но существуют подходы, использующие для этого псевдокод [16, 17] или программный код [15]. На втором этапе агент выполняет полученный план. Такая постановка подразумевает, что план поведения генерируется полностью до начала выполнения в среде и не модифицируется в процессе выполнения. Это может привести к ситуации, когда агент застревает на одном из этапов выполнения плана, что, в свою очередь, может привести к невыполнению исходной задачи. Решением этой проблемы может быть использование обратной связи от среды (пунктирная стрелка на рис. 1) на каждой итерации генерирования следующего действия агента. В качестве обратной связи может использоваться информация о возможности выполнения конкретного действия [13] или отчет о кор-

Таблица 1. Сравнения алгоритмов для задач воплощенного искусственного интеллекта на основе предобученных БЯМ

Алгоритм	Языковая модель	Среда	Робот. реализация	Тип запроса	Навигация	Взаимод. со средой
Zero-Shot Planners [7]	GPT-3 175B [5] Codex 12B [29]	VirtualHome [8]	–	Примеры задач и их выполнения	+	+
G-PlanET [10]	TaPEX [30]	ALFRED [11] (модификация)	–	Описание сцены, описание задачи	–	–
EA-APG [12]	GPT-3 [5]	VirtualHome [8]	–	Описание сцены, примеры задач и их выполнения	+	+
SayCan [13]	PALM 540B [4]	Естественная среда (кухня)	Everyday Robots	Примеры задач их решения, сформулированные в виде диалога	+	+
ProgPromt [15]	GPT-3 175B [5]	VirtualHome [8]	–	Python-подобный код с описанием доступных действий, сцены, примерами задач и их выполнением	+	+
Socratic [17]	GPT-3 175B [5]	Pybullet [19]	–	Описание сцены, примеры задач и их выполнение на псевдокоде	–	+
Inner Monologue [18]	InstructGPT [31]	Pybullet [19]	–	Описание сцены, примеры задач и их выполнения	–	+
	InstructGPT [31]	Естественная среда (стол)	Everyday Robots	Описание сцены, примеры задач и их выполнения	–	+
	PALM 540B [4]	Естественная среда (кухня)	Everyday Robots	Примеры задач и их выполнения в виде диалога	+	+
CaP [16]	Codex [29] code-davinci-002	Естественная среда (рисование на белой доске)	UR5e	Python-подобный код с описанием доступных действий, сцены, примерами задач и их выполнением	+	–
	Codex [29] code-davinci-002	Естественная среда (стол)	UR5e	Python-подобный код с описанием доступных действий, сцены, примерами задач и их выполнением	+	–
	Codex [29] code-davinci-002	Естественная среда (кухня)	Everyday Robots	Python-подобный код с описанием доступных действий, сцены, примерами задач и их выполнением	+	+
LM-Nav [23]	GPT-3 [5]	Естественная среда (улица)	Clearpath Jackal UGV	Примеры задач и их выполнения	+	–

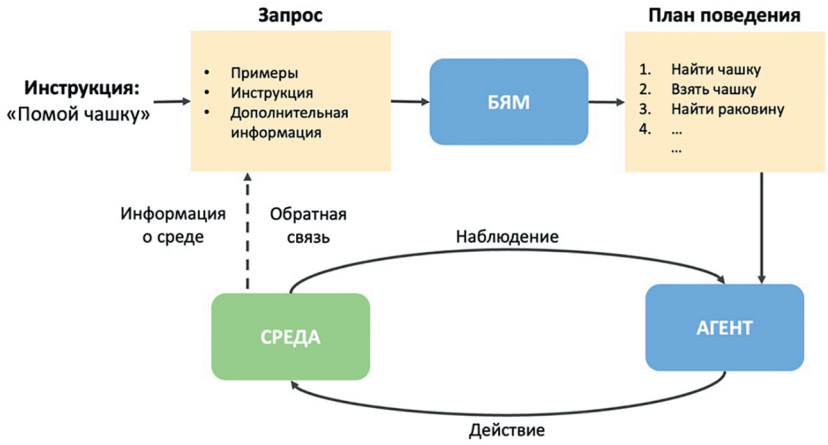

Рис. 1. Общая архитектура использования больших языковых моделей (БЯМ) для задачи построения плана поведения воплощенного агента.

ректном выполнении действия или об изменении состояния объекта [18].

4. ЗАКЛЮЧЕНИЕ

Рассмотренные подходы использования БЯМ для планирования поведения демонстрируют неплохие результаты как в виртуальных средах, так и при имплементации на робототехнических платформах. Тем не менее количественное сравнение этих работ осложнено тем, что в них используются (за исключением нескольких работ) различные среды. При этом существуют и хорошо известны задачи с установленными метриками качества и таблицами сравнений, на которых тестируются воплощенные интеллектуальные агенты, например ALFRED [11] и TEACh [26] для следования инструкциям на естественном языке, RoomR [27] и Benchbot [28] для перестановки объектов, и другие. К сожалению, пока в этих бенчмарках большинство указанных работ не представлены, что во многом объясняется сложностью организации запросов в разнообразных средах с большим количеством объектов и действий.

Перспективным будущим направлением работ являются, во-первых, усложнение и обучение обратной связи для определения качества выдаваемых ответов БЯМ, и, во-вторых, модификация процесса обучения самих БЯМ, чтобы добавить регуляризатор, моделирующий реалистичностость генерируемых ответов, в функцию потерь всей языковой модели.

СПИСОК ЛИТЕРАТУРЫ

1. *Min S.Y. et al.* Film: Following instructions in language with modular methods //arXiv preprint arXiv:2110.07342. 2021.
2. *Liu H. et al.* LEBP – Language Expectation & Binding Policy: A Two-Stream Framework for Embodied Vision-and-Language Interaction Task Learning Agents // arXiv preprint arXiv:2203.04637. 2022.
3. *Devlin J. et al.* Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. 2018.
4. *Chowdhery A. et al.* Palm: Scaling language modeling with pathways //arXiv preprint arXiv:2204.02311. 2022.
5. *Brown T. et al.* Language models are few-shot learners // Advances in neural information processing systems. 2020. Т. 33. С. 1877–1901.
6. *Wei J. et al.* Finetuned language models are zero-shot learners //arXiv preprint arXiv:2109.01652. 2021.
7. *Huang W. et al.* Language models as zero-shot planners: Extracting actionable knowledge for embodied agents // arXiv preprint arXiv:2201.07207. 2022.
8. *Puig X. et al.* Virtualhome: Simulating household activities via programs //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. С. 8494–85.
9. *Reimers N., Gurevych I.* Sentence-bert: Sentence embeddings using siamese bert-networks //arXiv preprint arXiv:1908.10084. 2019.
10. *Lin B.Y. et al.* On Grounded Planning for Embodied Tasks with Language Models // arXiv preprint arXiv:2209.00465. 2022.
11. *Shridhar M. et al.* Alfred: A benchmark for interpreting grounded instructions for everyday tasks //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. С. 10740–10749.
12. *Gramopadhye M., Szafir D.* Generating Executable Action Plans with Environmentally-Aware Language Models //arXiv preprint arXiv:2210.04964. 2022.
13. *Ahn M. et al.* Do as i can, not as i say: Grounding language in robotic affordances //arXiv preprint arXiv:2204.01691. 2022.
14. *Wei J. et al.* Chain of thought prompting elicits reasoning in large language models //arXiv preprint arXiv:2201.11903. 2022.

15. *Singh I. et al.* ProgPrompt: Generating Situated Robot Task Plans using Large Language Models //arXiv preprint arXiv:2209.11302. 2022.
16. *Liang J. et al.* Code as policies: Language model programs for embodied control //arXiv preprint arXiv:2209.07753. 2022.
17. *Zeng A. et al.* Socratic models: Composing zero-shot multimodal reasoning with language //arXiv preprint arXiv:2204.00598. 2022.
18. *Huang W. et al.* Inner monologue: Embodied reasoning through planning with language models //arXiv preprint arXiv:2207.05608. 2022.
19. *Coumans E., Bai Y.* Pybullet, a python module for physics simulation for games, robotics and machine learning. GitHub Repository – 2016.
20. *Gu X. et al.* Open-vocabulary object detection via vision and language knowledge distillation //arXiv preprint arXiv:2104.13921. 2021.
21. *Shridhar M., Manuelli L., Fox D.* Cliport: What and where pathways for robotic manipulation //Conference on Robot Learning. PMLR, 2022. С. 894–906.
22. *Zeng A. et al.* Transporter networks: Rearranging the visual world for robotic manipulation //arXiv preprint arXiv:2010.14406. 2020.
23. *Shah D. et al.* Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action // arXiv preprint arXiv:2207.04429. 2022.
24. *Radford A. et al.* Learning transferable visual models from natural language supervision //International Conference on Machine Learning. PMLR, 2021. С. 8748–8763.
25. *Shah D. et al.* Ving: Learning open-world navigation with visual goals //2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021. С. 13215-13222.
26. *Padmakumar A. et al.* Teach: Task-driven embodied agents that chat //Proceedings of the AAAI Conference on Artificial Intelligence. 2022. Т. 36. №. 2. С. 2017–2025.
27. *Weihs L. et al.* Visual room rearrangement //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021. С. 5922–5931.
28. *Talbot B. et al.* Benchbot: Evaluating robotics research in photorealistic 3d simulation and on real robots // arXiv preprint arXiv:2008.00635. 2020.
29. *Chen M. et al.* Evaluating large language models trained on code //arXiv preprint arXiv:2107.03374. 2021.
30. *Liu Q. et al.* Tapex: Table pre-training via learning a neural sql executor //arXiv preprint arXiv:2107.07653. 2021.
31. *Ouyang L. et al.* Training language models to follow instructions with human feedback //arXiv preprint arXiv:2203.02155. 2022.

EDN: GEJHHI
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 100–101

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ТЕОРЕТИЧЕСКИЕ ПРЕДПОСЫЛКИ ФИЗИЧЕСКИ ОБОСНОВАННОГО МАШИННОГО ОБУЧЕНИЯ И ЕГО ПРИЛОЖЕНИЯ К ГИДРОДИНАМИКЕ

© 2022 г. А. В. Корнаев[1,*], Е. П. Корнаева[2], И. С. Стебаков[3]

Представлено академиком РАН А.Л. Семеновым
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

Некоторые законы физики постулируют, что некоторая величина в рассматриваемом физическом процессе должна принимать экстремальное значение. В работе предложен вариант обобщения одного из таких законов и представлен подход применения искусственных нейронных сетей в качестве инструмента минимизации мощности внутренних сил и моделирования гидродинамических процессов для различных приложений.

Ключевые слова: физически обоснованное машинное обучение, глубокое обучение, сегментация изображений, вариационная задача, целевой функционал

DOI: 10.31857/S2686954322070128

1. ВВЕДЕНИЕ

Исследование течений вязких жидкостей обычно сопряжено с решением краевых задач, включающих дифференциальные уравнения в частных производных, в том числе уравнения Навье–Стокса, уравнения неразрывности и других. Стоит отметить, что поиск аналитического решения уравнения Навье–Стокса входит в перечень задач тысячелетия. В настоящее время задачи гидродинамики решаются численно с применением методов конечных разностей, конечных элементов и контрольных объемов. Их применение сопряжено с необходимостью разработки сложных программных комплексов. Однако существует альтернативный подход к решению задач гидродинамики, основанный на поиске экстремумов целевых функционалов. Такой подход подразумевает аппроксимации неизвестных функций в области течения среды. Минимизация целевого функционала посредством аппроксимации неизвестных функций – типичная задача для машинного обучения.

Основной **целью** исследования являются поиск и реализация физически обоснованного целевого функционала для машинного обучения, позволяющего моделировать течение неньютоновских и магнитореологических жидкостей.

Основные вызовы исследования:

– необходимо доказать, что минимизация предложенного целевого функционала эквивалентна решению краевой задачи гидродинамики;

– необходимо реализовать предложенный физически обоснованный функционал в виде алгоритмов и программ расчета гидродинамических задач;

– необходимо найти применения предложенным разработкам.

2. МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Альтернативный подход моделирования течений вязких сред сложной реологии путем решения вариационных задач требует обоснования эквивалентности классическому подходу. Наиболее распространенный способ доказательства эквивалентности сводится к тому, что для подынтегрального выражения исследуемого функционала записываются дифференциальные уравнения Эйлера–Лагранжа, и если они совпадают с уравнениями классической постановки задачи, то считается, что решение этих уравнений сообщают функционалу стационарное значение (в частности, минимальное или максимальное). Поэтому в

[1] *Исследовательский центр в сфере искусственного интеллекта, Университет Иннополис, Иннополис, Россия*
[2] *Кафедра информационных систем и цифровых технологий, Орловский государственный университет имени И.С. Тургенева, Орел, Россия*
[3] *Кафедры Мехатроники, механики и робототехники, Орловский государственный университет имени И.С. Тургенева, Орел, Россия*
*****E-mail: a.kornaev@innopolis.ru*

процессе проводимого исследования возникла необходимость установить вид обобщенных уравнений Эйлера–Лагранжа для функционалов, зависящих от многих функций многих переменных, их производных первого и второго порядка. Поиск функционала осуществлялся эвристически. В качестве базового варианта использовался функционал Лагранжа. В поиске нового функционала было необходимо отказаться от допущения о ньютоновских свойствах среды и о малости действия массовых сил (например, электромагнитной природы). В результате был предложен вариант обобщения функционала Лагранжа, поиск минимума которого эквивалентен решению уравнения переноса вихря. В качестве неизвестной функции используется векторная функция (пси-функция), ротором которой является поле скоростей.

Для алгоритмической реализации решения вариационной задачи были предложены два типа алгоритмов представления области течения: на основе координат области и на основе изображения области. Второй оказался более универсальным, так как для его реализации необходимо использование изображения области течения с масками для его границ. Далее речь идет о втором алгоритме.

На вход в сеть подается изображение области течения с масками для границ, при этом значения масок определяют величину расхода жидкости в области течения и являются аналогами граничных условий. В области течения инициализируется неизвестная пси-функция, например, в виде линейного распределения. Программная реализация алгоритма выполнена с использованием искусственной нейронной сети архитектуры типа U-Net. Особенностью архитектуры является то, что на выходе сети также определяется изображение. В данном случае это преобразованное в ходе прямого прохода изображение неизвестной пси-функции. Далее путем численного дифференцирования и соотношений гидродинамики определяются функции скорости течения жидкости, тензоры скоростей деформаций, интенсивности сдвиговых скоростей деформаций и касательных напряжений, и, наконец, значение целевого функционала. Затем реализуется обратный проход с определением компонент градиента целевого функционала и корректировкой параметров сети. Процедуры расчета повторяются до достижения минимума целевого функционала.

3. ОСНОВНЫЕ РЕЗУЛЬТАТЫ, ВЫВОДЫ

Предложенные физически обоснованный целевой функционал, алгоритм и программа его реализации позволяют моделировать течения неньютоновских жидкостей в каналах произвольной формы. В качестве вариантов тестовых задач использовались задачи о течении между параллельными пластинами (известно аналитическое решение, в том числе для неньютоновских жидкостей, реомагнитных жидкостей), течение между пластинами с углублениями, а также течение крови в капилляре ногтевого ложа. Последняя задача была решена с применением данных видеокапилляроскопии в виде изображения области течения.

Основными преимуществами разработанного метода моделирования являются: возможность моделирования неньютоновских жидкостей; отсутствие в необходимости датасета для обучения; простота реализации; универсальность; глобальная аппроксимация полей гидродинамических величин.

К недостаткам следует отнести: высокая вычислительная стоимость (длительность вычислений выше, чем при использовании конечно-элементных моделей и коммерческих продуктов для их реализации); зависимость точности от разрешения изображений; сложность моделирования нестационарных процессов.

Основные перспективы развития данного исследования связаны с обобщением алгоритмов и программ для обработки трехмерных изображений областей течения, применение разработанного инструментария в методах ‘in-silico’ диагностики состояния кровеносной системы человека, а также для моделирования процесса доставки лекарств по кровеносным сосудам.

СПИСОК ЛИТЕРАТУРЫ

1. *Patankar S.V.* Numerical Heat Transfer and Fluid Flow. 1st ed. Boca Raton: CRC Press, 1980. 1–214 p.
2. *Gelfand I.M., Fomin S.V.* Calculus of Variations / ed. Silverman R.A. Courier Corporation, 2000. 240 p.
3. *Schechter R.S., Newell G.F.* The Variational Method in Engineering // Journal of Applied Mechanics. American Society of Mechanical Engineers Digital Collection, 1968. Vol. 35, № 1. 200–200 p.
4. *Petrov A.G.* Variational principles and inequalities for the velocity of a steady viscous flow // Fluid Dynamics 2015 50:1. Springer, 2015. Vol. 50, № 1. P. 22–32.
5. *Petrov A.G.* Variational principles and inequalities for the velocity of a steady viscous flow // Fluid Dynamics 2015 50:1. Springer, 2015. Vol. 50, № 1. P. 22–32.
6. *Kornaeva E., Kornaev A., Egorov S.* Application of artificial neural networks to solution of variational problems in hydrodynamics // J Phys Conf Ser. Institute of Physics Publishing, 2020. Vol. 1553, № 1.
7. *Kornaev A.V. et al.* Application of variational approach to non-Newtonian fluid flow modelling // Proceedings of 10th International Scientific Conference BALTTRIB 2019. Vytautas Magnus University, 2019. P. 194–201.
8. *Kornaeva E. et al.* Physics-based loss and machine learning approach in application to non-Newtonian fluids flow modeling // 2022 IEEE Congress on Evolutionary Computation, CEC 2022 – Conference Proceedings. Institute of Electrical and Electronics Engineers Inc., 2022.

EDN: GCMPGE
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 102–103

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

AI-РЕЦЕНЗИРОВАНИЕ ПОЛИГРАФНЫХ СКРИНИНГОВ

© 2022 г. Д. В. Асонов[1], М. А. Крылов[2,*]

Представлено академиком РАН А.Л. Семеновым
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

Представлен короткий обзор и результаты научного исследования возможности AI-рецензирования полиграфных скринингов. Результаты исследования будут применяться на практике для укрепления внутренней безопасности в ПАО Сбербанк в конце 2022 г. Полностью исследования будут представлены в публикации [1].

Ключевые слова: детекция лжи, полиграф, скрининг, рецензирование, AI
DOI: 10.31857/S2686954322070025

Защита денежных средств и данных клиентов традиционно является одним из столпов банковской культуры и репутации. В качестве одного из инструментов защиты клиентов банки используют полиграфные скрининги (ПС). Финансовая отрасль – не единственная, использующая ПС. Такие критичные с т.з. возможных последствий от внутреннего мошенничества отрасли как авиация, промышленность, правоохранительные структуры, государственные органы во всем мире также используют ПС.

Тема детекции лжи (выявление скрываемой информации) все больше интересует общество, что косвенно видно по количеству научно-популярных “лонгридов” [2–9]. На фоне увеличения значимости корпоративной безопасности и культуры нетерпимости к внутреннему мошенничеству, интенсивность научных исследований в области детекции лжи растет по всему миру, последние 2–3 года публикуется около тысячи научных статей в год. Исследования по теме детекции лжи являются ультра-мультидисциплинарными, ведутся учеными в таких областях, как психофизиология, нейронауки, безопасность, юриспруденция, компьютерные науки и AI, и т.д.

Возможно, самый острый вопрос в научных сообществах по теме, который до сих пор не разрешен уже на протяжении нескольких десятилетий: точность полиграфа, в частности, как ее считать и какова природа ошибок в выводах (ложь выявлена/не выявлена). В докладе мы расскажем, как мы частично ответили на этот вопрос, решив конкретную и практичную научную задачу.

В рамках минимизации рисков внутреннего мошенничества на рисковых направлениях деятельности Банка подразделение внутренней безопасности проводит скрининговые проверки на полиграфе кандидатов на трудоустройство и действующих сотрудников, которые проводятся только с их согласия и в полном соответствии с законодательством. Ежегодно проводится около 6 тысяч таких скринингов. Если предположить, что порядка 5% скринингов имеют ошибочные выводы специалистов-полиграфологов, существуют риски неправильной оценки благонадежности порядка 300 сотрудников. Чтобы избежать подобных ошибок, внутренняя безопасность проводит рецензирование (запрос второго мнения) по неоднозначным результатам скринингов. Это увеличивает расходы на полиграфные скрининги. Мы задумались над тем, как можно запрашивать второе мнение не только по неоднозначным результатам, а по всем без исключения исследованиям, и одновременно снизить расходы и уменьшить вероятность противоречивых результатов скринингов. Так началось данное исследование.

Ошибки полиграфа часто связывают с недостатками конкретного полиграфического метода и вероятностной природой выводов. Сотни научных работ изучили недостатки различных методов. Мы же рассмотрели вид ошибки, исследования которого ранее не опубликовались, и связанного с неоднозначностью принятия решения специалистом, вне зависимости чем она была вызвана: применяемой методикой или внешними факторами.

[1] *ПАО Сбербанк, Блок “Технологии”, Управление исследований и инноваций, Москва, Россия*
[2] *ПАО Сбербанк, Управление внутрибанковской безопасности, Москва, Россия*
*****E-mail: makrylov@sberbank.ru*

С помощью AI мы построили, насколько нам известно – первый в мире, прототип второго мнения для выводов полиграфологов, и пропилотировали его на исторических данных (записях полиграфных скринигов и выводах полиграфологов по ним) [1]. Результаты пилота подтвердили практическую пользу от применения модели для Внутренней Безопасности еще до окончания научного исследования и позволят значительно снизить привлекаемые человеческие ресурсы, временные и финансовые ресурсы на проведение ручного рецензирования. На основе прототипа сейчас завершается внедрение MVP, и в конце Q4'22 MVP начнет вносить дополнительный вклад в укрепление внутренней безопасности в Банке.

Исследование принесло также широкий набор побочных, значимых результатов:

i. Мы первые замерили качество моделей не только на всем датасете (который включает множество риск-факторов), а также на каждом риск-факторе отдельно. Это позволило нам выдвинуть гипотезу о том, что люди реагируют по-разному на собственную ложь при ответе на вопросы по разным темам. Косвенно, мы смогли получить количественную оценку качества вопросов, которые содержательно наполняют проверку отдельной темы. Чем сложнее модели натренироваться – тем менее четко сформулированы вопросы.

ii. Результаты пилота показали, что модели не только находят противоречия в выводах полиграфологов, но и последующий их анализ позволяет выявлять классы ранее неизвестных системных и процессных ошибок.

iii. Пилот показал возможность определять сверхсложные случаи, характеризующиеся тем, что испытуемый применял методы противодействия полиграфной проверке.

iv. Как и во многих других областях исследований в области AI, в детекции лжи практически нет выверенных golden standard датасетов. Запуск MVP и ручная перепроверка скрингов, где мнения модели и полиграфолога разошлись, инициирует быстрое накопление самого большого в мире golden standard по теме.

v. Мы придумали и реализуем возможность для ученых по всему миру далее раздвигать границы познания в этой области и проводить собственные эксперименты на обезличенных данных 2100+ проверок, при этом соблюдая законодательство РФ по трансграничной передаче персональных данных. Самый большой датасет реальных полиграфных проверок, который был доступен ученым, ранее имел размер 149 проверок, и ученые (Carnegie Mellon University) не могли им делиться [10].

СПИСОК ЛИТЕРАТУРЫ

1. *Dmitri Asonov, Maksim Krylov, Vladimir Omelyusik, Anastasiya Ryabikina, Evgeny Litvinov, Maksim Mitrofanov, Maksim Mikhailov*. Albert Efimov. Building a Second-Opinion Tool for Classical Polygraph. 2022, Under review at Scientific Reports, Nature.
2. Truth vs Lies. Special issue of Scientific American, 2022.
3. True story? Lie detection systems go high-tech. BBC, 2022.
4. Lie detectors have been unreliable for over a century but more Britons are being subjected to polygraph tests. iNews.co.uk, 2022.
5. Will Your Cheatin' Heart Tell on You? As Americans Lose Trust in Each Other, They're Turning to Tech to Detect Lies. TheInformation.com, 2022.
6. Lie detectors have always been suspect. AI has made the problem worse. MIT Technology Review, 2020.
7. AI lie detector developed for airport security. Financial Times, 2019.
8. The race to create a perfect lie detector – and the dangers of succeeding. The Guardian, 2019.
9. A New AI That Detects "Deception" May Bring an End to Lying as We Know It. Futurism.com, 2018.
10. *Aleksandra Slavkovic*. Evaluating Polygraph Data. Carnegie Mellon University, 2002.

EDN: SOGRNG
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 104–105

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ruSciBERT: ЯЗЫКОВАЯ МОДЕЛЬ НА БАЗЕ АРХИТЕКТУРЫ ТРАНСФОРМЕР ДЛЯ ПОЛУЧЕНИЯ СЕМАНТИЧЕСКИХ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ НАУЧНЫХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

© 2022 г. Н. А. Герасименко[1,*], А. С. Чернявский[1], М. А. Никифорова[1]

Представлено академиком РАН А.Л. Семеновым
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

Значительный рост числа научных публикаций и количества научных отчетов делает задачу их обработки и анализа сложной и трудозатратной. Языковые модели, основанные на архитектуре Трансформер и предобученные на больших текстовых коллекциях, позволяют качественно решать множество задач анализа текстовых данных. Для работы с научными текстами на английском языке существуют модели SciBERT [1] и ее модификация SPECTER [2], однако они не поддерживают русский язык в связи с малым количеством текстов в обучающей выборке. Кроме того, способ оценки качества языковых моделей для научных текстов, бенчмарк SciDocs, также поддерживает только английский язык. Предлагаемая модель ruSciBERT позволит решать широкий спектр задач, связанных с анализом научных текстов на русском языке, а прилагаемый к ней бенчмарк ruSciDocs позволит оценивать качество языковых моделей применительно к этим задачам.

Ключевые слова: языковая модель, семантические представления, SciBERT, SciDocs
DOI: 10.31857/S2686954322070074

SciBERT является языковой моделью, обученной на многодоменном корпусе научных статей, написанных преимущественно на английском языке. Авторы предложили взять базовую модель BERT и дообучить ее на задаче предсказания маскированных токенов. Результаты, полученые авторами на нескольких задачах классификации и NER для научных статей, значительно превосходят результаты базовой модели. Дополнительно обученный токенизатор позволил улучшить получаемое качество language modeling. Мы используем похожие идеи и предлагаем дообучение модели RoBERTa на русскоязычном корпусе научных текстов с собственным токенизатором. Данную модель мы называем RuSciBERT. В качестве базовой модели нами была выбрана RoBERTa в связи с тем, что она обучена на расширенном количестве данных, большем количестве задач и достигла лучших результатов по сравнению с базовым BERT.

SciDocs является бенчмарком для оценки качества семантических векторных представлений, получаемых с помощью языковых моделей. Он включает в себя 4 типа задач:

1. Классификация на основе классификаторов MAG и MeSH

2. Предсказание цитирования на основе Semantic Scholar Academic Graph

2.1. Прямые цитаты (задача ранжирования)

2.2. Социтируемые статьи (задача ранжирования)

3. Предсказание активности пользователей Semantic Scholar

3.1. Сопросматриваемые статьи (задача ранжирования)

3.2. Сопрочитываемые статьи (задача ранжирования)

4. Рекомендации статей, похожих на статью-запрос (задача ранжирования).

ruSciBERT планируется обучить на датасете, включающем около 1 млрд токенов. Данные для обучения собраны из открытых источников, позволяющих использовать данные для некоммерческих целей (например, из датасета Semantic Scholar Academic Graph). Размер словаря токенизатора в нашем случае равен 50265 по аналогии с базовой моделью RoBERTa.

Бенчмарк ruSciDocs планируется составить из задач, аналогичных части задач оригинального SciDocs:

1. Классификация на основе классификаторов

[1] *ПАО "Сбербанк", Москва, Россия*
*****E-mail: nikgerasimenko@gmail.com*

1.1. MAG – верхний уровень Microsoft Academic Graph, таксономии областей знания, составленной специалистами из Microsoft и Allen institute for AI

1.2. OECD из ЕГИСУ НИОКТР, государственного сайта для учета научно-исследовательских работ

2. Предсказание цитирования на основе данных из Semantic Scholar Academic Graph.

На данный момент мы обучили модель RuSciBERT на 300 млн токенов на двух эпохах. Она показывала хорошие результаты при заполнении пробелов в текстовых фразах, а также гораздо более низкий уровень перплексии на отложенной выборке по сравнению с общей языковой моделью ruBERT, обученной на текстах всех тематик. Так, RuSciBERT имеет перплексию 4.81, причем она монотонна снижается на последних шагах обучения, вследствие чего модель можно дообучать дальше. В то же время ruBERT имеет перплексию 9.64.

Примеры работы нашей модели заполнения маскированных токенов показаны ниже. В них маскированные токены обозначены через “<mask>”, а модель предсказывает 3 наиболее вероятных варианта токенов для замены.

1) “при использовании в усилителе мощности адаптивной измерительной <mask> появится возможность” -> 'системы', 'аппаратуры', 'станции'

2) “указанные оппоненты не имеют <mask> проектов и публикаций с соискателем” -> 'совместных', 'собственных', 'аналогичных'

3) “новый метод управления <mask> характеристиками ао фильтров” -> 'техническими', 'технологическими', 'функциональными'

RuBERT также показывает неплохие результаты, но некоторые из его вариантов заполнения являются менее удачными. Так, в первом примере среди предсказанного множество токенов есть 'технологии' (меньше подходит по смыслу чем остальные варианты), во втором – 'своих', а в третьем – 'всеми' (возможные варианты, но более общие, и поэтому менее качественные).

Основываясь на текущих промежуточных результатах, можно предположить, что ruSciBERT, обученный на датасете в 1 млрд токенов, покажет наилучшие разультаты на бенчмарке ruSciDocs по сравнению с другими существующими подходами.

СПИСОК ЛИТЕРАТУРЫ

1. Iz Beltagy and Kyle Lo and Arman Cohan. SciBERT: Pretrained Language Model for Scientific Text // EMNLP, 2019.
2. Arman Cohan and Sergey Feldman and Iz Beltagy and Doug Downey and Daniel S. Weld. SPECTER: Document-level Representation Learning using Citation-informed Transformers // ACL, 2020.

EDN: UWFXHB
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 106–108

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ИНКРЕМЕНТАЛЬНОЕ ОБУЧЕНИЕ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ПОИСКА ТРЕНДОВЫХ ТЕМ В НАУЧНЫХ ПУБЛИКАЦИЯХ

© 2022 г. Н. А. Герасименко[1,*], А. С. Чернявский[1], М. А. Никифорова[1], М. Д. Никитин[1], К. В. Воронцов[2]

Представлено академиком РАН В.Б. Бетелиным
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

Стремительный рост числа научных публикаций, интенсивное появление новых направлений и подходов ставят перед научным сообществом задачу своевременного выявления трендов. Под трендом мы понимаем семантически однородную тему, которая характеризуется устойчивым во времени лексическим ядром и резким, зачастую экспоненциальным, ростом числа публикаций [1]. Примерами трендов в машинном обучении являются “LSTM”, “deep learning”, “word2vec”, “BERT”, “fake news detection”. Для выделения трендовых тем в потоке научных публикаций в реальном времени мы используем инкрементальные методы вероятностного тематического моделирования. При помощи подхода, основанного на ARTM, мы превзошли результаты популярных классических и нейросетевых подходов к задаче ранней детекции трендов. Для оценки качества мы вручную сформировали и сделали общедоступным датасет из 91 тренда.

Ключевые слова: инкрементальное тематическое моделирование, детектирование научных трендов, ARTM

DOI: 10.31857/S2686954322070086

Мы рассматриваем задачу ранней детекции трендовых тем. Эксперименты по выделению трендов производились на коллекции из 73 959 статей, опубликованных с 2000 по 2021 г. на конференциях по машинному обучению с h-индексом, превышающим 100. Валидационный датасет охватывает тренды в области машинного обучения и искусственного интеллекта 2009–2021 гг., каждый из которых характеризуется набором из не менее, чем 10 ключевых статей и 5 ключевых терминов. Обучение моделей производилось без учителя, а валидационная разметка использовалась только для финальной оценки качества.

Чтобы отслеживать появление новых тем, мы обучали отдельные модели для каждого временного шага. При поступлении новой порции документов D' словарь пополняется новыми терминами W' и могут образоваться новые темы T'. Предполагается, что новая лексика, появившаяся в новых документах, относится преимущественно к новым темам (рис. 1). Дополнительные ограничения на тематическую модель накладываются в рамках подхода аддитивной регуляризации ARTM с использованием библиотеки BigARTM [2]. В частности, для повышения различности тем используется регуляризатор декоррелирования.

Для определения количества новых тем для каждого временного шага использовалась метрика на основе относительного изменения количества токенов в словаре на текущем временном шаге, регулируемая гиперпараметром β. Это было сделано из предположения о том, что вместе с новыми темами так же появляются новые слова или начинают увеличиваться в употреблении уже известные.

На выходе модели каждой теме соответствуют ранжированные списки документов D_{topic} и ключевых слов W_{topic}. Валидационный датасет, предложенный в этой работе, также состоит из множества трендов, которым соответствуют ранжированные списки D_{trend} и ключевых слов W_{trend}, а так же названия трендов S_{trend}, для полученных моделью тем мы используем ключевые слова как возможные варианты названия тренда $S_{topic} := W_{topic}$. Чтобы сопоставить результаты реальным трендам, мы считаем три метрики Recall@k:

[1] *ПАО “Сбербанк”, Москва, Россия*
[2] *Федеральный исследовательский центр “Информатика и управление” Российской академии наук, Москва, Россия*
*****E-mail: nikgerasimenko@gmail.com*

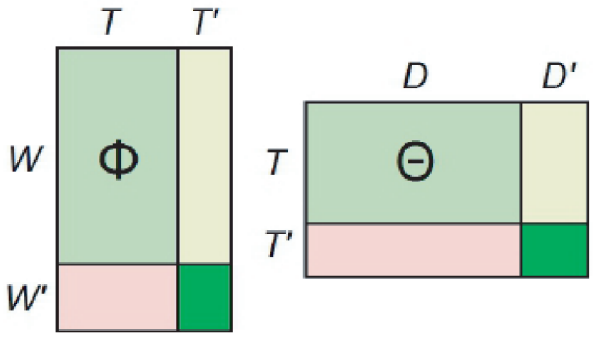

Рис. 1. Инкрементная тематическая модель. Нулевые блоки выделены красным цветом, а сильно разреженные – светло-зеленым.

$$XRecall@k = \frac{\left|X_{topic}[:k] \cap X_{trend}\right|}{K},$$

где X[:k] – первые k элементов списка X, а X заменяется на W, D, S. Для подсчета метрики для документов, слов и названий используются разные значения k, которые обозначаются как k_D, k_W и $k_S \le k_W$ соответственно.

Мы провели серию экспериментов, где рассмотрели вероятностные тематические модели, такие как PLSA, LDA и ARTM с декоррелирующим регуляризатором матрицы Φ, и нейронные сети, в частности BERTopic. Несмотря на то что BERTopic поддерживает динамическое тематическое моделирование, модель не соответствует нашим целям и критериям. Сначала BERTopic создает общую тематическую модель, как если бы в документах не было временного аспекта. Затем для каждой темы и временного шага он вычисляет представление c-TF-IDF, что приводит к различным формулировкам одних и тех же тем на разных временных шагах.

Мы сравнили наше решение с вышеперечисленными на основе трех конфигураций, включающих в себя следующие параметры:

1) Config1: D$Recall@k > 0.1$

2) Config2: W$Recall@k > 0.3$ and S$Recall@k > 0$

3) Config3: D$Recall@k > 0.1$, W$Recall@k > 0.3$ and S$Recall@k > 0$

Здесь Config1 соответствует сопоставлению извлеченных тем и трендов по документам, Config2 – только по ключевым словам, и Config3 объединяет в себе две предыдущие опции.

Модель BERTopic достигла наилучших результатов для Config1, и выявила почти все тренды: 90 из 91. Это связано с тем, что модель имеет большее количество тем и способна успешно различать документы среди них. Однако модель показывает результаты хуже для извлечения ключевых слов в других конфигурациях, так как это не является ее основной задачей.

ARTM детектирует правильные темы достаточно быстро даже в наиболее сложной конфигурации Config3, хотя в некоторых случаях может извлекать суммарно меньше трендов. В конфигурации Config1, основной целью которой является правильное разделение документов по темам, ARTM извлекает почти половину трендов за первые два месяца. Таким образом, модель подходит для качественного выявления трендов в задаче ранней детекции. Также стоит отметить, что модели BERTopic и ARTM способны извлекать тренд прямо в момент его возникновения для Config1. Это связано с тем, что некоторые из трендов относятся к типу “задача” и не имеют конкретного первого документа.

В нашем подходе есть несколько вариантов выбора набора данных для переобучения на каждом этапе. Эксперименты проводились для двух возможных вариантов: обучение по всей истории документов и обучение только по новым документам (инкрементально). Мы обозначили подход с инкрементальным обучением ARTMi. По результатам экспериментов ARTMi извлекает в общей сложности больше трендов, чем ARTM во всех рассматриваемых конфигурациях.

В своей работе мы исследовали задачу ранней детекции научных трендов. Мы адаптировали стандартный подход, основанный на ARTM, и предложили инкрементальное обучение, состоящее из инкрементальной инициализации, инкрементального набора данных и обновления количества тем на основе текущего словаря трендовых словосочетаний. Кроме того, мы включили дополнительную регуляризацию разреженности в наш подход для достижения наилучших результатов. Наш подход универсален и не зависит от конкретной модели.

Эксперименты показали, что базовая модель ARTM получила один из лучших результатов по

сравнению с другими базовыми моделями во всех рассмотренных конфигурациях подсчета качества. Более того, методы инкрементального обучения и дополнительная регуляризация позволили значительно улучшить качество. Итоговый подход, основанный на ARTM, способен выделить наибольшее количество трендов на ранних стадиях их развития и может работать в режиме реального времени.

СПИСОК ЛИТЕРАТУРЫ

1. *Kontostathis A., Galitsky M.L., Pottenger M.W., Roy S., Phelps J.D.* A Survey of emerging trend detection in textual data mining // Springer New York, 2004. p. 185–224.

2. *Vorontsov K., Frei O., Apishev M., Romov P., and Dudarenko M.* BigARTM: Open source library for regularized multimodal topic modeling of large collections // AIST, Springer International Publishing, 2015. p. 370–381.

EDN: EGEPTA
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 109–110

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ТЕХНОЛОГИИ КОМПЬЮТЕРНОГО ЗРЕНИЯ В ЗАДАЧАХ СИНТЕЗА ВЫСОКОКАЧЕСТВЕННОГО МУЛЬТИМЕДИЙНОГО КОНТЕНТА

© 2022 г. А. В. Кузнецов[1,*], Д. В. Димитров[1], А. Ю. Грошев[1], П. П. Парамонов[1], А. А. Мальцева[1]

Представлено академиком РАН А.Л. Семеновым
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

Развитие технологий глубокого обучения неизбежно порождает новые задачи и их решения в таких направлениях, как компьютерное зрение, VR/AR технологии, видеоаналитика, мультимодальное обучение и др. С ростом доступности высокопроизводительных вычислительных устройств многие современные методы и средства обработки цифровых данных становятся широко применимыми в том числе в рамках частных прикладных исследований. Данную тенденцию можно легко проследить по росту количества open-source решений, которые без труда запускаются на таких известных ресурсах, как, например, Google Colab. В рамках данного материала мы поделимся полученными результатами в части разработки и исследования прорывных технологий синтеза высококачественного мультимедийного контента, которые имеют широкое применение в таких задачах, как перенос лица.

Ключевые слова: перенос лица, GHOST, one shot, синтез фото, синтез видео
DOI: 10.31857/S2686954322070141

Существование технологии автоматического переноса лица человека на фото или видеоконтент всегда было и будет объектом спора научных и общественных групп по той причине, что алгоритмы создания deep fake контента часто звучат в СМИ в негативном контексте и описываются как средства создания компромата, манипуляции общественным мнением, дезориентации общества в части интерпретации каких-либо событий. Все это в первую очередь несет репутационные риски для физических, юридических лиц, так и государственных структур. Более того, даже сам термин “fake”, входящий в общеупотребимое словосочетании “deep fake”, обозначает подделку, что безусловно вызывает естественное негативное восприятие этого слова. Более того, в условиях распространенной в различных источниках информации о парадигме “цифровой гигиены” или чистоте данных, общество все меньше начинает доверять контенту, демонстрируемому в сети Интернет, что приводит к очевидному снижению степени доверия к мультимедийным данным, в особенности новостного характера.

Несмотря на потенциальный вред, который может нести технология синтеза мультимедийного контента, аппарат таких методов, как перенос лиц на фото или видео безусловно позволяет приносить пользу в различных задачах: съемка фильмов и видеороликов с участием актеров, которые по тем или иным причинам не могут физически присутствовать на съемках, создание нового образовательного, развлекательного и рекламного контента для привлечения пользователей, повышение качества мультимедийного контента и т.д. Стоит перенести фокус внимания с термина “fake” на технологические преимущества, которые дают методы переноса лиц, как становятся понятны очевидные плюсы для общества.

Наша команда давно занимается технологией переноса лица, и одним из главных достижений является алгоритм GHOST (Generative High-fidelity One Shot Transfer) [1]. Он позволяет выполнять перенос лица всего лишь с одного изображения-источника на целевые изображение или видео. Стоит отметить, что превалирующее большинство существующих методов решают задачу переноса лица на видео посредством использования набора кадров, на которых обучается специальная модель извлечения признаков. В основе решения лежит базовая архитектура FaceShifter [2] (перенос лица с изображения на изображение), которая была значительно улучшена в ходе исследований за счет таких нововведений, как функция потерь для области глаз, алгоритм сглаживания маски лица, алгоритм замены лица на видео,

[1] *Sber AI, Москва, Россия*
***E-mail: AVladimirKuznetsov@sberbank.ru*

а также новый метод стабилизации ключевых точек лица для уменьшения его дрожания на соседних кадрах и этап повышения разрешения. Все это позволило обойти существующие SoTA решения по известным метрикам на 1–2%, а также снизить вычислительную сложность технологии переноса за счет One Shot подхода.

Как мы уже сказали ранее, модель, способная генерировать качественный синтезированный фото и видео контент, при неправильном умысле может нести риски особенно в эпоху информационных противостояний. Поэтому, чувствуя бремя этой большой ответственности, мы разработали модель обнаружения синтезированного моделью GHOST фото и видео контента, которая в отличие от существующих в открытом доступе детекторов deepfake с высокой точностью определяет контент, сгенерированный алгоритмом GHOST. Не ставя перед собой задачу поиска научной новизны, а рассматривая детекцию как чисто инженерную задачу, мы взяли за основу модель [3], которая выиграла на соревновании Kaggle в 2020 г. Архитектура состоит из сверточной сети извлечения векторов признаков изображений EfficientNet-B7 и добавленного слоя классификатора. Данная модель предобучалась на данных, предоставленных организаторами соревнования DeepFakeDetectionChallenge (DFDC). Датасет DFDC – это огромный набор оригинальных видео и полученных на их основе дипфейков с помощью различных доступных на тот момент алгоритмов генерации. В сжатом виде этот датасет занимает почти 500 Гб. Несмотря на то что автором лучшего решения была проделана огромная кропотливая работа по предобработке данных и различным аугментациям, его модель была ориентирована только на дипфейки, представленные в рамках DFDC, а для других методов, включая наш метод GHOST, она была нечувствительна. В ходе экспериментов на основе синтезированной моделью GHOST выборке мы получили обновленные веса модели детекции, которые позволяют значительно повысить качество (F1_Score) обнаружения контента, синтезированного моделью GHOST, с 0.14 до 0.98, сохранив при этом исходные значения качества обнаружения других способов создания deepfake фото и видео.

В заключение хочется отметить, что важно устанавливать нормативные рамки разработки любой технологии, чтобы минимизировать возможные риски использования ее злоумышленниками. Мы показали важность развития аппарата методов синтеза мультимедийного контента для решения полезных обществу задач и планируем дальше повышать качество модели GHOST путем учета оценки положения головы в трехмерном пространстве для улучшения переноса лица в экстремальных углах поворота.

СПИСОК ЛИТЕРАТУРЫ

1. *Groshev A. et al.* GHOST – A New Face Swap Approach for Image and Video Domains // IEEE Access. 2022. Т. 10. С. 83452–83462.
2. *Li L. et al.* Faceshifter: Towards high fidelity and occlusion aware face swapping // arXiv preprint arXiv:1912.13457. 2019.
3. *Das S. et al.* Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. С. 3776–3785.

EDN: SRJMOO
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 111–127

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ФОРМАЛИЗАЦИЯ ТЕОРИИ ПРОГРАММИРОВАНИЯ ПРИНЦИПОВ РАБОТЫ МОЗГА С ИНФОРМАЦИЕЙ

© 2022 г. Е. Е. Витяев[1,2,*], А. Г. Колонин[2], А. В. Курпатов[3], А. А. Молчанов[3]

Представлено академиком РАН С.С. Гончаровым
Поступило 05.11.2022 г.
После доработки 07.11.2022 г.
Принято к публикации 08.11.2022 г.

В монографии “Сильный искусственный интеллект. На подступах к сверхразуму” содержится обзор общего искусственного интеллекта (AGI). В качестве антропоморфного направления исследований, он включает Brain Principles Programming (BPP) – формализацию универсальных механизмов (принципов) работы мозга с информацией, которые реализуются на всех уровнях организации нервной ткани. В этой монографии содержится формализация этих принципов в терминах теории категорий. Однако этой формализации недостаточно для разработки алгоритмов работы с информацией. В данной работе для описания и моделирования BPP предлагается применять разработанные нами ранее математические модели и алгоритмы, моделирующие когнитивные функции, которые основаны на известных физиологических, психологических и других естественнонаучных теориях. В работе используются математические модели и алгоритмы следующих теорий: Теории Функциональных Систем работы мозга П.К. Анохина, прототипической теории категоризации Eleanor Rosch, теории причинных моделей Bob Rehder и “естественная” классификация. В результате получена формализация BPP и приведены компьютерные эксперименты, демонстрирующие работу алгоритмов.

Ключевые слова: мозг, классификация, кластеризация, теория функциональных систем, понятия, когнитивные функции

DOI: 10.31857/S2686954322070219

1. ВВЕДЕНИЕ

В монографии “Сильный искусственный интеллект. На подступах к сверхразуму” [1] приводится первое кросс-дисциплинарное исследование по общему искусственному интеллекту, где говорится, что “Общий искусственный интеллект – это следующая ступень в развитии ИИ, не обязательно наделенная самосознанием, но, в отличие от современных нейросетей, способная справляться с широким кругом задач в разных условиях”. В качестве антропоморфного направления исследований в ней рассматривается Brain Principles Programming (BPP) – формализация универсальных механизмов (принципов) работы мозга с информацией, сформулированная А.В. Курпатовым. В книге приводится формализация этих принципов в языке теории категорий. Однако из этой формализации не следуют алгоритмы работы с информацией.

В данной работе сделана попытка применить разработанные ранее математические модели и алгоритмы, моделирующие когнитивные функции и опирающиеся на известные физиологические, психологические и другие естественно-научные теории, для описания и моделирования Brain Principles Programming. Мы будем опираться на следующие теории: Теорию Функциональных Систем работы мозга П.К.Анохина [2–4], теорию интегрированной информации G. Tononi [5], прототипическую теорию категоризации Eleonor Rosh [6–8], теорию причинных моделей Bob Rehter [9–11] и работы по “естественной” классификации [12].

В части I мы приведем математические модели и алгоритмы этих теорий, а затем в части II формализацию BPP в терминах этих моделей и алгоритмом, опираясь на формализацию BPP в терминах теории категорий, приведенную в [1].

Начнем формализацию с некоторых элементарных единиц восприятия внешнего мира. В прототипической теории категоризации и теории

[1] *Институт математики им. С.Л. Соболева, Новосибирск, Россия*
[2] *Новосибирский государственный университет, Новосибирск, Россия*
[3] *ПАО Сбербанк, Лаборатория нейронаук и поведения человека, Москва, Россия*
*****E-mail: vityaev@math.nsc.ru*

“естественных” понятий Eleonor Rosh ими являются “естественные” понятия и прототипы классов, в теории причинных моделей Bob Rehter – причинные модели, в теории интегрированной информацией G. Tononi – концепты, формирующиеся в сознании в виде систем причинных связей с высоко интегрированной информацией.

Эти единицы восприятия описываются в этих подходах как субъективные единицы, однако есть теория, которая описывает объективные свойства объектов, проявляющиеся в этих единицах восприятия – это “естественная” классификация. Далее мы приводим краткое описание этих теорий, начиная с “естественной” классификации и показываем, что все они могут быть формализованы с помощью вероятностных формальных понятий, приводимых далее. Затем приводится метод обнаружения вероятностных формальных понятий.

Часть I. Базовые теории и формальные модели

2. БАЗОВЫЕ ЭЛЕМЕНТЫ ВОСПРИЯТИЯ, СОЗНАНИЯ И МИРА

2.1. “Естественная” классификация. Она описывает способ “естественного” формирования понятий об объектах внешнего мира и, как будет показано далее, соответствует исследованиям по формированию “естественных” понятий в когнитивных науках. “Естественная” классификация опирается на объективные свойства внешнего мира и позволяет понять процесс отражения реальности в субъективном опыте.

Первый достаточно подробный философский анализ “естественной” классификации принадлежит Дж.Ст. Миллю [13]. По Дж.Ст. Миллю “искусственные” классификации отличаются от “естественных” тем, что они могут быть основаны на любом одном или нескольким признаках, так что разные классы различаются только тем, что включают объекты, обладающие различными значениями этих признаков. Но если рассмотреть классы “животных” или “растений”, то они отличаются столь большим (потенциально бесконечным) количеством свойств, что их нельзя перечислить. И все эти свойства будут основаны на утверждениях, подтверждающих это различие.

Дж.Ст. Милль дает следующее определение “естественной” классификации: это такая классификация, которая ... основывается на таких свойствах, которые служат причинами многих других или по крайней мере составляют их верные признаки. Он определяет также понятие “образа” класса, которое является предтечей “естественных” понятий в когнитивных науках: “наше понятие о классе – тот образ, которым этот класс представлен в нашем уме, – есть понятие о некотором образце, обладающем всеми признаками данного класса ... в самой высокой степени”.

Рассуждения Дж.Ст. Милля были подтверждены естествоиспытателями. О схожести свойств у “естественных” классов пишет Л. Рутковский [14]: “Чем в большем числе существенных признаков сходны сравниваемые предметы, тем вероятнее их одинаковость и в других отношениях”. Е.С. Смирнов [15]: “Таксономическая проблема заключается в “индикации”: от бесконечно большого числа признаков нам нужно перейти к ограниченному их количеству, которое заменило бы все остальные признаки”.

Из исследований по “естественной” классификации следует, что признаки в “естественных” классах сильно коррелированы. Например, если у нас есть 128 классов и признаки двоичные, то независимыми “индикаторными” признаками среди них могут быть только 7 признаков, потому что $2^7 = 128$, а остальные 121 признаков, в силу схожести свойств по Рутковскому, могут быть предсказаны по этим 7 признакам, что означает наличие для них 121 закономерности. Так как “индикаторными” признаками могут быть разные 7 признаков, выбранные из 128, и для каждого набора выбранных 7 признаков есть 121 закономерность, предсказывающая все остальные признаки, то общее число закономерностей может быть не намного меньше, чем $121 \cdot C_{128}^{7} = 11437621219200$. В работе [12] приведены формальная модель “естественной” классификации и пример ее построения.

2.2. “Естественные” понятия и прототипическая теория категоризации. Высокая коррелированность признаков для “естественных” классов была подтверждена в когнитивных исследованиях. В работах Eleanor Rosch [6–8] был сформулирован следующий принцип категоризации “естественных” категорий, подтверждающий высказывания Дж.Ст. Милля и естествоиспытателей: “Perceived World Structure. ... perceived world – is not an unstructured total set of equiprobable co-occurring attributes. Rather, the material objects of the world are perceived to possess ... ***high correlational structure*** (Курсив ЕЕ). ... In short, combinations of what we perceive as the attributes of real objects do not occur uniformly. Some pairs, triples, etc., are quite probable, appearing in combination sometimes with one, sometimes another attribute; others are rare; others logically cannot or empirically do not occur”.

Непосредственно воспринимаемые объекты (basic objects) – информационно богатые связки наблюдаемых и функциональных свойств, которые образуют естественную разрывность, создающую категоризацию. Эти связки формируют “прототипы” объектов классов (образ у Дж.Ст. Милля). В дальнейшем теория “естественных” понятий Eleanor Rosch получила название прототипической теории понятий (prototype theory).

2.3. Теория причинных моделей. В дальнейших исследованиях было обнаружено, что моделей, основанных на признаках, сходстве и прототипах, недостаточно для описания классов. Необходимо учитывать теоретические, причинные и онтологические знания, относящиеся к объектам классов. Например, люди не только знают, что птицы имеют крылья, могут летать и вить гнезда на деревьях, но также и то, что птицы вьют гнезда на деревьях, потому что могут летать, и летать, потому что имеют крылья.

Исследования показали, что знания людей о категориях не сводятся к перечню свойств, а включают богатое множество причинных связей между этими свойствами. Важность свойств категории зависит от их причинных взаимосвязей. В некоторых экспериментах [11] было показано, что свойство важнее, если оно сильнее включено в причинную сеть взаимосвязей признаков.

Учитывая эти исследования, Bob Rehder выдвинул теорию причинных моделей (causal-model theory), в соответствии с которой отношение объекта к категории основывается уже не на множестве признаков и близости по признакам, а на основании сходства порождающего причинного механизма [9]. Для представления причинного знания были использованы Байесовские сети [10]. Однако они не могут моделировать циклические причинные связи, потому что Байесовские сети не поддерживают циклов. Разработанные нами вероятностные формальные понятия, приведенные далее, непосредственно моделируют циклические причинные связи с помощью неподвижных точек предсказаний по причинным связям.

2.4. Теория интегрированной информации G. Tononi. Сознание как интегрированная информация. Если “естественная” классификация описывает объекты внешнего мира, а когнитивные науки – восприятие объектов внешнего мира, то теория интегрированной информации сознания G. Tononi анализирует информационные процессы мозга по восприятию объектов внешнего мира.

G. Tononi определяет сознание как первичное понятие, которое обладает следующими феноменологическими свойствами: composition, information, integration, exclusion [5]. Для более точного определения этих свойств G. Tononi вводит понятие интегрированной информации: “это информация, генерируемая системой, которая приходит в определенное состояние после причинно-следственного взаимодействия между ее частями, которая превосходит информацию, генерируемую независимо самими ее частями” [5]. В терминах интегрированной информации феноменологические свойства формулируются следующим образом. В скобках мы приводим интерпретацию этих свойств с точки зрения “естественной” классификации.

• composition – elementary mechanisms (causal interactions) can be combined into higher-order ones (“естественные” классы формируются в виде причинных циклов и иерархии “естественных” классов);

• information – only mechanisms that specify ‘differences that make a difference’ within a system count (только система “резонирующих” причинных связей, формирующая класс, является значимой. См. иллюстрацию на примере ниже);

• integration – only information irreducible to non-interdependent components counts (значима только система “резонирующих” причинных связей, свидетельствующая об избытке информации и восприятии высоко коррелированной структуры “естественного” объекта);

• exclusion – only maxima of integrated information count (только значения признаков, которые максимально взаимосвязаны причинными связями формируют “образ” или “прототип”).

Поскольку у G. Tononi нет внешнего мира и его “естественной” классификации, то приведенные свойства определяются как внутренние свойства системы. Мы рассмотрим эти свойства не как внутренние свойства системы, а как способность системы отражать комплексы причинных связей объектов, а сознание – как способность комплексного иерархического отражения “естественной” классификации внешнего мира.

3. ВЕРОЯТНОСТНЫЕ ФОРМАЛЬНЫЕ ПОНЯТИЯ И ИХ ОБНАРУЖЕНИЕ

Нами выдвигается гипотеза о том, что “естественная” классификация, “естественные” понятия и интегрированная информация G. Tononi описываются одним и тем же формализмом. С нашей точки зрения информационные процессы работы мозга и сознание настроились в процессе эволюции на извлечение высоко коррелированной структуры признаков “естественных” объектов путем формирования “естественных” понятий объектов. Мозг с помощью интегрированной информации настраивается на восприятие “естественных” объектов внешнего мира, отражая их высоко коррелированную структуру. Причинные связи при восприятии “естественных” объектов замыкаются на себя, образуя определенный “резонанс”, что является системой с высоко интегрированной информацией в смысле G. Tononi. При этом “резонанс” возникает тогда и только тогда, когда эти причинные связи отражают некоторый целостный “естественный” объект, в котором потенциально бесконечное множество признаков взаимно предсказывают друг друга. Возникающие при этом циклы выводов по при-

чинным связям математически описываются "неподвижными точками" взаимно предсказывающихся свойств, что дает "образ" класса и "прототип" объекта. Поэтому мозг воспринимает "естественный" объект не набором признаков, а как "резонирующую" систему причинных связей. Ниже приведен пример моделирования обнаружения "естественных" классов, "естественных" понятий и интегрированной информации на примере закодированных цифр.

Приведем формализацию циклических причинных связей в виде вероятностных формальных понятий [16–18]. Одновременно будет дано определение Максимально Специфических Вероятностных Причинных Связей (МСВПС), для которых доказано, что логический вывод по ним и, соответственно, вывод предсказаний непротиворечив [19, 20]. В теории G. Tononi явно не сказано, каким биологическим субстратом моделируются причинные связи – нейронами, кортикальными колонками или как-то еще. Мы будем предполагать, что причинные связи обнаруживаются нейронами в соответствии с формальной моделью нейрона, изложенной в [21], которая обнаруживает МСВПС. Кроме того, определенные ниже МСВПС удовлетворяют определению Cartwright [22] вероятностной причинности. "Резонанс" причинных связей в виде неподвижных точек предсказаний по МСВПС причинным связям дан ниже в определении 18, что сразу же приводит к вероятностным формальным понятиям в определении 19. Как будет показано далее, вероятностные формальные понятия в то же время моделируют понятие контекста.

Приведенная здесь формализация вероятностных формальных понятий следует работам [16–18, 23].

Определение 1. *Формальный контекст* $\mathrm{K} = (\mathrm{G}, \mathrm{M}, \mathrm{I})$ представляет собой тройку, где G и M – произвольные наборы объектов и атрибутов, и $I \subseteq G \times M$ – бинарное отношение, выражающее принадлежность атрибута объекту.

В формальном контексте операторы производных связывают подмножества объектов и атрибутов контекста.

Определение 2. $A \subseteq G, B \subseteq M$, тогда:

$$A^{\uparrow} = \{m \in M \mid \forall g \in A, (g,m) \in I\}$$

$$B^{\downarrow} = \{g \in G \mid \forall m \in B, (g,m) \in I\}$$

Определение 3. Пара (A, B) – формальное понятие, если $A^{\uparrow} = B$ и $B^{\downarrow} = A$.

Переопределим контекст в логических терминах. Будем рассматривать только конечные контексты.

Определение 4. Для контекста $\mathrm{K} = (\mathrm{G}, \mathrm{M}, \mathrm{I})$ определяем сигнатуру Ω_K контекста, которая содержит символы предикатов $m(x)$ для каждого $\mathrm{m} \in \mathrm{M}$, $K \vDash m(x) \Leftrightarrow (x,m) \in I$.

Определение 5. Для сигнатуры Ω_K определим следующий вариант логики первого порядка:

1. X_K – множество переменных;

2. At_K – множество атомарных формул (атомов) $\mathrm{m}(x)$, $m \in \Omega_K$, $x \in X_K$;

3. L_K – набор литералов, включающий атомы $\mathrm{m}(\mathrm{t})$ и их отрицания $\neg m(t)$;

4. Φ_K – набор формул, определяемый индуктивно: литерал – формула, для $\Phi, \Psi \in \Phi_K$ выражения $\Phi \wedge \Psi$, $\Phi \vee \Psi$, $\Phi \to \Psi$, $\neg\Phi$ – также формулы.

Определим коньюнкцию $\wedge L$ и отрицание $\neg L = \{\neg P \mid P \in L\}$ набора литералов $L \subseteq \mathrm{L}_K$.

Определение 6. Единичный элемент $\{\mathrm{g}\}$, $g \in G$, представленный в сигнатуре Ω_K, образует модель K_g этого объекта. Истинность формулы ϕ на модели K_g определяется как $g \vDash \phi \Leftrightarrow K_g \vDash \phi$.

Определение 7. Определим *вероятностную меру* μ на множестве G в смысле Колмогорова. Тогда мы можем определить вероятностную меру на множестве формул как:

$$\nu : \Phi_K \to [0,1], \quad \nu(\phi) = \mu(\{g \mid g \vDash \phi\}).$$

Мы предполагаем, что в контексте нет несущественных объектов, таких что $\mu(\{g\}) = 0$, $g \in G$.

Определение 8. Пусть $\{H_1, H_2, \ldots, H_k, C\} \in \mathrm{L}_K$, $C \notin \{H_1, H_2, \ldots H_k\}$, $k \geq 0$.

Отношение есть $R = (H_1 \wedge H_2 \wedge \ldots \wedge H_k \to C)$;

Посылка $R^{\leftarrow}$ отношения R – это набор литер $\{\mathrm{H}_1, \mathrm{H}_2, \ldots, \mathrm{H}_k\}$;

Заключение отношения это $R^{\rightarrow} = C$;

Длина отношения это $|R^{\leftarrow}|$;

Определение 9. *Вероятность* η отношения R – это величина

$$\eta(R) = \nu(R^{\rightarrow} \mid R^{\leftarrow}) = \nu(R^{\leftarrow} \wedge R^{\rightarrow})/\nu(R^{\leftarrow}).$$

Если знаменатель $\nu(R^{\leftarrow})$ отношения равен 0, то вероятность не определена.

Определение 10. Отношение R_1 является *подотношением* отношения R_2, обозначается как $R_1 \sqsubset R_2$, если $R_1^{\rightarrow} = R_2^{\rightarrow}$, $R_1^{\leftarrow} \subset R_2^{\leftarrow}$.

Определение 11. Отношение R_1 *уточняет* отношение R_2, обозначим как $R_2 < R_1$, если $R_2 \sqsubset R_1$ и $\eta(R_1) > \eta(R_2)$.

Определение 12. Отношение R является *вероятностной причинной связью*, если для каждого $\tilde{R}$ выполнено $(\tilde{R} \sqsubset R) \Rightarrow (\tilde{R} < R)$.

Определение вероятностной причинности, данное Cartwright [22] относительно некоторого бэкграунда, может быть сформулировано в приведенных терминах следующим образом. Если посылкой $R^{\leftarrow}$ отношения R является набором литералов $\{H_1, H_2, ..., H_k\}$ и мы рассматриваем этот набор как бэкграунд, то каждый литерал посылки является вероятностной причиной заключения $R^{\rightarrow}$ отношения R относительно этого бэкграунда, то есть

$$\nu(R^{\rightarrow}/R^{\leftarrow}) > \nu(R^{\rightarrow}/(R^{\leftarrow}\backslash H))$$
$$\text{для каждого } H \in \{H_1, H_2, ..., H_k\}.$$

Легко видеть, что это определение следует из определения 12.

Определение 13. *Сильнейшей вероятностной причинной связью* будет называться отношение R, для которого не существует такой вероятностной причинной связи $\tilde{R}$, что $(\tilde{R} > R)$.

Определение 14. *Семантический Вероятностный Вывод* (СВВ) предсказаний некоторого литерала C есть последовательность вероятностных причинных связей $R_0 < R_1 < R_2 ... < R_m$, $R_0^{\rightarrow} = R_1^{\rightarrow} = R_2^{\rightarrow} ... = R_m^{\rightarrow} = C$, $R_0^{\leftarrow} = \varnothing$, R_m – сильнейшая вероятностная причинная связь.

Определение 15. *Дерево семантического вероятностного вывода* Tree(C) некоторого литерала C – это совокупность всех СВВ, предсказаний литерала C.

Определение 16. *Максимально специфичное причинное отношение* для предсказания некоторого C – это сильнейшее вероятностное причинное отношение дерева Tree(C), имеющее максимальную условную вероятность.

Обозначим через MSCR множество всех максимально специфичных причинных отношений. Под *системой причинных отношений* будем понимаем любое подмножество $\mathcal{R} \subseteq$ MSCR.

Определение 17. Определим *оператор предсказания* для системы $\mathcal{R}$ как:

$$\Pi_{\mathcal{R}}(L) = L \cup \{C \mid \exists R \in \mathcal{R} : R^{\leftarrow} \subseteq L, R^{\rightarrow} = C\}.$$

Определение 18. *Замыканием* набора литералов L назовем наименьшую неподвижную точку оператора предсказания, содержащую L:

$$\Pi_{\mathcal{R}}^{\infty}(L) = \bigcup_{k \in \mathbb{N}} \Pi_{\mathcal{R}}^{k}(L).$$

Набор литералов L *непротиворечив*, если он не содержит одновременно атом C и его отрицание $\neg C$. Набор литералов L *совместен,* если $\nu(\wedge L) \neq 0$.

Теорема 1. [18, 19]. Если L – совместно, то $\Pi_{\mathcal{R}}(L)$ совместно и непротиворечиво для любой системы $\mathcal{R}$.

Определение 19. *Вероятностное формальное понятие* на контексте K – это пара (A, B), удовлетворяющая следующим условиям:

$$\Pi_{\mathcal{R}}^{\infty}(B) = B, \quad A = \bigcup_{\Pi_{\mathcal{R}}^{\infty}(C)=B} C^{\downarrow}.$$

Определение множества A основано на следующей теореме, связывающей вероятностные и стандартные формальные понятия на контексте K.

Теорема 2. [18, 19]. Пусть $K = (G, M, I)$ – формальный контекст, тогда:

Если (A,B) – формальное понятие на K, то существует вероятностное формальное понятие (S,T) на K такое, что $A \subseteq S, B \subseteq T$.

Если (S,T) – вероятностное формальное понятие на K, то существует семейство $\mathcal{C}$ формальных понятий на K, такое что

$$\forall (A, B) \in \mathcal{C}(\Pi_{\mathcal{R}}^{\infty}(B) = T), \quad S = \bigcup_{(A,B)\in\mathcal{C}} A.$$

4. АЛГОРИТМ СТАТИСТИЧЕСКОЙ АППРОКСИМАЦИИ ВЕРОЯТНОСТНЫХ ФОРМАЛЬНЫХ ПОНЯТИЙ

В практических задачах мы не можем предполагать, что вероятностная мера нам известна. Поэтому нам необходимо использовать некоторый статистический критерий для определения вероятностных неравенств в семантическом вероятностном выводе и обнаружении МСВПС [12, 24]. Для этого мы используем точный критерий независимости Фишера с уровнем значимости α. Результирующий набор $\mathcal{R}_\alpha$ вероятностных максимально специфических причинно-следственных связей, полученный с уровнем значимости α, может вызывать противоречия в неподвижных точках вероятностных формальных понятий. Следовательно, для аппроксимации оператора $\Pi_{\mathcal{R}}(L)$ необходимо ввести дополнительный критерий согласованности максимально специфических причинных связей $\mathcal{R}_\alpha$ на множестве L.

Определение 20. Причинное отношение $R \in \mathcal{R}_\alpha$ *подтверждается* на множестве литералов L, если $R^{\leftarrow} \subset L$ и $R^{\rightarrow} \in L$. Тогда $R \in \text{Sat}(L) \subseteq \mathcal{R}_\alpha$.

Определение 21. Причинное отношение $R \in \mathcal{R}_\alpha$ *опровергается* на множестве литералов L, если $R^{\leftarrow} \subset L$ и $R^{\rightarrow} \in \neg L$. Тогда $R \in \text{Fal}(L) \subseteq \mathcal{R}_\alpha$.

Теперь мы можем определить критерий максимальной согласованности предсказаний по максимально специфическим причинным связям $\mathcal{R}_\alpha$ на некотором множестве литералов L.

Определение 22. *Критерием максимальной согласованности предсказаний* по максимально специфическим причинным связям $\mathcal{R}_\alpha$ на множестве литералов L является значение:

$$\mathrm{Int}(L) = \sum_{R\in \mathrm{Sat}(L)} \gamma(R) - \sum_{R\in \mathrm{Fal}(L)} \gamma(R).$$

Выбор оценки причинной связи γ может зависеть от специфики задачи. В наших экспериментах мы руководствовались соображениями Шеннона:

$$\gamma(R) = -\log(1 + \epsilon - \eta(R)), \quad \epsilon > 0, \quad \epsilon \ll 1.$$

Теперь мы можем аппроксимировать оператор $\Pi_{\mathcal{R}}(L)$, используя критерий согласованности предсказаний.

Определение 23. Определим *оператор максимальной согласованности предсказаний* $\Upsilon(L)$ для множества $\mathcal{R}_\alpha$ максимально специфических причинных связей, который аппроксимирует оператор $\Pi_{\mathcal{R}}(L)$. Он изменяет набор литер L на один элемент так, чтобы строго увеличить критерий $\mathrm{Int}(L)$:

1. Для всех $G \in L_K \backslash L$ вычислить максимальное увеличение критерия от добавления G к L: $\Delta^+ = \mathrm{Int}(L \cup \{G\}) - \mathrm{Int}(L)$ при условии, что в $\mathrm{Sat}(L)$ есть закономерность $R \in \mathrm{Sat}(L)$ такая, что $R^{\leftarrow} \subset L$ и $R^{\rightarrow} = G$.

2. Для всех $G \in L$ вычислить максимальное увеличение критерия от удаления G из L: $\Delta^- = \mathrm{Int}(L\backslash\{G\}) - \mathrm{Int}(L)$;

3. Оператор $\Upsilon(L)$ добавляет литерал G к L, если $\Delta^+ > 0$ и $\Delta^+ > \Delta^-$;

4. Оператор $\Upsilon(L)$ удаляет литерал G из L, если $\Delta^- > 0$ и $\Delta^- > \Delta^+$.

5. Если $\Delta^- = \Delta^+$ и $\Delta^- > 0$, оператор $\Upsilon(L)$ удаляет литерал G;

6. Если $\Delta^+ \le 0$ и $\Delta^- \le 0$, оператор $\Upsilon(L)$ возвращает L и, следовательно, мы получили неподвижную точку оператора максимальной согласованности предсказаний.

Определение 24. Под *статистической аппроксимацией вероятностных формальных понятий* контекста K для максимально специфических причинных связей $\mathcal{R}_\alpha$ мы понимаем набор всех неподвижных точек $\Upsilon^\infty(L)$, которые могут быть получены в результате многократного применения оператора $\Upsilon(L)$ к некоторому набору литералов L, представляющему некоторый объект $L = \{g\}^{\uparrow}$.

Докажем, что предельном случае, когда множество закономерностей $\mathcal{R}_\alpha$ совпадает с системой причинных отношений $\mathcal{R} \subseteq \mathrm{MSCR}$, неподвижная точка оператора предсказания $\Pi^\infty_{\mathcal{R}}(B)$ и оператора максимальной согласованности предсказаний $\Upsilon(L)$ совпадают. Поэтому статистическая аппроксимация вероятностных формальных понятий является прямым обобщением исходных вероятностных формальных понятий на случай работы с зашумленными данными.

Теорема 3. Пусть $\mathcal{R}_\alpha = \mathcal{R} \subseteq \mathrm{MSCR}$. Тогда для любого совместного набора литер L $\Upsilon^\infty(L) = \Pi^\infty_{\mathcal{R}}(L)$.

Доказательство: В силу теоремы 1 для $\mathcal{R}_\alpha = \mathcal{R}$ и L совместного у нас всегда будет $\mathrm{Fal}(L) = \varnothing$ на любом шаге применения оператора $\Upsilon(L)$. Тогда, поскольку $\gamma(R) = -\log(1 + \epsilon - \eta(R)) = -\log(\epsilon) > 0$, то $\mathrm{Int}(L) = \sum_{R\in \mathrm{Sat}(L)} \gamma(R) > 0$, при $\mathrm{Sat}(L) \neq \varnothing$.

Тогда всегда будет выполняться неравенство $\Delta^- = \mathrm{Int}(L\backslash\{G\}) - \mathrm{Int}(L) \le 0$, поскольку $\mathrm{Sat}(L\backslash G) \subseteq \mathrm{Sat}(L)$. Поэтому в соответствии с определением 23 оператор $\Upsilon(L)$ не будет удалять литеры из L.

С другой стороны, оператор $\Upsilon(L)$ будет добавлять новые литеры G в L при условии, что $\Delta^+ = \mathrm{Int}(L \cup \{G\}) - \mathrm{Int}(L) > 0$. Это значит, что $\mathrm{Sat}(L) \subset \mathrm{Sat}(L \cup \{G\})$ и существует закономерность $R \in \mathrm{Sat}(L \cup \{G\})$ такая, что $R^{\leftarrow} \subset L$ и $R^{\rightarrow} = G$. Это означает, что оператор $\Upsilon(L)$ всегда будет добавлять к L одну из литер G, для которой существует закономерность $R^{\leftarrow} \subset L$ и $R^{\rightarrow} = G$.

Таким образом, в нашем случае оператор $\Upsilon(L)$ можно записать так:

$$\Upsilon(L) = L \cup \frac{\arg\max \eta(G)}{\{G \mid \exists R \in \mathcal{R} : R^{\leftarrow} \subseteq L, R^{\rightarrow} = G,\}}.$$

Напомним, что оператор предсказания имеет аналогичный вид:

$$\Pi_{\mathcal{R}}(L) = L \cup \{C \mid \exists R \in \mathcal{R} : R^{\leftarrow} \subseteq L, R^{\rightarrow} = C\}.$$

Отличие операторов состоит только в последовательности добавления литер. Однако они всегда добавляют к множеству L те и только те литералы G, для которой существует отношение $\mathrm{R} \in \mathcal{R}$ такое, что $R^{\leftarrow} \subset L$ и $R^{\rightarrow} = G$. Поскольку порядок добавления литер не влияет на возможность включения других литер, то получаемые в результате неподвижные точки $\Upsilon^\infty(L)$ и $\Pi^\infty_{\mathcal{R}}(L)$ совпадают.

5. ОБНАРУЖЕНИЕ “ЕСТЕСТВЕННЫХ” ПОНЯТИЙ И КОНТЕКСТОВ

Приведем пример работы алгоритма статистической аппроксимации вероятностных формаль-

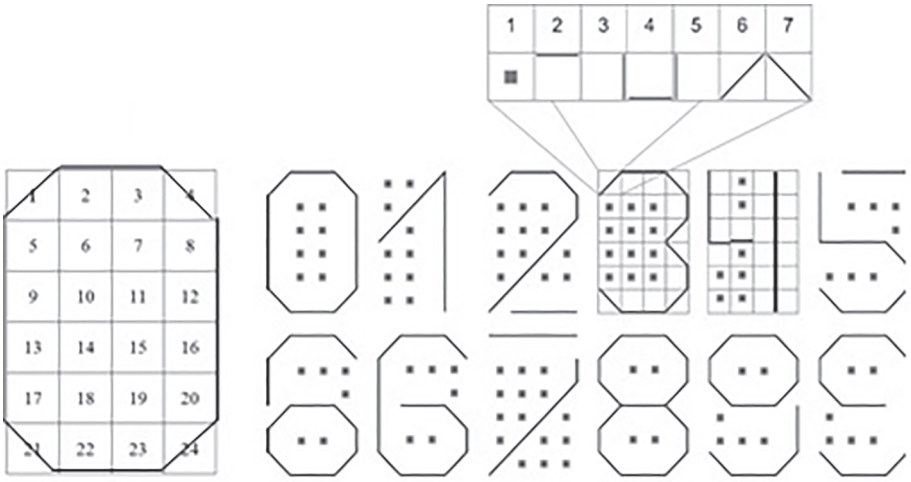

Рис. 1. Кодировка цифр.

ных понятий для некоторого контекста K = (G, M, I), где G – это множество закодированных цифр, как показано на рис. 1, M – это множество признаков цифр (см. рис. 1а) и I отношение, связывающее признаки и цифры. Для эксперимента было взято множество из 360 перетасованных цифр (12 цифр рис. 1 продублированных в 30 экземплярах без указания, где какая цифра). На этом множестве было обнаружено множество $\mathcal{R}_\alpha$ из 55089 вероятностных максимально специфических причинно-следственных связей, полученных с уровнем значимости $\alpha = 0.01$. Пусть L – это множество литералов, определенных для всех значений всех признаков. По причинно-следственным связям $\mathcal{R}_\alpha$ оператором $\Upsilon(L)$ было обнаружено 12 статистических аппроксимаций вероятностных формальных понятий точно соответствующих 12 цифрам.

Пример неподвижной точки для цифры 6 приведен на рис. 2. Рассмотрим, что представляет собой эта неподвижная точка. Занумеруем признаки цифр, как указано на рис. 1. Первая закономерность цифры 6 рис. 2, представленная в первом прямоугольнике после фигурной скобки, означает, что, если в квадрате 13 стоит признак 6 (обозначим 13-6), то в квадрате 3 должен стоять признак 2 (обозначим как (3-2)). Предсказываемый признак обозначается точечной линией. Запишем это отношение как (13-6 ⇒ 3-2). Нетрудно проверить, что это отношение выполнена на всех цифрах. Второе отношение означает, что из признака (9-5) и отрицания значения 5 первого признака ¬(1-5) (первый признак не должен быть равен 5) следует признак (4-7). Отрицание обозначается на рисунке пунктирной линией, как показано в нижней части рис. 2. Получим отношение (9-5&¬(1-5) ⇒ 4-7). Последующие 3 отношения в первой строке цифры 6 будут соответственно (13-6 ⇒ 4-7), (17-5&¬(13-5) ⇒ 4-7), (13-6 ⇒ 16-7).

На рис. 2 видно, что отношения и признаки цифры 6 образуют неподвижную точку – взаимно предсказывают друг друга. Заметим, что отношения, используемые в неподвижной точке, выполнены на всех цифрах, а сама неподвижная точка выделяет только одну цифру. Это иллюстрирует феноменологическое свойство 2 G. Tononi ‘differences that make a difference’, в котором система причинных связей воспринимает “осознает” целостный объект. Поэтому цифры выделяются не закономерностями самими по себе, а их системной взаимосвязью. Неподвижная точка формирует “прототип” по Eleanor Rosch или “образ” по Дж.Ст. Миллю. Программа не знает заранее, какие сочетания признаков максимально коррелируют между собой.

Важно отметить, что причинные связи в неподвижной точке предсказывают не только наличие некоторого другого признака в этой неподвижной точке, но и невозможность наличия какого-то другого признака в этой неподвижной точке. Таким образом, неподвижная точка характеризуется не только наличием признаков в соответствующих квадратах, но и необходимостью отсутствия признаков в каких-то других квадратах, т.е. неподвижная точка моделирует процесс выталкивания признаков и соответствующих прототипов других классов.

Формирование контекста на примере. Покажем, каким образом алгоритм статистической аппроксимации применим для обнаружения контекста. В исследовании приняло участие 2784 респондента, которые являются пользователями одной или нескольких социальных сетей. Учитывалось 36 характеристик, в особенности возраст, пол, число друзей, подписчиков и подписок, различ-

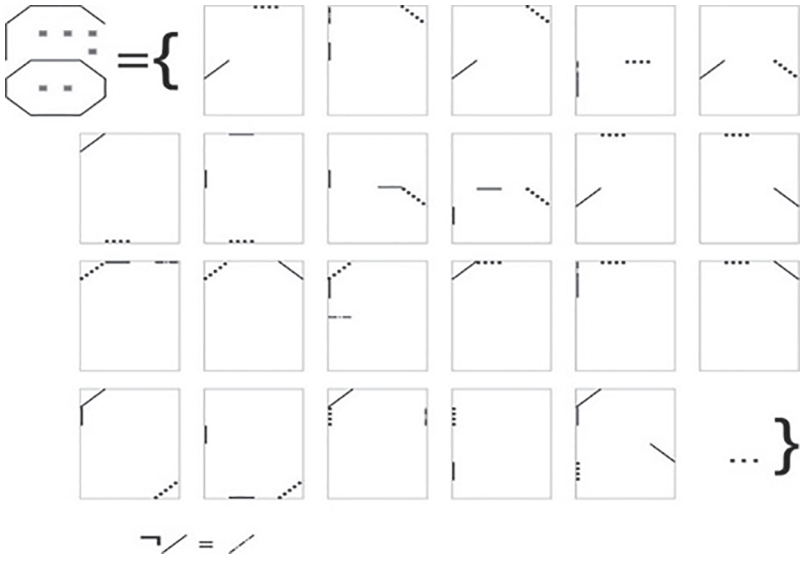

Рис. 2. Неподвижная точка цифры 6.

ных медиафайлов (заметки, фото, видео и т.д.), а также количество постов и лайков.

В результате работы алгоритма статистической аппроксимации была получена 21 неподвижная точка оператора $\Upsilon(L)$, которые в данном случае представляют собой контексты или типы пользователей. Приведём примеры наиболее характерных типов, выявленных алгоритмом. Первый - это замужняя женщина, проживающая в своем родном городе, для которой главное в жизни семья и дети, она ценит в людях доброту и честность, отношение к алкоголю и курению негативное, а число медиафайлов небольшое. Второй же можно охарактеризовать как неженатого мужчину, возможно подростка, который ценит в людях доброту и честность, но главное для него в жизни – это саморазвитие, при этом к курению и алкоголю он относится компромиссно.

Часть II. Формализация принципов программирования мозга (Brain Principles Programming)

6. “ИНТЕЛЛЕКТУАЛЬНЫЙ ОБЪЕКТ” И “ИНТЕЛЛЕКТУАЛЬНАЯ ФУНКЦИЯ”

В основе формализации “Brain Principles Programming, (BPP)”, изложенной в [1] и осуществленной в теории категорий лежат понятия “интеллектуальный объект” и “интеллектуальная функция”.

Приведем сначала неформальные определения “интеллектуального объекта” и “интеллектуальной функции” из [1, 25]:

• “интеллектуальный объект”, под которым мы понимаем любую единичную целостность, выделяемую нами в этом пространстве – например, когда мы видим стол, сигналы от зрительного нерва обрабатываются мозгом, и сочетание отдельных линий опознается как стол;

• “интеллектуальная функция”, которая описывает все возможные операции в рассматриваемой системе – это все, что психика может сделать с интеллектуальным объектом. Когда мы опознаем стол как объект, мы можем оценить его размер или придумать, как его использовать;

• “сущность” – специфическое значение объекта для психики. То есть, знание о том, для чего можно использовать стол.

Основная идея формализации этих понятий в рамках теории категорий заключается в утверждении: “некий наблюдатель – субъект опыта, которому является Мир – присваивает значения вещам исключительно через призму взаимодействия с ними” или иначе “вещи, которыми мне является Мир, существуют для меня и лишь в отношении со мной” [1].

Формально “интеллектуальный объект” определяется как отображение, где:

• некоторый набор данных (A);

• наблюдатель, как отражение мира после взаимодействия с ним (Ω);

• отношение мира с наблюдателем, являющееся функцией внутреннего состояния/ожидания f – интеллектуальная функция $A \xrightarrow{f} \Omega$ [1].

Здесь A – набор данных, характеризующий любую единичную целостность; Ω – различительная способность: “отношение интеллектуального объекта со “мной” еще не означает какой-либо осознанности, представленности дан-

ного интеллектуального объекта в сознании – достаточно того, чтобы нечто было хоть как-то воспринято и распознано в степени, достаточной для того, чтобы это "нечто" в будущем было так или иначе учтено, принято в расчет... поскольку основной задачей мышления, как уже отмечалось выше, является предсказание, или производство конкурентного будущего, то каким нам в итоге представится воспринимаемый объект, будет зависеть от нашей настроенности, или, как сказали бы феноменологи, от нашей интенциональности" [26].

Более формально "К объекту Ω, моделирующему различительную способность, таким образом, будут предъявлены некоторые требования: это должно быть, во-первых, частично-упорядоченное множество, элементам которого соответствуют "более" или "менее" высокие значения. То есть на элементах данного множества должна иметься структура порядка. Иначе говоря, мы будем использовать Ω как некую экзистенциальную меру или, попросту говоря, линейку, которой мы будем измерять различия" [26].

Далее определяется функция ожидания $Exp_A : A \times A \to \Omega$ (от англ. expectation) и даются следующие пояснения этой функции: "на всех уровнях восприятия ... мы по сути имеем дело с ситуацией, с некоторым ожидаемым положением дел ... *Наша психика непреодолимо тяготеет к тому, чтобы сложить весь набор раздражителей в некую понятную, ясную и как бы непротиворечивую картину реальности* (курсив 1 – Е.Е.) ... Эти представления о реальности, в свою очередь, являются специфическим фильтром-интерпретатором – *всякие новые раздражители, оказываясь, образно говоря, в поле тяготения соответствующей системы представлений, неизбежно как бы изменяют свою траекторию – одни отталкиваются (игнорируются), другие, комплементарные, напротив, притягиваются, третьи – видоизменяются (интерпретируются) в угоду господствующим установкам* (курсив 2 – Е.Е.) ... В результате в отношении любого элемента x, входящего в состав интеллектуального объекта A, осмысленно говорить, насколько он, во-первых, отличен от самого себя в смысле того, что мы ожидаем увидеть на его месте, и, во-вторых, насколько он уместен в ситуации вообще, т.е. насколько он близок остальным элементам, различенным в ситуации".

Более формально [26], функция ожидания $Exp_A : A \times A \to \Omega$ сопоставляет каждой паре элементов $x, y \in A$ меру их согласованности (когерентности) на нашей экзистенциальной частично-упорядоченной шкале Ω. При этом мера согласованности объекта $x \in A$ с самим собой $Exp_A(x, x)$ может пониматься как мера близости x к своей сущности ... и обозначаться как $Ess_A(x)$ (от англ. essence). Если считать прототип объектов класса как "инвариант" объектов класса, то неподвижная точка оператора $\Upsilon^{\infty}(X(y))$, полученная на множестве свойств, заданных предикатами $X(y) = \{P_1 \,\&\, \ldots \,\&\, P_m\}$ для некоторого элемента $y \in A$, будет отличаться от признаков $X(y)$ самого элемента в точности как мера согласованности объекта с самим собой. Поэтому оператор $\Upsilon^{\infty}(X(y))$ дает определенную меру близости объекта к своей сущности (инварианту) "то есть мозг учится неким шаблонам восприятия — формирует в себе некие идеальные (инвариантные данной "сущности") модели, которые впоследствии помогают ему быстро объединять разрозненные данные, чтобы идентифицировать те или иные объекты, как бы вкладывая их в соответствующий инвариант" [25].

Функция ожидания позволяет полнее определить интеллектуальный объект [26]. "Под интеллектуальным объектом мы будем понимать ... объект $\mathcal{A} := (A, Exp_A)$, включающий в себя множество данных A и функцию ожидания $Exp_A : A \times A \to \Omega$, существенным образом зависящую от субъекта опыта Ω и его внутреннего состояния".

Строение интеллектуального объекта, описанное курсивом 1 выше, фактически означает, что интеллектуальный объект представляет собой контекст, представленный вероятностным формальным понятием, в котором оператор $\Upsilon^{\infty}(A)$, имеющим тот же смысл – минимизации противоречий в наборе раздражителей – по входному множеству раздражителей A генерирует максимально непротиворечивую картину реальности, дополняя ее всей, соответствующей ситуации информацией. При этом (см. курсив 2) новые раздражители либо меняют свою траекторию, либо отталкиваются, либо притягиваются. Все эти эффекты, которые учитываются в функции ожидания $Exp_A : A \times A \to \Omega$, моделируются взаимодействием вероятностных формальных понятий элементов множества A. Изменение траектории – это перевод признаков ближе к прототипу, отталкивание – это вытормаживание признаков, о котором говорилось выше и притяжение – это взаимная поддержка в неподвижной точке.

Интеллектуальная функция. Работа интеллектуальной функции состоит не только в том, чтобы воссоздать интеллектуальные объекты, связанные с элементами данных A и самим множеством A, но и соединить эти интеллектуальны объекты со всеми другими интеллектуальными объектами, имеющимися в психике и имеющими отношение к данной ситуации, например, к потребности, имеющейся в данной ситуации или некоторой задаче (цели) ""мир интеллектуальной функции" — это все возможные "интеллектуальные объекты", которые могут оказаться в пространстве психического

(по существу, речь идет о культурно-историческом содержании, как его понимал Л.С. Выготский). Причем они воспроизводятся конкретной психикой через отношение — интеллектуальную функцию — с другими, уже существующими в ней интеллектуальными объектами" [25].

Итогом работы интеллектуальной функции является создание "тяжелого интеллектуального объекта" путем "укрупнения имеющихся у нас знаний, которые мы полагаем относящимися к некоторой занимающей нас проблеме" [25]. Таким образом, интеллектуальный объект $\mathcal{A} := (A, Exp_A)$ "как бы возводится в степень тех знаний (интеллектуальных объектов), которыми мы обладаем, и обретает для нас соответствующее значение" [26].

Возведение некоторого интеллектуального объекта $\mathcal{A}$ "в степень" знаний оператором $\Upsilon^{\infty}(A)$ можно представить как отношение: "Если отношение мыслить, как определенного вида направленную связь, то кажется вполне естественным обозначать интеллектуальные объекты буквами $\mathcal{A}, \mathcal{B}, \mathcal{C}$ ..., а отношения между ними стрелками ... $r : \mathcal{A} \to \mathcal{B}$" [26]. Кроме того, это отношение "должно уважать те различия и отождествления, которые были положены функцией ожидания Exp_A" и удовлетворять следующим условиям:

$$\forall a, b \in A(Ess_B(r(a)) \le Ess_A(a)),$$
$$Exp_A(a, b) \le Exp_B(r(a), r(b)).$$

Все стрелки отношения r, показывающие путь обогащения некоторого интеллектуального объекта, образуют "конус". Пределом диаграммы обогащений является конус, порожденный "тяжелым" объектом, содержащимся во всех других конусах.

Оператор $\Upsilon^{\infty}(A)$ автоматически формирует контекст, порожденный A и вероятностными формальными понятиями ее элементов, поскольку по всем причинным связям, связывающим элементы A с другими знаниями, имеющимися в психике, другие ожидаемые элементы психики автоматически будут включены в контекст по этим причинным связям, если конечно они не будут сильно противоречить имеющейся информации, что проверяется этим оператором. Если, при этом, активируются некоторые высокоуровневые понятия, например, работа, учеба, то не будет извлекаться вся связанная с ними информация, кроме самой общей, а будет извлекаться информация, привязанная к контексту и свойствам ситуации, имеющимся в A. Пределом работы оператора $\Upsilon^{\infty}(A)$ является в этом случае контекст, соответствующий "тяжелому интеллектуальному объекту".

7. ТЕОРИЯ ФУНКЦИОНАЛЬНЫХ СИСТЕМ

Понятия "интеллектуальный объект" и "интеллектуальная функция" в рамках Brain Principles Programming должны описывать когнитивные функции и прежде всего мышление. Покажем на примере ведущей в России физиологической теории целенаправленной деятельности – Теории Функциональных Систем (ТФС), как такая теория может быть описана в терминах "интеллектуальных объектов" и "интеллектуальных функций".

Само по себе мышление целенаправленных действий не предполагает. Мы можем планировать достижений каких-то целей лежа на диване. Поэтому разобьем целенаправленное поведение по удовлетворению некоторой потребности на два этапа – этап планирования действий и принятия решения, который осуществляется еще до всяких действий, как формирование контекста, включающего "образ потребного будущего", и этап осуществления целенаправленного поведения в соответствии с принятым решением вместе с контролем достижения промежуточных и конечного результатов в соответствии с акцептором результатов действий [4].

Когда возникает некоторая потребность, а как правило, всегда доминирует некоторая потребность, если учесть, что спектр потребностей достаточно широк, то, во-первых, во множестве A должны быть элемент и соответствующий интеллектуальный объект, сформированный мотивационным возбуждением, которое активирует процесс поиска решения по удовлетворению потребности, а во-вторых, частично-упорядоченное множество Ω, моделирующее нашу различительную способность по ожиданию функцией $Exp_A : A \times A \to \Omega$ удовлетворения нашей потребности, будет оценивать элементы A и их взаимодействие с точки зрения удовлетворения потребности и соответствующим образом влиять на конструирование интеллектуального объекта "образа потребного будущего". Поэтому первый этап можно рассматривать как контекст функциональной системы по удовлетворению потребности, сформированный мотивационным возбуждением и "образом потребного будущего".

В теории функциональных систем достаточно подробно описан процесс формирования "образа потребного будущего". Полная и подробная формализация функциональных систем приведены в работах [27, 28].

Любая функциональная система имеет следующую архитектуру, первый этап которой мы проинтерпретируем как формирование контекста.

Первый этап включает:

Афферентный синтез. Включающий в себя синтез мотивационного возбуждения, памяти, обстановочной и пусковой афферентации:

• **Мотивационное возбуждение.** Постановка цели в целенаправленном поведении осуществляется возникшей потребностью, которая трансформируется в мотивационное возбуждение – возбуждение “центральных мозговых структур”, инициируемое возникшей потребностью. Мотивационное возбуждение формирует базовый интеллектуальный объект $\mathcal{M} := (M, Exp_M)$, задающий различительную способность Ω, которая будет определять, что нужно, а что не нужно для удовлетворения потребности. Элементами этого интеллектуального объекта являются возбуждения соответствующих мозговых структур.

• **Память.** Мотивационное возбуждение “извлекает из памяти” все последовательности действий, которые ранее приводили к достижению цели. Таким образом, интеллектуальный объект $\mathcal{M} := (M, Exp_M)$ “возводится в степень” – обогащается опытом тех случаев, которые ранее приводили к удовлетворению данной потребности. В результате получаем множество $\{\mathcal{C} := (C, Exp_C)\}$ обогащенных интеллектуальных объектов, соответствующих каждому случаю.

• **Обстановочная афферентация.** Мотивационное возбуждение с учетом текущей обстановки извлекает из памяти только тот опыт по достижению цели, который возможен в данной обстановке. Поэтому выбираются только интеллектуальные объекты тех способов $\mathcal{C} := (C, Exp_C)$ достижения цели, которые возможны в данной обстановке.

• **Пусковая афферентация.** По смыслу эта афферентация также является обстановочной афферентацией, только она связанна со временем и местом достижения результата.

Принятие решений. На стадии афферентного синтеза мотивационным возбуждением может быть извлечено из памяти несколько способов достижения цели и сформировано соответствующее множество $\{\mathcal{C} := (C, Exp_C)\}$ интеллектуальных объектов. В соответствии с формализацией [27–28] эти способы включают в себя правила вида:

$$P_1 \,\&\, \ldots \,\&\, P_n \,\&\, PG_1 \,\&\, \ldots \,\&\, PG_m \,\&\, A_1 \,\&\, \ldots \,\&\, A_k \Rightarrow PG_0,$$

где $P_1 \,\&\, \ldots \,\&\, P_n$ – условие обстановки, требуемое этим правилом для достижения цели, $PG_1 \,\&\, \ldots \,\&\, PG_m$ – подцели, которые нужно достичь для достижения конечной цели PG_0 и $A_1 \,\&\, \ldots \,\&\, A_k$ – действия, которые наряду с достижением подцелей требуется выполнить, для достижения конечной цели PG_0. Когда некоторый интеллектуальный объект $\mathcal{C} := (C, Exp_C)$ “возводится в степень” оператором $\Upsilon^{\infty}(A)$, используя имеющийся опыт с учетом обстановки, то он обогащается такими правилами, но без учета действий, предполагая, что они будут осуществлены в будущем. Поэтому знания, которыми обогащается интеллектуальный объект о способе достижения цели, имеют вид $P_1 \,\&\, \ldots \,\&\, P_n \,\&\, PG_1 \,\&\, \ldots \,\&\, PG_m \Rightarrow PG_0$, не содержащий действий. Это правило должно входить в вероятностное формальное понятие данного интеллектуального объекта.

На стадии принятия решений выбирается только один из способов достижения цели, формирующий план действий. Интеллектуальные объекты $\mathcal{C} := (C, Exp_C)$ “возведенные в степень” имеющегося опыта и включающие определенный способ достижения цели, дают “тяжелые” контексты, соответствующие разным “образам потребного будущего”. Среди этих “тяжелых” контекстов выбирается самый “тяжелый” $\mathbb{C} := (O, Exp_O)$ с наиболее желаемым “образом потребного будущего”. Он и есть результирующий контекст первого этапа работы функциональной системы.

Акцептор результатов действия. Выбранный план действий, соответствующий выбранному “образу потребного будущего”, включает в себя также последовательность и иерархию результатов, которые должны быть получены для достижения цели. Критерии достижения этих результатов, как совокупность определенных стимулов, которые должны быть получены при их достижении, формируют акцептор результатов действия. Это определенные “интеллектуальные объекты” со своей стимуляцией, функцией ожидания и различительной способностью, фиксирующие достижение этих результатов.

Второй этап выполнения плана действий вместе с контролем достижения промежуточных и конечного результатов акцептором результатов действий выполняется в точном соответствии с полученным контекстом. В случае какого-либо отклонения от плана включается ориентировочно исследовательская реакция, которая пересматривает план действий и возможно “образ потребного будущего” в результате чего сформированный контекст пересматривается.

Если в соответствии с выбранным планом действий цель достигается и потребность удовлетворяется, то этот план действий подкрепляется и заносится в память. В серии компьютерных экспериментов [27–33] была подтверждена работоспособность данной схемы.

8. ПЕРВЫЙ ПРИНЦИП ВРР – ПРИНЦИП ГЕНЕРАЦИИ СЛОЖНОСТИ

Принцип генерации сложности в [1, стр. 217] формулируется так: “Мозг работает с весьма ограниченным объемом информации от окружающей его реальности, поступающим на его сенсоры ... По мере использования этой, изначально скудной информации мозг, на всех уровнях своей организации многократно увеличивает ее объем, соотнося полученные вводные с уже существующими в нем данными ... Принцип генерации сложности позволяет мозгу, получив самый незначительный внешний сигнал, воспроизвести в сознании человека знание (интеллектуальный объект) несопоставимо большей мощности, обогатив модель этого объекта информацией, которая актуальна для мозга в рамках его задач (его целей)”.

Такую генерацию сложности выполняет введенная ранее в [1, стр. 214] интеллектуальная функция, которая “в рассматриваемом нами контексте выступает единственным инструментом мышления, используя которую мы создаем новые отношения между интеллектуальными объектами”.

Формализация “интеллектуального объекта” и “интеллектуальной функции” вероятностными формальными понятиями дает следующие модели генерации сложности:

1. Если рассматривать признаки цифр рис. 1, как признаки, воспринимаемые первичной зрительной корой, а множество A, как множество воспринимаемых цифр, то множество вероятностных формальных понятий, которые были обнаружены для этих цифр порождает множество интеллектуальных объектов $\mathscr{A}_0 \coloneqq (0, Exp_0)$, $\mathscr{A}_1 \coloneqq (1, Exp_1)$, ..., $\mathscr{A}_9 \coloneqq (9, Exp_9)$ – инвариантов этих цифр. Причинные связи между признаками цифр на множестве A будут найдены автоматически, поскольку мозг всегда и везде обнаруживает причинные связи. Найденные инварианты – это пример сгенерированной сложности на базе простейших свойств.

2. Формирование контекстов как вероятностных формальных понятий, что порождает, например, типологию пользователей социальной сети (см. пример выше). Контексты могут быть разные, например, вербальный контекст в виде законченного отрывка текста, смысл которого уточняет значения входящих в него слов или ситуативный контекст, включающий обстановку, время, место и т.д., помогающий более точно интерпретировать значения высказываний об обстановке.

3. Формирование ситуативного контекста некоторой функциональной системой с целью формирования плана действий по удовлетворению некоторой потребности.

В общем случае, когда решается некоторая задача или достигается определенная цель генерация сложности соответствующей интеллектуальной функции, будет состоять в генерации контекста по заданным начальным условиям A путем “возведения их в степень” тех знаний, которые имеют к ним прямое отношение. Формально это представляет собой генерацию некоторого вероятностного формального понятия по условиям A оператором $\Upsilon(A)^{\infty}$ с использованием всех относящихся к задаче или цели знаний, представленных совокупностью МСВПС правил.

9. ВТОРОЙ ПРИНЦИП ВРР – ПРИНЦИП ОТНОШЕНИЯ

В психологии этот принцип изначально получил название – принцип гештальта. Мозг, как мы знаем, реагирует не на конкретный стимул, а на то, каким становится этот стимул при соотнесении его с той информацией, которая в мозге уже содержится [1]. “Оценка возникающей в мозге информации ... осуществляется исключительно через акт соотнесения одной информации с другой, а сам мозг реагирует не на объект реальности как таковой, а на то, как он соотносится с другой информацией, находящейся в мозге” [1].

В качестве основного объяснительного принципа гештальтпсихология выдвигает принцип целостности. “Целостность восприятия – свойство восприятия, состоящее в том, что всякий объект, а тем более пространственная предметная ситуация воспринимаются как ***устойчивое системное целое***, даже если его некоторые части в данный момент нельзя наблюдать (например, тыльная часть вещи): актуально не воспринимаемые признаки все же оказываются интегрированными в целостный образ этого объекта” (Википедия). Целостность восприятия, которая формируется в процессе восприятия “естественного” понятия или прототипа класса, а также контекста некоторой задачи, формально представлена в вероятностном формальном понятии взаимным предсказанием свойств понятия или элементов контекста. Поэтому вероятностное формальное понятие образует то самое “устойчивое системное целое”, которое характеризует целостность.

Поэтому формально оператор $\Upsilon(A)^{\infty}$ и есть то самое “устойчивое системное целое”, в котором воспринимаются не отдельные элементы A, а их неразрывная взаимосвязь с остальными элементами неподвижной точки $\Upsilon(A)^{\infty}$.

10. ТРЕТИЙ ПРИНЦИП ВРР – ПРИНЦИП АППРОКСИМАЦИИ ДО СУЩНОСТИ

Принцип аппроксимации в [1] описывается следующим образом: “... в реальности не существует абсолютно идентичных объектов, поэтому мозг осуществляет аппроксимацию, то есть игнорирует отличия, если ему удается по специфическим признакам присвоить объекту ту или иную “сущность”. При этом, под “сущностью” понимается функционал объекта – то, какое значение он имеет для мозга (какую роль он выполняет) в рамках решаемых им задач (его целей). Наглядным примером в этом случае является использование какого-либо объекта в качестве другого, путем наделения первого функционалом второго под актуализированную потребность: когда человек устал и хочет отдохнуть – в лесу пень может служить стулом, так как на нем можно сидеть”.

Формирование “сущностей” происходит в контексте решаемых задач или функциональных систем. Всякий контекст уточняет и взаимно соотносит элементы контекста. Это приводит к формированию “сущностей”, связанных с контекстом. Например, нож в разных контекстах: приготовления пищи, боевой ситуации, офисной работы и походных условиях должен обладать разными свойствами, вытекающими из контекста: для кухонного ножа важна взаимосвязь ширины, веса и острия лезвия, для канцелярского ножа – малость веса, длина и безопасность, для перочинного ножа – относительная малость размеров. Поэтому формируются “сущности” “кухонный нож”, “боевой нож”, “канцелярский нож”, “перочинный нож”, которые автоматически в соответствующих ситуациях порождают различные вероятностные формальные понятия, поскольку свойства и закономерности их взаимосвязи различны.

Контекст решаемой задачи, цели или потребности будет автоматически заставлять выбирать наиболее подходящие для этого объекты с соответствующим “функционалом”. Этот функционал, имеющий определенное значение для мозга в рамках решаемых им задач и целей, определенным образом отразится на совокупности свойств объекта, которые, взаимно предполагая друг друга, автоматически сформируют соответствующее вероятностное формальное понятие, соответствующее его функциональной “сущности”.

Поэтому “сущность” – это вероятностное формальное понятие $\Upsilon(A)^{\infty}$, порожденное такими элементами A – свойствами используемых объектов, которые будут выбираться в соответствии с контекстом решаемой задачи или достигаемой цели.

11. ЧЕТВЕРТЫЙ ПРИНЦИП ВРР – ПРИНЦИП ЛОКАЛЬНОСТИ-РАСПРЕДЕЛЕННОСТИ (ПРИНЦИП СИМУЛЬТАННОСТИ)

Принцип локальности-распределенности [1]: “Вся информация, поступающая в мозг, может в нем многократно дублироваться, и ее копии обрабатываются параллельно разными структурами самостоятельно, и лишь затем эта информация интегрируется в целостный образ. Иными словами, мозг обрабатывает одну и ту же информацию разными способами (в разных отделах), чтобы получить несколько результатов и объединить их в рамках одного, целостного интеллектуального объекта, в соответствии с определенной им сущностью”.

Мозг обрабатывает информацию о некотором объекте параллельно сразу в нескольких модальностях – зрительной, слуховой, тактильной и т.д. В каждой из этих модальностей образуется иерархия простейших “естественных” классов и понятий, например, в зрительной коре на основании воспринятых палочек могут формироваться образы цифр, как в приведенном выше примере, а также “вторичные” признаки – линии, углы, окружности и т.д., в слуховой коре – фонемы, слова, текст и т.д. Согласование модальностей образа осуществляется уже на верхнем уровне через восприятие целостности объекта, которая интегрирует и связывает восприятие частей в “устойчивое системное целое”, что и осуществляют вероятностные формальные понятия целостных объектов.

Работа интеллектуальной функции по принципу локальности-распределенности (симультанности) состоит не только в том, чтобы интегрировать модальности некоторого образа и воссоздавать интеллектуальные объекты, связанные с элементами воспринимаемых данных A, но и соединять эти интеллектуальные объекты со всеми другими интеллектуальными объектами, имеющимися в психике и имеющими отношение к данной ситуации, например, к некоторой потребности или задаче (цели). Таким образом, интеллектуальный объект $\mathcal{A} := (A, Exp_A)$ “как бы возводится в степень” тех знаний (интеллектуальных объектов), которыми мы обладаем и в результате работы интеллектуальной функции параллельно создаются “конусы” – множества $\{\mathcal{C} := (C, Exp_C)\}$ интеллектуальных объектов, как и случае функциональных систем, обогащающих исходный интеллектуальный объект $\mathcal{A} := (A, Exp_A)$ до некоторых целостных контекстов, определяющих возможные смыслы воспринимаемой ситуации A.

Поэтому формально этот принцип также представляется оператором $\Upsilon(A)^{\infty}$, генерирующим вероятностные формальные понятия целостных контекстов воспринимаемой ситуации A.

12. ПЯТЫЙ ПРИНЦИП ВРР – ПРИНЦИП ТЯЖЕСТИ

Принцип тяжести [1]: “Количество нейронных связей, включенных в создание модели объекта, количество отношений между элементами континуума интеллектуальных объектов, объем привносимой в объект информации (атрибуты сущности), количество способов расчета информации об объекте и объединение разноканальной (модальности) информации о нем в единое целое, соотнесенные с актуальностью задачи (цели) системы, определяют “тяжесть” интеллектуального объекта. “Тяжесть” интеллектуального объекта предопределяет решение системы. Так, например, если человек голоден – он будет искать пищу, которая утолит голод, однако если ему начнет угрожать непосредственная опасность (например, от хищника), то начнет главенствовать оборонительная стратегия, и он перестанет искать еду и начнет спасаться, так как без еды он проживет еще какое-то время, а если его настигнет хищник – он умрет сразу. То есть, приоритет отдается наиболее актуальной и выраженной в каждой конкретной ситуации стратегии”.

Еще в 1911 г. А.А. Ухтомским был выдвинут принцип доминанты [34]. Он сохранился и в теории функциональных систем [3, 4], как принцип доминирующей функциональной системы, которая и создает наиболее “тяжелый” контекст.

В общем случае, когда речь идет о решении некоторой задачи или достижении определенной цели, возможные решения по принципу локальности-распределенности (симультанности) получаются разными путями обогащения исходного интеллектуального объекта “постановка задачи” (цели) и образуют соответствующие “конусы” и порождаемые ими контексты. Выбор из них самого “тяжелого” определяется выбором наиболее желаемого “тяжелого” решения.

Поэтому формально принцип тяжести состоит в выборе наиболее желаемого “тяжелого” интеллектуального объекта, порожденного одним из контекстов, которые генерируются оператором $\Upsilon^{\infty}(M \cup C)$ в зависимости от исходной постановки задачи/цели $\mathcal{M} := (M, Exp_M)$ и имеющегося опыта C решения подобной задачи/цели.

13. СИСТЕМНАЯ ВЗАИМОСВЯЗЬ ПРИНЦИПОВ

Когда мы берем в руки яблоко (см. рис. 3), то получаем первичную информацию A о нем – оно твердое, имеет средний вес, форма круглая, поверхность гладкая, размер средний. Сначала начинает работу первый принцип – генерации сложности: “Мозг работает с весьма ограниченным объемом информации от окружающей его реальности, поступающим на его сенсоры ... По мере использования этой, изначально скудной информации мозг, на всех уровнях своей организации многократно увеличивает ее объем, соотнося полученные вводные с уже существующими в нем данными ...” [1]. Поэтому мы сразу понимаем, что это не апельсин, поскольку апельсин не имеет гладкой поверхности и это не мячик, т.к. мячики не твердые и это не бильярдный шар, поскольку они тяжелые и это не теннисный мячик, поскольку его поверхность не шершавая и т.д.

Далее работает принцип отношений: “Оценка возникающей в мозге информации ... осуществляется исключительно через акт соотнесения одной информации с другой, а сам мозг реагирует не на объект реальности как таковой, а на то, как он соотносится с другой информацией, находящейся в мозге” [1], “всякий объект... воспринимаются как устойчивое системное целое, даже если его некоторые части в данный момент нельзя наблюдать” (Википедия). Воспринятая информация должна соотносится сама с собой и образовывать некоторое системное целое – поэтому из имеющихся в памяти образов извлекается именно образ яблока и этот образ также дополнительными (перцептивными) действиями проверяется на целостность, например, на наличие хвостика и диаметрально расположенных ямок.

Далее по принципу генерации сложности, который работает всегда, воспринятая информация еще более обогащается имеющимися знаниями и возникает понимание “сущности” воспринятого яблока по третьему принципу аппроксимации до сущности: под ““сущностью” понимается функционал объекта – то, какое значение он имеет для мозга (какую роль он выполняет) в рамках решаемых им задач” [1]. Это яблоко “съедобное”, если мы собираемся его есть, это яблоко “красивое”, если мы им просто любуемся или собираем натюрморт, это яблоко “сортовое”, если нам нужны его косточки для дальнейшего разведения и т.д.

Далее, все также по принципу генерации сложности воспринятая информация еще более обогащается “возводится в степень” интеллектуальной функцией параллельно по разным каналам обработки информации вплоть до целостного ее восприятия в соответствии с четвертым принципом локальности-распределенности (симультанности): “Вся информация, поступающая в мозг, может в нем многократно дублироваться, и ее копии обрабатываются параллельно разными структурами самостоятельно, и лишь затем эта информация интегрируется в целостный образ” [1]. Этот целостный образ формирует контекст восприятия – в каком контексте мы воспринимаем яблоко – в контексте поесть, в контексте полюбоваться, в контексте семеноводства и т.д.

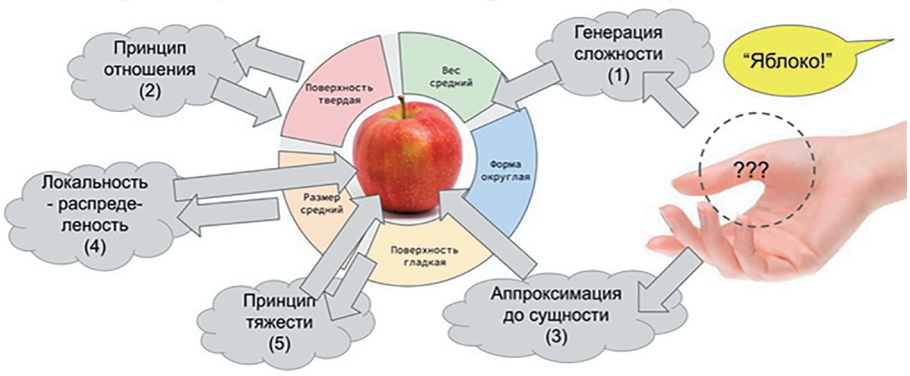

Рис. 3. Системная взаимосвязь принципов.

Пятый принцип – принцип тяжести определяет наш выбор наиболее тяжелого контекста из воспринятых, который в данный момент наиболее желаем для нас или больше всего нас интересует и который определит наши дальнейшие размышления или действия: "объединение разноканальной информации в единое целое, соотнесенные с актуальностью задачи (цели) системы, определяют "тяжесть" интеллектуального объекта. "Тяжесть" интеллектуального объекта предопределяет решение системы" [1].

С формальной точки зрения непрерывная работа интеллектуальной функции в соответствии с принципом генерации сложности по обогащению воспринятой информации описывается оператором $\Upsilon(A)$, интеллектуальные объекты и контексты, получаемые по второму, третьему и четвертому принципам, описываются вероятностными формальными понятиями, т.е. оператором $\Upsilon^{\infty}(A)$. Устойчивое системное целое, характеризующее целостность этих интеллектуальных объектов, следует из строения вероятностных формальных понятий, выявляющих системный закон (в виде согласованной системы причинных связей) строения интеллектуальных объектов. Следует отметить, что интеллектуальные объекты, получаемые вероятностными формальными понятиями, описываются не только как набор (синдром) признаков, но и как устойчивое системное целое "инвариант", характеризующийся системной взаимосвязью причинных связей, взаимно предсказывающих признаки интеллектуального объекта.

В основе формализации BPP в рамках теории категорий также лежит только два понятия – "интеллектуальный объект" и "интеллектуальная функция", функционирование которых разворачивается в принципах. В нашей формализации интеллектуальному объекту соответствует оператор $\Upsilon^{\infty}(A)$, а интеллектуальной функции – оператор $\Upsilon(A)$.

В работе [34] приведена англоязычная версия описания базовых моделей и принципов программирования мозга.

14. ПРАКТИЧЕСКИЕ СООБРАЖЕНИЯ И ПРИМЕНИМОСТЬ В ПРИКЛАДНЫХ ЗАДАЧАХ

Алгоритмы построения вероятностных формальных понятий и соответственно прототипов классов, "естественных" понятий и контекстов практически подтверждены [12, 16–20, 24]. Модель функциональных систем также разработана и показала свою эффективность [26–32]. Интеграция этих алгоритмов и моделей может быть осуществлена путем использования в контекстах функциональных систем правил без действий (см. раздел 6) в предположении, что необходимые действия будут выполнены и проконтролированы в соответствии с включенными в контекст функциональными системами. Интегрированный алгоритм достаточно точно моделирует основные когнитивные функции человека и животных, упомянутые в первых частях статьи, поэтому область применимости может быть широка. Для этого необходимо провести масштабирование данного подхода.

15. ВЫВОДЫ

Данный подход может быть обобщен до *задачного подхода к общему искусственному интеллекту*, как это и планировалось в [1] путем обобщения функциональных систем до систем решения задач [35–39]. Этот подход вполне справляется с задачей AGI, сформулированной в [1] как: "способность достигать целей в широком диапазоне сред с учетом ограничений".

Поэтому Brain Principles Programming, сформулированные в [1] как принципы программирования мозга, опираясь на исследования в когнитивных науках, могут быть реализованы как за-

дачный подход к AGI, который одновременно способен решать достаточно широкий класс задач, а с другой стороны, достаточно точно соответствует моделям когнитивных процессов.

СПИСОК ЛИТЕРАТУРЫ

1. Сильный искусственный интеллект. На подступах к сверхразуму // Александр Ведяхин [и др.]. М.: Интеллектуальная Литература, 2021. 232 с.
2. *Анохин П.К.* Опережающее отражение действительности // Философские аспекты теории функциональных систем. М.: Наука, 1978. С. 7–27.
3. *Anokhin P.K.* Biology and neurophysiology of the conditioned reflex and its role in adaptive behavior. Oxford., Pergamon press, 1974. 574 p.
4. *Судаков К.В.* Общая теория функциональных систем М., Медицина, 1984. С. 222.
5. *Masafumi Oizumi, Larissa Albantakis, Giulio Tononi*. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0 // PLOS Computational Biology, May 2014, V.10. Issue 5.
6. *Rosch E.H.* Natural categories // Cognitive Psychology 4, 1973 P. 328–350.
7. *Rosch, Eleanor and Lloyd, Barbara B.* (eds), Cognition and categorization. Hillsdale, NJ: Lawrence Erlbaum, 1978, P. 27–48.
8. *Rosch E.* Principles of Categorization // Rosch, E. & Lloyd, B.B. (eds), Cognition and Categorization, Lawrence Erlbaum Associates, Publishers, (Hillsdale), 1978. P. 27–48.
9. *Bob Rehder*. Categorization as causal reasoning // Cognitive Science 27 (2003) 709–748.
10. *Rehder B.* (2003). A causal-model theory of conceptual representation and categorization. J. of Exper. Psych.: Learning, Memory, and Cognition, 29, 1141–1159.
11. *Bob Rehder, Jay B. Martin*. Towards A Generative Model of Causal Cycles // 33rd Annual Meeting of the Cognitive Science Society 2011, (CogSci 2011), Boston, Massachusetts, USA, 20–23 July 2011, V. 1. P. 2944–2949.
12. *Витяев Е.Е., Мартынович В.В.* Формализация “естественной” классификации и систематики через неподвижные точки предсказаний. Сибирские электронные математические известия. Том 12, ИМ СО РАН, 2015, С. 1006–1031. Mill, J.S. System of Logic, Ratiocinative and Inductive. L., 1983.
13. *Mill J.S.* System of Logic, Ratiocinative and Inductive. L., 1983.
14. *Рутковский Л.* Элементарный учебник логики. Санкт-Петербург, 1884.
15. *Смирнов Е.С.* Конструкция вида таксономической точки зрения // Зоол. Журн. 1938. Т. 17. № 3. С. 387–418.
16. *Витяев Е.Е., Демин А.В., Пономарев Д.К.* Вероятностное обобщение формальных понятий // Программирование. 2012. Т. 38. № 5. С. 219–230.
17. *Vityaev E.E., Demin A.V., Ponomaryov D.K.* Probabilistic Generalization of Formal Concepts // Programming and Computer Software. 2012, V. 38. № 5. P. 219–230.
18. *Vityaev E.E., Martinovich V.V.* Probabilistic Formal Concepts with Negation // A. Voronkov, I. Virbitskaite (Eds.): PCI 2014, LNCS 8974, P. 385–399.
19. *Evgenii Vityaev.* The logic of prediction // Mathematical Logic in Asia. Proceedings of the 9th Asian Logic Conference (August 16–19, 2005, Novosibirsk, Russia), edited by S.S. Goncharov, R. Downey, H. Ono, World Scientific, Singapore, 2006, P. 263–276.
20. *Vityaev E., Odintsov S.* How to predict consistently? Trends in Mathematics and Computational Intelligence In: Studies in Computational Intelligence, 796, Mar?a Eugenia Cornejo (ed), 2019, 35–41.
21. *Vityaev E.E.* A formal model of neuron that provides consistent predictions // Biologically Inspired Cognitive Archi-tectures 2012. Proceedings of the Third Annual Meeting of the BICA Society (A. Chella, R.Pirrone, R. Sorbello, K.R. Johannsdottir, Eds). In Advances in Intelligent Systems and Computing, v.196, Springer: Heidelberg, New York, Dordrecht, London. 2013. P. 339–344.
22. *Cartwright N.* Causal Laws and Effective Strategies. Noûs. 1979. 13. P. 419–437.
23. *Ganter B.* Formal Concept Analysis: Methods, and Applications in Computer Science. TU Dresden, Germany, 2003.
24. *Витяев Е.Е., Неупокоев Н.В.* Формальная модель восприятия и образа как неподвижной точки предвосхищений // Подходы к моделированию мышления. УРСС Эдиториал, Москва, 2014г., стр. 155–172.
25. *Андрей Курпатов*. Мышление. Системное исследование. ООО “Дом Печати Издательства Книготорговли “Капитал””, 2019.
26. *Егорычев И.Э.* Категорный анализ текста “методология мышления” А.В. Курпатова в контексте перспективных разработок AGI. Научное мнение № 7–8 (2020).
27. *Evgenii E.* Vityaev Purposefulness as a Principle of Brain Activity // Anticipation: Learning from the Past, (ed.) M. Nadin. Cognitive Systems Monographs, V. 25, Chapter No.: 13. Springer, 2015. P. 231–254.
28. *Витяев Е.Е.* Логика работы мозга. Подходы к моделированию мышления. (сборник под ред. д.ф.-м.н. В.Г. Редько). УРСС Эдиториал, Москва, 2014, С. 120–153.
29. *Демин А.В., Витяев Е.Е.* Логическая модель адаптивной системы управления. Нейроинформатика. 2008. Т. 3. № 1. С. 79–107.
30. *Demin A.V., Vityaev E.E.* Adaptive control of multiped robot // Procedia Computer Science 145C (2018). P. 629–634.
31. *Alexander V. Demin and Evgenii E. Vityaev.* Adaptive Control of Modular Robots // A.V. Samsonovich and V.V. Klimov (eds.), Biologically Inspired Cognitive Architectures (BICA) for Young Scientists, Advances in Intelligent Systems and Computing 636, Springer. 2018. P. 204–212.
32. *Anton Kolonina, Evgenii Vityaev, Yuriy Orlov.* Cognitive Architecture of Collective Intelligence Based on Social Evidence // 7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA

2016, 16-19 July, 2016 in New York City, NY, USA. Procedia Computer Science. 2016. V. 88. P. 475–481.

33. *Demin A.V., Vityaev E.E.* Learning in a virtual model of the C. elegans nematode for locomotion and chemotaxis. Biologically Inspired Cognitive Architectures. 2014. V. 7. P. 9–14.
34. *Ухтомский А.А.* Доминанта. Статьи разных лет. 1887-1939. СПб.: Питер, 2002. 448 с.
35. *Sergei S. Goncharov, Dmitrii I. Sviridenko, Evgenii E. Vityaev.* Task Approach to Artificial Intelligence // Proceedings of the Workshop on Applied Mathematics and Fundamental Computer Science 2020 (AMFCS 2020), Omsk, Russia, April 23–30, 2020. CEUR Workshop Proceeding. Vol. 2642, pp. 1–6.
36. *Витяев Е.Е., Гончаров С.С., Свириденко Д.И.* О задачном подходе в искусственном интеллекте // Сибирский философский журнал. 2019. Т. 17. № 4. С. 5–25.
37. *Витяев Е.Е., Гончаров С.С., Свириденко Д.И.* О задачном подходе в искусственном интеллекте и когнитивных науках // Сибирский философский журнал. 2020. Т. 18. № 2. С. 5–29.
38. *Свириденко Д.И., Витяев Е.Е.* Задачный подход к искусственному интеллекту и его теоретическая и технологическая база // 18 Национальная конференция по искусственному интеллекту с международным участием (КИИ-2020). Труды конференции / Под ред. В.В. Борисова, О.П. Кузнецова. М.: МФТИ, 2020. (326с.). С. 36–44.

EDN: WUJHXZ
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 128–133

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ЦИФРОВОЙ КОВЧЕГ ЗНАНИЙ

© 2022 г. В. В. Горячко[1], А. С. Бубнов[1], Е. В. Раевский[1], Академик РАН А. Л. Семенов[1,2,*]

Поступило 30.10.2022 г.
После доработки 06.11.2022 г.
Принято к публикации 08.11.2022 г.

В работе рассматривается общая проблема создания и использования энциклопедического знания в цифровой цивилизации. Создание и использование цифровой энциклопедии являются примером интеллектуальной деятельности коллективной расширенной личности. Представлены основные положения, относящиеся к созданию и функционированию "Цифрового ковчега знаний" МГУ на примере математических наук. Рассмотрены возникающие проблемы и пути их решения.

Ключевые слова: энциклопедия, БРЭ, расширенная личность, расширенное сознание, википедия, коллективное знание, открытые общедоступные информационные ресурсы, профессиональное сообщество, МГУ им. М.В. Ломоносова, математика

DOI: 10.31857/S2686954322070098

1. ВВЕДЕНИЕ

Как уже отмечалось [1], развитие человечества можно охарактеризовать информационными революциями. Эти революции определялись возникновением:

- Сознания,
- Речи,
- Письма,
- Автоматизации интеллектуальной деятельности – искусственного интеллекта.

Каждая из революций сопровождалась изменением того, как человек мыслит, общается, действует. Можно сказать, что каждая из них приводила к расширению сознания, личности человека [2, 3]. Одновременно каждая из этих революций приводила к сокращению, иногда – отмиранию тех или иных форм деятельности и связанных с ними способностей человека, как это отмечает Л.С. Выготский [4–6]. По отношению к возникновению письменности об этом писал Платон в диалоге "Федр", ссылаясь на (бесписьменного) Сократа [7].

Концепция расширенной личности является продуктивной метафорой для:

- описания современного человека как участника производственных, социальных и экономических отношений,
- проектирования целей и содержания образования и оценивания их достижения.

Последняя из перечисленных революций началась около 100 лет назад внутри математики – когда математики, начав с построения формальных математических систем описания математических рассуждений, математической коммуникации построили универсальную формализацию интеллектуальной деятельности вообще [8]. Эта формализация, благодаря успехам естественных наук и технологии, материализовалась в первой половине XX века в создании компьютера. Сегодня человек передает компьютеру все большие фрагменты интеллектуальной деятельности, начиная с ее рациональных элементов, таких, как решение вычислительных задач, включая численное моделирование и компьютерную алгебру. Взрывной рост цифровых технологий привел в XXI веке к автоматизации интуитивной деятельности человека – машинному обучению, с которым часто отождествляется и весь искусственный интеллект. Мы не будем вдаваться здесь в терминологическую дискуссию.

Внутри самой революции искусственного интеллекта также произошли революционные события. Важнейшее из них – это накопление информации в общем пространстве, ее повсеместная доступность и организация простого доступа к ней. Эта "суб-революция" знания (подобная

[1] *Московский государственный университет имени М.В. Ломоносова, Москва, Россия*
[2] *Московский физико-технический институт (национальный исследовательский университет), г. Долгопрудный, Московская область, Россия*
*****E-mail: alsemno@ya.ru*

суб-революции Гутенберга внутри революции письменности) основывалась на:

- интернете,
- мобильных устройствах (смартфонах),
- поисковых системах.

Возможно, следующая революция, которая нам предстоит, будет базироваться на прямом интерфейсе "мозг – компьютер". Сейчас этот интерфейс частично реализован в виде помощника для восприятия устной речи при некоторых видах глухоты (кохлеарный имплант), протезировании зрения и примитивной (пока) передачи нервных импульсов в электронные устройства, в частности, для людей с ограниченной подвижностью.

К важнейшим, действительно революционным событиям XXI века относится и запуск принципиально новых форм коллективной интеллектуальной деятельности – формирование коллективной расширенной личности. Наиболее ясной и убедительной формой такой деятельности стала Википедия.

2. СОЗДАНИЕ ВИКИПЕДИИ КОЛЛЕКТИВНОЙ РАСШИРЕННОЙ ЛИЧНОСТЬЮ

Собрания знаний о мире играли принципиальную роль в развитии науки, культуры и образования. Достаточно упомянуть "Orbis pictus" Коменского [9] – начальный текстово-картиночный полилингвальный инструмент ориентации ребенка в мире, и "Энциклопедию" Дидро и Д'Аламбера – ознаменовавшую Век Просвещения [10].

Википедия – симбиоз Человечества и Технологии. Это – дело расширенной глобальной человеческой личности, которое стало возможной благодаря цифровым технологиям. Благодаря ей человеческое знание стало доступно большинству населения Земли: теперь этим знанием фактически пользуется намного больше людей, чем всеми справочными изданиям бумажного века. При этом количество использований Википедии в единицу времени выросло на порядки по сравнению с бумажными энциклопедиями.

Основой для социальной технологии Википедии стали ее правила – Столпы, разумные ограничения: энциклопедичность, нейтральность, общедоступность, взаимное уважение участников, отсутствие жестких правил.

Таким образом, Википедия является важнейшим массовым ответом ИИ на невозможность для человека и человечества справиться с экспоненциальным информационным взрывом.

Серьезные проблемы и недостатки Википедии, критически существенные для отдельных вопросов и сообществ (некоторые из них обсуждаются ниже), незначительно влияют на ее важность в целом.

3. ИСПОЛЬЗОВАНИЕ ВИКИПЕДИИ РАСШИРЕННОЙ ЛИЧНОСТЬЮ

Важнейшим эффектом проекта Википедии стало активное использование Википедии населением Земли. Без этого проект, конечно, потерял бы смысл. Сегодня, когда ученый, литератор, деятель культуры читает лекцию в широкой аудитории и ссылается на какую-то неожиданную идею, не слишком известную историческую личность, использует специфическое понятие, концепцию, в аудитории бывает заметно движение: несколько человек достают свои мобильные телефоны. Это значит, что аудитория состоит из расширенных личностей, частью индивидуального знания которых являются энциклопедические статьи Википедии. Такое было невозможно в доцифровую эпоху: так ли часто любознательный молодой человек середины XX века обращался к "Большой советской энциклопедии"? Сегодня же даже средний школьник, получив задание, выходящее за рамки учебника, немедленно обращается к Википедии. Не менее важным является то, что учащиеся уже в начальной школе могут создавать и использовать собственные микро-википедии: коллективное знание о ближайшем окружении (природном, техногенном, социальном) учеников одного класса (а потом – шире, школы); этот подход мы реализуем в проекте курса "Будущий мир" для начальной школы [11].

Таким образом, упомянутая "суб-революция" знания сегодня охватила существенную долю населения Земли – куда большую, чем охваченную письменной грамотностью в XIX веке (вы помните, что тогда в России она не превосходила 15%). По существу, как отмечают многие авторы, изменилось представление о том, что значит, что человек что-то знает. Это представление мигрирует в направлении расширенной личности. Пример отражения такой миграции в научном дискурсе сегодня можно видеть уже в бакалаврских работах [12].

Как это уже бывало неоднократно в последнее столетие, технологическое достижение, возможность которого доказана на Западе, было освоено и интерпретировано Китаем. Созданная там на авторитарной основе альтернатива Википедии – Байдупедия (см. [13]) – быстро обогнала всю мировую Википедию по объемам.

Еще одним явлением, существенным для мирового научного сообщества, являются общедоступные (Open Access) публикации и культура электронных препринтов. Показателен, хотя и уникален, пример Г. Перельмана, получившего высшие математические награды за такие архив-

ные, в частности, не рецензируемые (и не проверяемые на плагиат)) публикации.

Наконец, есть огромное море размещенной в интернете профессиональной литературы, часто сомнительной с точки зрения качества оцифровки и правообладания.

4. НЕДОСТАТКИ И ПРОБЛЕМЫ ВИКИПЕДИИ

Вот наугад взятый из интернета пример критики Википедии:

“Однако, главная беда Википедии заключается в том, что модераторы и администраторы этого ресурса, являясь безусловными эрудитами и специалистами по вики-разметке, зачастую не являются профильными специалистами в тех областях, о которых пишут. То есть, если сильно упростить, то они лишь компилируют разрозненную по источникам информацию в единый текст. Правила Википедии не запрещают слесарю из штата Техас писать статьи о достижениях онкологии в России, например” [14]. Автор в своей публикации указывает не только на недостаток “демократичности” Википедии, но и на возможность эксплуатации этого недостатка для неблаговидных, в том числе, личных, или политических целей, противоречащих исходной идее и идеологии Википедии.

Замечательно при этом, что во многом именно исходная демократическая идея обеспечила “доказательство возможности – proof of concept” важнейшего технолого-социального проекта.

Ключевые проблемы Википедии сегодня:

• различие в уровне мотивации и квалификации отдельных авторов (те, кто может написать профессионально, не пишут),

• качество статей,

• полнота ссылок и материала, куда ведут ссылки.

Несмотря на эти проблемы, ясно, что сегодня Википедия стала явлением, значительно дополняющим энциклопедические ресурсы, созданные по традиционным издательским и социальным технологиям и переведенные “в цифру”.

5. КОВЧЕГ ЗНАНИЙ. ОСОБЕННОСТИ ПРОЕКТА МГУ

В доцифровую эпоху нашей стане удалось достичь значительных энциклопедических достижений, таких, как энциклопедия Брокгауза и Эфрона, Большая советская энциклопедия (БСЭ) в трех изданиях. Математическая энциклопедия [15], создание которой совпало с завершением советского периода развития отечественной математики, стала уникальным мировым явлением. Доказательством этого является создание на ее основе The Encyclopedia of Mathematics Европейского математического общества [16].

Большая российская энциклопедия (БРЭ) продолжила традицию БСЭ в бумажном формате, а последнее десятилетие и – в цифровом [17]. Сотрудничество с БРЭ обеспечивает для профессионального сообщества возможность выхода на все общество и поддержку государства. Далее мы описываем идущие сейчас процессы взаимодействия профессионального научного сообщества с БРЭ на примере, прежде всего, математики и информационных технологий.

Заметим, что параллельно с БРЭ развивается российская википедия, сегодня входящая в первую пятерку википедий мира. Также расширяется система открытого доступа к публикациям.

Начав свое сотрудничество с БРЭ, мы столкнулись с (общемировой) проблемой низкой мотивации профессионалов к написанию энциклопедических статей. Прямое использование механизмов википедии приводило бы к снижению качества результата.

В целях решения проблемы были выработаны следующие подходы:

• возможность отделения функции создания статей от других функций (заказ, рецензирование, передача в БРЭ);

• возложение задач обеспечения качества статей на профессиональное сообщество, так, как это делается в журналах, советах по защите диссертаций и т.п.;

• создание **Ковчега**, где готовятся и предварительно размещаются рабочие материалы, которые после их одобрения передаются в БРЭ;

• использование как фундамента уже существующих источников, прежде всего “Математической энциклопедии”;

• технологическая поддержка комфортной платформы коллективной работы с документами – вики-редактирования, размещенной на серверах в РФ;

• учет публикуемых в БРЭ статей как научных публикаций в ведущих журналах.

Структура процессов, из которых формируется общий процесс создания энциклопедии и участники этих процессов, складывается следующим образом:

• **профессионал**, которому профессиональное сообщество уже вручило ответственность за какую-то область знания, и он эту ответственность принял: заведующий кафедрой (передача знания новым поколениям профессионалов), заведующий отделом исследовательского института (новые исследования и результаты), председатель ученого совета (оценка вклада других), главный редактор журнала (представление новых резуль-

татов сообществу), член академии, профессор РАН (избран профессиональным сообществом):

• Получает из научно-авторитетного в своей области источника (например, отделения РАН) приглашение принять участие в формировании содержания Энциклопедии в его области; такое участие не предполагает личного написания статей: речь идет о формировании перечня статей, выборе авторов, реакции на написанные статьи.

• Непринятие такого приглашения, или его полное игнорирование ставит этого профессионала в данном конкретном отношении вне своей референтной группы. Аналогия: не такая уж большая доля специалистов отказывается от вхождения в редколлегию научного журнала.

• Принятие предложения означает, что профессионал вместе со своими коллегами разделяет ответственность за полноту и качество представления их области в Энциклопедии.

• Принявшие приглашение профессионалы каждой области образуют **редакционную группу** (РГ) этой области (например, теории чисел, или генетики), куда они могут кооптировать коллег, как правило, имеющих академический статус (научную степень).

• Каждый из членов РГ, если не получает возражений от членов РГ, может написать статью сам или предложить любому **автору** написать статью на тему из области, относящейся к РГ, или разместить на странице группы открытое предложение написать статью на предлагаемую им тему.

• Каждый желающий может написать статью на тему, предлагаемую какой-то РГ, или предложить свою тему и направить статью для размещения в Ковчеге, при этом он может разместить ее, например, и в русской Википедии.

• Статья, направленная для размещения, попадает в Ковчег после того, как она получила одобрение хотя бы одного члена РГ.

• Создаваемые в Ковчеге статьи могут использовать в качестве основы (со ссылкой) статьи из Математической энциклопедии, БСЭ, Математического энциклопедического словаря, открытых свободно распространяемых источников. РГ может предложить БРЭ разместить какую-то существующую или переработанную статью.

• **Научно-редакционная коллегия БРЭ** по данной области знания, например, математике, направляет в РГ для размещения в Ковчеге и переработки существующие статьи из энциклопедических источников. При отсутствии предложений от всех РГ БРЭ размещает статью у себя, в Ковчеге появляется соответствующая ссылка.

• Любой член редакционной группы может предложить имеющуюся в Ковчеге статью для передачи в БРЭ. Статья, не получившая возражений от членов РГ, передается в БРЭ, с автором заключается договор, авторство может быть указано в БРЭ, автор получает от БРЭ разрешение опубликовать статью со ссылкой на первоначальный источник частично или полностью, в оригинале или переводе на иностранные языки, после опубликования статьи в БРЭ, со ссылкой на БРЭ.

Если члены РГ предлагают написать статью авторитетному ученому, специалисту в теме статьи, то в качестве возможной формы организации работы по написанию статьи таким ученым предлагается следующая технологическая схема:

• РГ может предложить рекомендации по написанию, включая, например, объем того варианта статьи, который будет размещен в БРЭ, а также список вопросов для освещения в статье, информацию о ближайших в возможной сетевой структуре статьях: более общие статьи, откуда идет ссылка на данную, более специальные, частные статьи, соседние статьи, с которыми возможно пересечение. Автор в ходе работы над статьей может существенно отклониться от этих рекомендаций, предложив альтернативы.

• Автор привлекает для написания статьи одного или нескольких своих сотрудников и учеников и делает устный доклад-интервью, используя предложенный ему РГ список вопросов, слушатели доклада также задают нужные им вопросы. Доклад записывается (при необходимости с фиксацией письменных заметок, формул и т.п.)

• Запись доклада расшифровывается, при необходимости, с использованием ресурсов Ковчега.

• Расшифровка и запись обрабатываются слушателями – участниками доклада, уточняются формулировки, вставляются формулы, ссылки и т.п.

• Результат обработки представляется автору, который дорабатывает текст; он может включить кого-то из слушателей доклада в соавторы и т.д.

Таким образом:

1. Традиционная структура, на которой базируется Ковчег – **профессиональное сообщество** с его связями, иерархией, мотивами и т.п., непрерывно выстраивающимися столетиями. У его членов может не хватать ресурсов и мотивов для создания материалов Ковчега своими силами.

2. Эта структура усиливается **вики-сообществом**, соседствующим с профессиональным, расширяющим и дополняющим его. Это сообщество динамически формируется всеми людьми, обладающими мотивацией к вхождению в проект. У этого сообщества и отдельных его членов может не хватать компетентности и влияния для создания авторитетного и качественного источника знания.

Наше научно-образовательное сообщество может дать пример сбалансированного сочетания традиционных и цифровых механизмов.

Качество материалов на Ковчеге в БРЭ по той или иной области знания становится отражением качества научного знания профессионалов и их заинтересованности в поддержании этого качества, количество и объем материалов будут отражать также интерес широкого сообщества к данной проблематике.

6. МЕТОДОЛОГИЯ И ТЕХНОЛОГИЯ КОВЧЕГА

Ковчег представляет собой структуру, предназначенную для непрерывного сетевого создания сетевого продукта для сетевого потребителя с использованием профессиональных иерархий для обеспечения качества.

Она может рассматриваться как реализация метафоры машинного обучения в техно-социальной среде коллективной расширенной личности. В соответствующей “нейро-сети” присутствует слой потребителей разрабатываемого содержания, слой разработчиков – авторов-редакторов, слой экспертов – РГ, слой заказчика – БРЭ. Мы планируем процесс обучения, совершенствования правил взаимодействия, – как часть функционирования сети.

Ядро платформы построено на базе вики-системы, разработанной на факультете Вычислительной математики и кибернетики МГУ. Выбор этого вида программного обеспечения обусловлен целым рядом преимуществ:

• Обеспечение возможности одновременной работы большого числа удаленных пользователей.

• Сохранение подробной истории всех редакций формируемого пользователями информационного потока.

• Возможность для пользователей работать с системой “тонкий клиент”, поскольку программное обеспечение и информация хранятся в центральном ядре.

• Возможность использования системы унифицированного простого языка вики-разметки для формирования текстов, включая формулы.

• Возможность простого создания гипертекстовых документов и шаблонов, организации ссылок.

7. ДАЛЬНЕЙШИЕ НАПРАВЛЕНИЯ РАЗВИТИЯ КОВЧЕГА

Дальнейшие направления развития Ковчега могут охватывать:

• Размещение архивов препринтов, аналогично зарубежным архивам.

• Размещение материалов открытого доступа, в том числе, например, учебных мульти-медиа курсов, видео-хостинг, в том числе – интервью, воспоминаний, относящихся к истории математики и истории науки.

• Размещение материалов с иными правилами разработки и доступа.

• Размещение учебных сред (LMS) или стыковка с ними.

• Система видео-конференций с автоматизированным аннотированием, расшифровкой.

РГ могут стать активными участниками издания электронных научных журналов, различных форм экспертизы.

8. ЗАКЛЮЧЕНИЕ

Научно-образовательный проект МГУ “Ковчег знаний” базируется на цифровой платформе, разработанной специалистами МГУ им. М.В. Ломоносова. Платформа обеспечивает возможность одновременной работы над базами данных практически неограниченного числа специалистов формирующих, развивающих и поддерживающих единую сетевую структуру.

Интерфейсом проекта с российским и русскоязычным сообществом является портал “Большой российской энциклопедии”.

БЛАГОДАРНОСТИ

Авторы выражают благодарность академику В.А. Садовничему за постановку задачи, С.Л. Кравцу за полезное обсуждение и поддержку.

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Работа А.С. Бубнова и Е.Н. Раевского была поддержана Школой математических методов анализа сложных систем МГУ, А.Л. Семенова – грантом РНФ 22-11-00177.

СПИСОК ЛИТЕРАТУРЫ

1. *Семенов А.Л., Зискин К.Е.* Расширенная личность как основной субъект и предмет философского анализа. Следствия для образования // Человек и системы искусственного интеллекта, ред. Лекторский В.А. СПб.: ООО “Издательство "Юридический центр””, 2022. С. 172–200. ISBN 978-5-94201-835-1.
2. *Clark A.* Being there: Putting brain, body, and world together again // MIT press, 1998. URL: http://www2.econ.iastate.edu/tesfatsi/BeingThere.AClark1998.EntireBook.pdf (дата обращения 08.11.2022).
3. *Серр М.* Девочка с пальчик // М.: Ад Маргинем Пресс, 2016. Оригинал: Serres M. Petite Poucette. Éditions Le Pommier, Paris, 2012.
4. *Выготский Л.С.* Инструментальный метод в психологии // Собр. соч. В 6 т. Т. 1. 1982. URL:

http://elib.gnpbu.ru/text/vygotsky_ss-v-6tt_t1_1982/go,108;fs,1 (дата обращения 08.11.2022).

5. *Vygotsky L.S.* The instrumental method in psychology // 1981, URL: https://www.marxists.org/archive/vygotsky/works/1930/instrumental.htm (дата обращения 08.11.2022).
6. *Vygotsky L.S.* Mind in Society: The Development of Higher Psychological Processes // Harvard University Press, 1980.
7. Платон. Федр // https://www.plato.spbu.ru/TEXTS/PLATO/Academia/005-02.pdf. См. также обсуждение: URL: https://cyberleninka.ru/article/n/mif-ob-izobretenii-pismennosti-v-dialoge-platona-fedr-sovremennoe-osmyslenie-idei-sokrata-i-platona-o-preimuschestve-ustnogo-slova-2 (дата обращения 08.11.2022).
8. *Turing A.M.* On Computable Numbers, with an Application to the Entscheidungsproblem // Proc. Lond. Math. Soc., series 2, vol. 42, 1937. Pp. 230–265. Correction: ibid. vol. 43, 1937. Pp. 544–546.
9. *Komensky J.A.* Orbis Pictus. Wikipedia page // https://en.wikipedia.org/wiki/Orbis_Pictus
10. Энциклопедия Дидро и Д'Аламбера. Виртуальная выставка // Электронные выставки Иностранки, URL: https://press-libfl.tilda.ws/enciklopediya-didro-i-dalambera-triumf-prosveshcheniya (дата обращения 08.11.2022).
11. *Семенов А.Л., Булин-Соколова Е.И, Муранов А.А. и др.* Цифровые технологии в начальной школе. Вход в будущий мир // Информатизация образования и методика электронного обучения: цифровые технологии в образовании. Материалы VI Международной науч. конф., г. Красноярск, 20–23 сентября 2022 г. В 3 ч. Ч. 2 / под общ. ред. М.В. Носкова. – Красноярск : КГПУ им. В.П. Астафьева, 2022. С. 325–329. ISBN 978-5-907558-24-3.
12. *Choi G.* The Internet-Extended Mind: The Psychological Ramifications and Philosophical Implications of Cognitive Offloading // Scripps Senior Theses, 1675. The Claremont Colleges Library, USA, 2021. URL: https://scholarship.claremont.edu/scripps_theses/1675 (дата обращения 08.11.2022).
13. Энциклопедия Байду. Страница Википедии // URL: https://ru.wikipedia.org/wiki/Энциклопедия_Байду (дата обращения 08.11.2022).
14. *Кузнецов С.* А редактор кто? Что не так со статьями в Википедии // Портал RuPosters, 26 марта 2020 г. URL: https://ruposters.ru/news/26-03-2020/oruzhie-informatsionnoi-voini (дата обращения 08.11.2022).
15. Математическая энциклопедия // В 5 томах. Гл. ред. И.М. Виноградов. М.: Сов. энциклопедия, 1977.
16. Encyclopedia of Mathematics // The European Mathematical Society, Springer Verlag. URL: http://encyclopediaofmath.303.si (дата обращения 08.11.2022).
17. Большая российская энциклопедия – электронная версия // URL: https://bigenc.ru (дата обращения 08.11.2022).

EDN: SUSAXK
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 134–145

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

eco2AI: КОНТРОЛЬ УГЛЕРОДНОГО СЛЕДА МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ В КАЧЕСТВЕ ПЕРВОГО ШАГА К УСТОЙЧИВОМУ ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ

© 2022 г. С. А. Буденный[1,2,*], В. Д. Лазарев[2], Н. Н. Захаренко[1], А. Н. Коровин[2], О. А. Плосская[1], Д. В. Димитров[1], В. С. Ахрипкин[1], И. В. Павлов[1], И. В. Оселедец[2,3], И. С. Барсола[4], И. В. Егоров[4], А. А. Костерина[4], Л. Е. Жуков[5]

Представлено академиком РАН А.П. Кулешовым
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

На сегодняшний день в самых различных областях науки и производства возрастает значение искусственного интеллекта (ИИ), в частности, моделей глубокого обучения. Вместе с развитием вычислительных систем наблюдается экспоненциальный рост сложности моделей ИИ, увеличивается их энергопотребление в процессе обучения и инференса. В статье представляется библиотека на Python с открытым исходным кодом eco2AI, который поможет исследователям и аналитикам контролировать потребления энергии и эквивалентную эмиссию CO_2 моделей ИИ. В eco2AI делается акцент на точности отслеживания энергопотребления и правильном региональном учете эмиссии CO_2. Авторы библиотеки призывают исследовательское сообщество к поиску более энергоэффективных архитектур моделей ИИ, а также предлагают концепцию циклического снижения парниковых газов комбинацией концепций устойчивого развития и зеленого ИИ. Код библиотеки и документация размещены в репозитории Github под лицензией Apache 2.0 https://github.com/sb-ai-lab/Eco2AI.

Ключевые слова: ESG, ИИ, Устойчивое развитие, Углеводный след, Экология, Эмиссия CO_2, Парниковые газы

DOI: 10.31857/S2686954322070232

1. ВВЕДЕНИЕ

Несмотря на то, что глобальная повестка ESG (окружающая среда, социальное и корпоративное управление) руководствуется соглашениями, заключенными между странами [1], фактическое развитие принципов ESG происходит посредством внедрения корпоративных, исследовательских и академических стандартов. По этой причине многие компании начали разрабатывать свои стратегии ESG, создавать полноценные отделы, публиковать ежегодные отчеты по устойчивому развитию, выделять дополнительные средства на исследования, в т.ч. цифровых технологий и искусственного интеллекта.

При этом остается актуальной проблема прозрачной и объективной количественной оценки прогресса ESG в области охраны окружающей среды. Это имеет большое значение для ИТ-индустрии, поскольку уже около одного процента мировой электроэнергии потребляется облачными вычислениями и их доля продолжает расти [31]. Искусственный интеллект и машинное обучение (МО) являются важной частью современной ИТ-индустрии, это быстро развивающиеся технологии с огромным потенциалом для прорывного развития. Существует ряд способов, с помощью которых ИИ и МО могли бы смягчить экологические проблемы и антропогенное воздействие. В частности, их можно было бы использовать для генерации и обработки больших данных, для более точного изучения Земли и прогнозирования поведения окружающей среды в различных сценариях [43] с целью улучшения понимания экологических процессов и принятия более обоснованных решений. Существует также потенциал использования ИИ и МО для моделирования результатов вредоносных процессов для экологии, таких как вырубка лесов, эрозия почвы,

[1] *Sber AI Lab, Москва, Россия*
[2] *Институт искусственного интеллекта AIRI, Москва, Россия*
[3] *Сколковский институт науки и технологий, Москва, Россия*
[4] *Sber ESG, Москва, Россия*
[5] *Национальный исследовательский университет "Высшая школа экономики", Москва, Россия*
*****E-mail: sanbudenny@sberbank.ru*

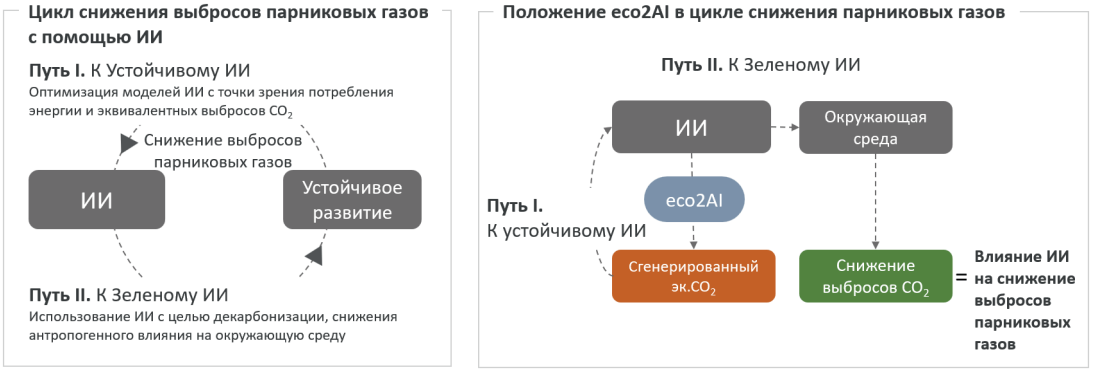

Рис. 1. Обобщенная схема снижения выбросов парниковых газов с помощью ИИ (схема слева), роль *eco2AI* в этой концепции (схема справа).

наводнения, увеличение содержания парниковых газов в атмосфере и т.д. В конечном счете, эти технологии обладают огромным потенциалом для улучшения нашего понимания окружающей среды и контроля над ней.

В настоящее время разрабатывается ряд решений на основе искусственного интеллекта для достижения углеродной нейтральности в рамках концепции "Зеленого искусственного интеллекта". Конечной целью этих решений является сокращение эмиссии парниковых газов (ПГ). В действительности, искусственный интеллект может помочь уменьшить последствия климатического кризиса, например, путем проектирования интеллектуальных систем, развития инфраструктуры с низким уровнем эмиссии и моделирования изменений климата [8]. Также крайне важно учитывать эмиссию CO_2, генерируемую самим ИИ в результате обучения моделей и их применения. Искусственный интеллект развивается в сторону моделей с большей вычислительной сложностью и потреблением электроэнергии и, как следствие, возрастающим в связи с этим косвенным эквивалентным углеродным следом. Экологическое воздействие искусственного интеллекта является основным фактором, который необходимо учитывать при определении возможных рисков. Чтобы модели ИИ/МО были экологически устойчивыми, они должны быть оптимизированы не только с точки зрения точности прогнозирования, но и с точки зрения потребления энергии и воздействия на окружающую среду. Таким образом, отслеживание воздействия ИИ на окружающую среду является первым шагом на пути к концепции устойчивого ИИ. Четкое понимание воздействия ИИ на окружающую среду мотивирует сообщество ИТ специалистов к поиску оптимальных архитектур, потребляющих меньше вычислительных ресурсов [38].

В статье представлена концепция цикла снижения выбросов парниковых газов с помощью ИИ, которая описывает возможности ИИ для достижения целей устойчивого развития (рис. 1). Концепция устойчивого развития обусловливает спрос на более энергоэффективные модели ИИ (рис. 1, путь "К устойчивому ИИ"). С другой стороны, ИИ создает дополнительные возможности для достижения целей устойчивого развития, и мы предлагаем назвать этот путь "К зеленому ИИ". Роль библиотеки *eco2AI* в этом цикле указана в правой части рис. 1. Во-первых, *eco2AI* мотивирует оптимизировать саму технологию ИИ. Во-вторых, если ИИ направлен на снижение выброса ПГ, то общий эффект следует оценивать с учетом генерируемого экв. CO_2, по крайней мере во время обучения модели (и в лучшем случае во время инференса модели). В рамках этой статьи авторы ограничись рассмотрением только пути "К устойчивому ИИ" (см. примеры в главе "Эксперименты").

Научный вклад работы:

• Во-первых, представлена *eco2AI*, библиотека Python с открытым исходным кодом, разработанная для оценки эквивалентной эмиссии CO_2 во время обучения моделей ИИ.

• Во-вторых, описана роль *eco2AI* в контексте концепции цикла снижения выбросов парниковых газов с помощью искусственного интеллекта.

• В-третьих, продемонстрированы примеры использования *eco2AI* в качестве средства оптимизации сложных fusion ИИ моделей.

Статья состоит из следующих разделов: в разделе **2** рассматриваются существующие решения для контроля уровня эквивалентного CO_2 при обучении моделей ИИ и описываются отличия от библиотеки *eco2AI*. В разделе **3** представлена методология вычисления эквивалентного CO_2. В разделе **4** продемострированы варианты использования

библиотеки. Наконец, в разделе **5** подводятся итоги работы. В приложении приводятся примеры использования библиотеки в коде.

2. СМЕЖНЫЕ ИССЛЕДОВАНИЯ

В этой главе описываются современные методы оценки косвенной эквивалентной эмиссии CO_2 для моделей ИИ, приводится краткое описание существующих пакетов с открытым исходным кодом.

2.1. Практика контроля эквивалентной эмиссии CO_2 связанной с работой моделей ИИ

С момента появления моделей глубокого обучения в 2012 г. их сложность росла в геометрической прогрессии, количество параметров удваивалось каждые 3–4 мес и достигло более триллиона параметров в 2022 г. Наиболее известными моделями являются BERT-Large (октябрь 2018 г., 3.4×10^8), GPT-2 (2019 г., 1.5×10^9), T5 (октябрь 2019 г., 1.1×10^{10}), GPT-3 (2020 г., 1.75×10^{11}), Megatron Turing (2022 г., 5.30×10^{11}), Switch Transformer (2022 г., 1.6×10^{12}).

На накопление, разметку, хранение, обработку и использование данных в течение срока их жизни от генерации до утилизации затрачивается значительное число ресурсов, масштаб которых можно оценить на примере инфраструктуры компании Amazon [24]. При этом их эффективный мониторинг важен для разработки сводов правил и законодательства [20].

В [38] проведено крупномасштабное исследование, направленное на количественную оценку приблизительных экологических издержек, связанных с обучением моделей ИИ, широко используемых для задач обработки текстов на естественном языке (NLP). Среди рассмотренных архитектур глубокого обучения таких, как Transformer, ELMo, BERT, NAS, GPT-2 оценивалось комбинированное энергопотребление GPU, CPU и DRAM, скорректированное на показатель эффективности использования энергии (PUE), специфического для конкретного центра обработки данных. Энергопотребление CPU и GPU определялось специализированными программными пакетами: Intel Running Average Power Limit и NVIDIA System Management. Произведением общего потребления энергии и коэффициента эмиссии углерода пересчитывают энергию в косвенную эмиссию CO_2. Авторы подсчитали, что углеродный след для обучения базового BERT составляет около 652 кг, что сравнимо с эмиссией CO_2 при авиаперелете “Нью-Йорк <–> Сан-Франциско” на одного пассажира.

Изучена возможность повышения энергоэффективности моделей NLP (T5, Meena, GShard, Switch Transformer, GPT-3) [30]. Показана возможность повышения энергоэффективности при обучении моделей нейронных сетей с помощью ряда методов, таких как: sparsely activating DL; distillation techniques [22]; pruning, quantization, efficient coding [19]; fine-tuning и transfer-learning [9]; обучение крупных моделей в конкретном регионе с низким энергопотреблением, использование облачных центров обработки данных, оптимизированных с точки зрения энергопотребления. Авторы ожидают, что принятие во внимание данных мер может сократить углеродный след в 10^2–10^3 раз.

2.2. Обзор существующих библиотек для отслеживания углеродного следа моделей ИИ

К настоящему времени разработан ряд библиотек для отслеживания косвенной эквивалентной эмиссии CO_2, связанной с обучением моделей ИИ (см. табл. 1).

Cloud Carbon Footprint [5] – это приложение, которое оценивает потребление энергии и углеродный след поставщиков общедоступных облачных сервисов. Приложение также предоставляет оценки как потребления энергии, так и углеродного следа для всех типов облачных сервисов, с возможностью детализации эмиссии по поставщику услуг, учетной записи, запущенной службе и периоду времени. Кроме того, для сервисов AWS и Google Cloud предоставляются рекомендации по экономии денег и минимизации эмиссии CO_2, прогнозируется экономия средств, а фактические сэкономленные ресурсы отображаются в посаженных деревьях. При этом для датацентров потребление энергии измеряется не на уровне среднего значения, а с использованием точных данных о реальной нагрузке на сервер в ходе его работы. Библиотека позволяет регистрируемыми показателями расширять существующие базы данных систем выставления счетов, конвейеров данных и систем мониторинга.

CodeCarbon [6] – это пакет Python для отслеживания эмиссии углерода, производимого при выполнении любого кода Python – от простых алгоритмов до глубоких нейронных сетей. CodeCarbon учитывает вычислительную инфраструктуру, местоположение, нагрузку на систему и время исполнения кода. Библиотека также позволяет проводить сравнение с эмиссией от обычных видов транспорта.

Carbontracker [4] – это пакет для отслеживания и прогнозирования энергопотребления и углеродного следа при обучении моделей глубокого обучения. Пакет направлен на использование прогноза энергопотребления для проактивного сокращения эмиссии CO_2. Например, обучение модели может быть остановлено по решению пользователя при превышении прогнозируемого

Таблица 1. Функции библиотек с открытым исходным кодом для оценки эквивалентной эмиссии CO_2 при обучении моделей ИИ

Библиотека	Cloud Carbon Footprint	Code Carbon	Carbon Tracker	Experimental Impact Tracker	Tracarbon	Green Algorithms	eco2AI
Общие сведения							
Первый выпуск	2020	2020	2020	2019	2022	2021	2022
Лизензия	Apache 2.0	MIT	MIT	MIT	Apache 2.0	CC-BY-4.0	Apache 2.0
Региональный коэффициент эмиссии	✓	✓	–	✓	✓	✓	✓*
Совместимая ОС							
Linux	✓	✓	✓	✓		✓	✓
Windows	✓	✓	✓			✓	✓
MacOS	✓	✓	✓	✓	✓	✓	✓
Совместимое оборудование							
RAM	✓	✓	✓	✓	✓	✓	✓
CPU	✓	✓	Неизвестно	✓	✓	✓	✓**
GPU	✓	✓	✓	✓	✓	✓	✓
Дополнительно							
Кодирование данных***							✓
WEB интерфейс	✓	✓				✓	

*База данных содержит коэффициенты интенсивности эмиссии для 365 территорий, в т.ч. данные по регионам Австралии ([13], [36]), Канады ([13], [40]), России ([34], [12]) и США ([13], [41]).

** *eco2AI* содержит базу данных CPU, состоящую из 3279 моделей от компаний Intel и AMD.

*** важно в случаях, когда необходимо подтвердить подлинность данных.

экологического ущерба. Библиотека поддерживает множество различных сред и платформ, таких как кластеры, настольные компьютеры и ноутбуки Google Colab, что позволяет работать по принципу plug-and-play [2].

Experiment impact tracker [16] – программный пакет, предоставляющий информацию об энергетическом, вычислительном и углеродном следе моделей машинного обучения. Он обладает следующими функциями: извлечение информации об устройствах CPU и GPU, определение времени начала и окончания эксперимента, учет региона оборудования, на котором проводится эксперимент (по IP-адресу), средняя интенсивность эмиссии углерода в регионе, расчет памяти и частоты процессора в реальном времени [20].

Green Algorithms [17] – это онлайн-инструмент, который позволяет пользователю оценивать и сообщать об углеродном следе в результате вычислений. Он интегрируется с вычислительными процессами и не взаимодействует с существующим кодом, а также учитывает модель CPU, GPU, облачных вычислений, локальных серверов и настольных компьютеров [26].

Tracarbon [39] – это Python библиотека, которая отслеживает энергопотребление устройства и рассчитывает углеродный след. Она автоматически определяет местоположение, модель CPU и GPU и может использоваться в качестве интерфейса командной строки (CLI) с предопределенными или рассчитанными с помощью API (интерфейс прикладного программирования) пользовательскими метриками.

При схожести *eco2AI* с описываемыми библиотеками в ней сделан акцент на следующем: учитываются только системные процессы, связанные непосредственно с обучением моделей (во избежание завышения оценки); база данных региональных коэффициентов интенсивности эмиссии (включено 365 территориальных объек-

тов) и база данных энергопотребления CPU (3279 моделей).

3. МЕТОДОЛОГИЯ

В главе рассматриваются вопросы расчета потребления электроэнергии, коэффициента эмиссии, эквивалентной эмиссии CO_2.

3.1. Расчет энергопотребления

Для расчета энергопотребления вычислительной системы необходимо оценить энергетический вклад каждого аппаратного блока [20]. В библиотеке *eco2AI* оценивается энергия, потребляемая графическим процессором (GPU), центральным процессором (CPU) и памятью (RAM) в силу их наибольшего вклада в энергопотребление среди всех аппаратных блоков. В процессе измерения пренебрегается вкладом краевых эффектов, связанных с завершающимися процессами, из-за их относительно небольшого влияния на общее энергопотребление. Также не учитывается энергопотребление систем хранения (SSD, HDD), так как они не имеют прямой связи с активностью процесса модели, но скорее всего о процессе постоянного хранения данных. Энергопотребление системы измеряется в джоулях (Дж), но чаще киловатт-часах (кВт·ч) – единице энергии, равной одному киловатту мощности, поддерживаемой в течение одного часа.

GPU. Библиотека *eco2AI* работает с GPU производства NVIDIA. Функционал обеспечивается библиотекой *Pynvml*, в которой реализован интерфейс Python для функций управления и мониторинга графических процессоров. Эта оболочка Python для библиотеки *nvml* от NVIDIA позволяет обнаруживать большинство GPU устройств NVIDIA, а также отслеживать количество активных устройств, их имена, используемую память, температуру, максимальную мощность (энергопотребление GPU может немного превышать это значение) и текущее энергопотребление каждого устройства. Для корректного обнаружения GPU требуется установка CUDA. Общее энергопотребление всех активных GPU E_{GPU} (кВт · ч) равно произведению энергопотребления графических процессоров на время их работы:

$$E_{GPU} = \int_0^T P_{GPU}(t)dt,$$

где P_{GPU} – суммарное энергопотребление(кВт) всех графических процессоров, определяемое функционалом *Pynvml*, T – время работы графических процессоров (ч). Если трекер не обнаруживает ни одного графического процессора, то энергопотребление GPU считается равным нулю.

CPU. Для мониторинга энергопотребления CPU использовался Python модуль *psutil*. Важным акцентом является то, что во избежание переоценки, в *eco2AI* реализован функционал, фильтрующий все фоновые процессы, библиотека учитывает только текущий процесс, связанный с обучением модели. Процент загрузки CPU определяется соотношением процента использования CPU и количества ядер. На данный момент создана самая полная база данных, содержащая 3279 уникальных процессоров для моделей Intel и AMD. Каждому наименованию модели CPU соответствует значение расчетной тепловой мощности (TDP), которое эквивалентно потребляемой мощности при длительных нагрузках. Суммарное энергопотребление всех активных процессорных устройств E_{CPU} (кВт · ч) равно произведению потребляемой мощности CPU устройств на время их загрузки $E_{CPU} = TDP\int_0^T W_{CPU}(t)dt$, где TDP – эквивалентная удельная мощность модели CPU при длительной нагрузке (кВт), W_{CPU} – суммарная загрузка всех процессоров. Если трекер не может определить ни одно процессорное устройство, энергопотребление процессора устанавливается равным 100 Вт [27].

RAM. Оперативная память является важным источником потребления энергии в современных вычислительных системах, особенно когда необходимо выделить или обработать значительный объем данных. Однако учет энергопотребления оперативной памяти проблематичен, так как ее энергопотребление сильно зависит от режима работы с данными: чтение, запись или хранение. В *eco2AI* энергопотребление RAM считается пропорциональным используемому количеству памяти и рассчитывается следующим образом: $E_{RAM} = 0.375\int_0^T M_{RAM_i}(t)dt$, где E_{RAM} – потребляемая энергия оперативной памяти (кВт · ч), M_{RAM_i} – используемая память (Гб), измеренная с помощью *psutil*, а 0.375 Вт/Гб – расчетное удельное энергопотребление модулей DDR3, DDR4 [27].

3.2. Региональный коэффициент интенсивности эмиссии CO_2

Эмиссия CO_2, связанная с производством электроэнергии, в значительной степени различается между странами и регионами. Для учета региональной зависимости эффективности эмиссии используют коэффициент интенсивности эмиссии γ, который определяется как масса выбрасываемого CO_2, выраженной в кг, на каждый мегаватт-час (МВт · ч) производимой электроэнергии. Коэффициент интенсивности эмиссии опреде-

Таблица 2. Коэффициенты интенсивности эмиссии CO_2 для некоторых регионов

Страна	ISO-Alpha-2 code	ISO-Alpha-3 code	UN M49 code	Коэффициент интенсивности эмиссии CO_2, кг/МВт · ч
Канада	CA	CAN	124	120.49
Франция	FR	FRA	250	67.53
Индия	IN	IND	356	625.57
Парагвай	PY	PRY	600	23.92
Замбия	ZM	ZMB	894	120.78

ляется региональным энергетическим балансом: $\gamma = \sum_i f_i e_i$, где i – индекс, относящийся к i-му источнику энергии (например, уголь, возобновляемые источники энергии, нефть, газ и т.д.), f_i – доля i-го источника энергии для конкретного региона, e_i – эмиссия, производимая сожжением килограмма массы этого источника энергии. Энергетический баланс в свою очередь определяется структурой производства электроэнергии, географическим положением, используемым топливом и технологическими процессами. Следовательно, чем выше доля возобновляемой энергии, тем меньше суммарный коэффициент интенсивности эмиссии. В противном случае большая доля углеводородных энергоресурсов в балансе приводит к более высокому значению коэффициента интенсивности эмиссии.

Библиотека *eco2AI* включает в себя базу данных коэффициентов интенсивности эмиссии для 365 регионов на основе общедоступных данных по 209 странам [11], а также региональных данных по таким странам, как Австралия ([13], [36]), Канада ([13], [40]), России ([34], [12], [28]) и США ([13], [41]). В настоящее время это самая большая база данных среди рассмотренных трекеров, что делает оценку энергопотребления более точной.

База данных имеет следующую структуру: название страны, код ISO-Alpha-2, код ISO-Alpha-3, код UN M49 и значение коэффициента интенсивности эмиссии. Коэффициенты интенсивности эмиссии CO_2 для некоторых регионов с различным энергетическим балансом приведен в табл. 2. Библиотека *eco2AI* автоматически определяет страну и регион пользователя по IP и находит в базе данных соответствующий им коэффициент эмиссии CO_2. Если коэффициент по какой-либо причине автоматически не определен, он устанавливается равным 436.5 кг/МВт · ч, среднее значение по миру [11]. В *eco2AI* можно вручную задать регион, страну и значение коэффициента интенсивности эмиссии.

3.3. Эквивалентное значение эмиссии CO_2

Наконец, эквивалентное значение эмиссии CO_2(кг), образующегося при обучении моделей, определяется путем умножения общего энергопотребления CPU, GPU и RAM на коэффициент эмиссии γ (кг/кВт · ч) и коэффициент PUE:

$$CF = \gamma \cdot PUE \cdot (E_{CPU} + E_{GPU} + E_{RAM}),$$

PUE – эффективность энергопотребления датацентра, необходимая, если процесс обучения выполняется в облаке. В *eco2AI* PUE является – опциональным задаваемым вручную параметром по умолчанию, равным 1.

4. ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

В этом разделе представлены эксперименты по отслеживанию эквивалентной эмиссии CO_2 с помощью *eco2AI* при обучении Malevich и Kandinsky. Malevich и Kandinsky – это большие мультимодальные text2image модели [18] с 1.3 миллиардом и 12 миллиардами параметров соответственно, способные генерировать произвольные изображения по текстовой строке.

В работе представлены результаты дообучения моделей Malevich и Kandinsky на наборе данных Emojis [37] и обучения Malevich с использованием оптимизированной функцией активации GELU [21].

4.1. Дообучение мультимодальных моделей

В этом разделе представлены примеры использования *eco2AI* для мониторинга дообучения моделей Malevich и Kandinsky (например, CO_2, кг; мощность, кВт·ч) на датасете Emojis. Malevich и Kandinsky – это мультимодальные трансформеры, которые обучаются условному распределе-

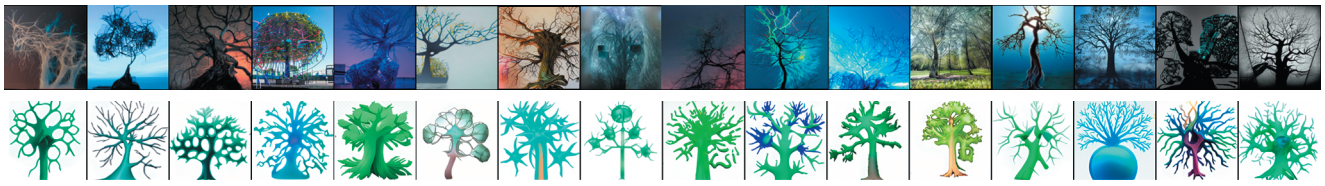

Рис. 2. Генерация изображений Malevich (сверху) vs Emojich XL (снизу) при вводе текста “Дерево в виде нейрона”.

Рис. 3. Генерация изображений Kandinsky (сверху) vs Emojich XXL (снизу) при вводе текста “Зеленый искусственный интеллект”.

нию изображений. Точнее, они авторегрессивно обрабатывают токены текста и изображения как единый поток данных (см., например, DALL-E [33]. Эти модели представляют собой декодеры-трансформеры [42] с 24 и 64 слоями, 16 и 60 attention heads, с размерностями скрытого пространства 2048 и 3840 соответственно и функцией активации GELU.

И Malevich, и Kandinsky работают со 128 текстовыми токенами, которые генерируются из текстовых входных данных с помощью токенизатора YTTM [44], и 1024 графическими токенами, которые получаются при кодировании входного изображения с помощью генеративно-состязательной сети Sber-VQGAN (предварительно обученный VQGAN [15] с Gumbel Softmax Relaxation [25].

Набор данных emoji [35] для дообучения содержит 2749 уникальных иконок emoji и 1611 уникальных текстов, которые были собраны с помощью парсинга веб-сайтов (разница в количествах связана с тем, что есть наборы, внутри которых смайлики отличаются только цветом, причем некоторые элементы являются омонимами).

Malevich и Kandinsky были обучены с точностями fp16 и fp32 соответственно. В обоих экспериментах используется оптимизатор Адам (8-бит) [7]. Эта реализация уменьшает объем памяти графического процессора, необходимой для хранения градиента. Выбраны следующие параметры обучения: начальное значение коэффициента скорости обучения(lr) 4×10^{-7}, максимальное значение – 10^{-5} и конечное значение – 2×10^{-8}. Модели обучались на 40 эпохах с коэффициентом изменения lr 0.3, размером батча 4 для Malevich и размером батча 12 для Kandinsky, с большим коэффициентом потери (loss coefficient) для изображения, равным 1000, и замороженными слоями с feed forward и attention.

Модели Malevich и Kandinsky обучались на 1 GPU Tesla A100 (80 ГБ) и 8 GPU Tesla A100 (80 ГБ) соответственно. Стоит отметить, что для обучения модели Kandinsky использовался оптимизатор распределенной модели DeepSpeed [45]. Исходный код, используемый для дообучения Malevich, доступен на Kaggle [14].

Результаты дообучения Malevich и Kandinsky называются Emojich XL и Emojich XXL соответственно. Сравнение результатов генерации Malevich и Emojich XL и Kandinsky и Emojicha XXL на некоторых текстовых токенах (см. рис. 2 и 3) позволяет визуально оценить качество дообучения (стиль сгенерированных изображений подстраивается под стиль emoji). Генерация изображения начинается с текстовой строки, описывающей ожидаемое содержимое. Когда токенизированный текст передается в Emojich, модель автоматически генерирует оставшиеся токены изображения.

Каждый токен изображения выбирается поэлементно из предсказанного полиномиального распределения вероятностей латентого пространства изображения с использованием nucleus sampling top-p и top-k с температурой [23] в качестве алгоритма декодирования. Изображение получается из сгенерированной последовательности скрытых векторов декодером Sber-VQGAN.

Все приведенные примеры генерируются автоматически со следующими гиперпараметрами: размер батча 16 и 6, top-k 2048 и 768, top-p 0.995 и 0.99, температура 1.0, 1 GPU Tesla A100 для Malevich (а также Emojich XL) и Kandinsky (а также Emojich XXL) соответственно.

Таблица 3. Эмиссия CO_2 и энергопотребление для дообучения моделей Malevich и Kandinsky

Модель	Время обучения	Энергия, кВт · ч	CO_2, кг	GPU	CPU	Размер батча
Malevich	4h 19m	1.37	0.33	A100 Graphics, 1	AMD EPYC 7742 64-Core	4
Kandinsky	9h 45m	24.50	5.89	A100 Graphics, 8	AMD EPYC 7742 64-Core	12

Параметры дообучения, результаты потребления энергии и эквивалентной эмиссии CO_2 приведены в табл. 3. Можно отметить, что при дообучении Kandinsky выделяется более чем в 17 раз больше CO_2, чем при дообучении Malevich.

Таким образом, можно видеть, что библиотека *eco2AI* позволяет контролировать энергопотребление при обучении (и дообучении) больших моделей не только на одном GPU, но и на нескольких GPU, что имеет важное значение в случае использования библиотек оптимизации для распределенного обучения, например DeepSpeed.

4.2. Предобучение мультимодальных моделей

Обучение больших моделей, таких как Malevich, требует затраты большого количества ресурсов. В этом разделе рассматривается случай повышения энергоэффективности модели с использованием квантованной функции активации GELU. Квантованная GELU [29] – это разновидность функции активации GELU [21], которая сохраняет градиенты модели с разрешением в несколько бит, тем самым занимая меньше памяти GPU и затрачивая меньше вычислительных ресурсов (см. рис. 4). Если быть точнее, сравниваются ошибка и энергоэффективность версии модели Malevich с обычным GELU и версией Malevich с квантованным GELU 4-бит, 3-бит, 2-бит и 1-бит с помощью библиотеки *eco2AI*. Для всех версий моделей использова-лись тот же оптимизатор, планировщик и алгоритм обучения, что и в экспериментах по дообучению. Для обеспечения воспроизводимости каждый эксперимент запускался 5 раз со случайным начальным числом. Набор данных для обучения состоял из 300 000 объектов. Каждый образец был пропущен через модель только один раз с размером батча, равным 4. Набор данных для валидации состоял из 100 000 объектов. Для отслеживания углеродного следа во время обучения в режиме реального времени использовалась библиотека *eco2AI*.

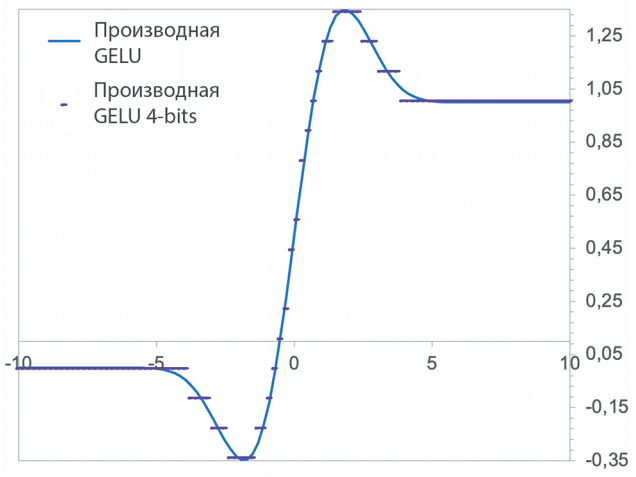

Рис. 4. Функция активации GELU и ее оптимизированная 4-битная аппроксимация.

Как видно из рис. 5а, потери при валидации Malevich с 4-битным, 3-битным GELU и Malevich с обычным GELU почти одинаковы (рост 0.06%), тогда как 2-битный, 1-битный GELU демонстрируют увеличение ошибки на валидационной выборке примерно на 0.42%. При этом, 4-битный GELU и 3-битный GELU примерно на 15 и 17% соответственно более энергоэффективны по сравнению с исходным GELU и приводят к соответственно меньшей эмиссии CO_2 на одном и том же шаге обучения (рис. 5б).

Использование 1-битного GELU позволило дополнительно уменьшить эмиссию только на 0.05% CO_2. Производительность моделей представлена на рис. 5в и в табл. 4. Таким образом, GELU 3-бит обеспечивает точность модели, близкую к оригинальной GELU, при этом потребляя на 17% меньше энергии и, следовательно, производя меньше эквивалентной эмиссии CO_2.

Таким образом, библиотека *eco2AI* может отслеживать энергопотребление и углеродный след при обучении моделей ИИ в режиме реального времени, помогает реализовывать и демонстрировать различные алгоритмы оптимизации энергопотребления (например, с помощью квантованых функций активации).

5. ЗАКЛЮЧЕНИЕ

Несмотря на большой потенциал ИИ в решении экологических проблем, сам ИИ может также быть источником углеродного следа. Библиотека *eco2AI* может помочь ИИ-сообществу понять влияние моделей ИИ на окружающую среду во время обучения и инференса и организовать систематический мониторинг эквивалентной эмиссии углерода. *eco2AI* – это библиотека с открытым

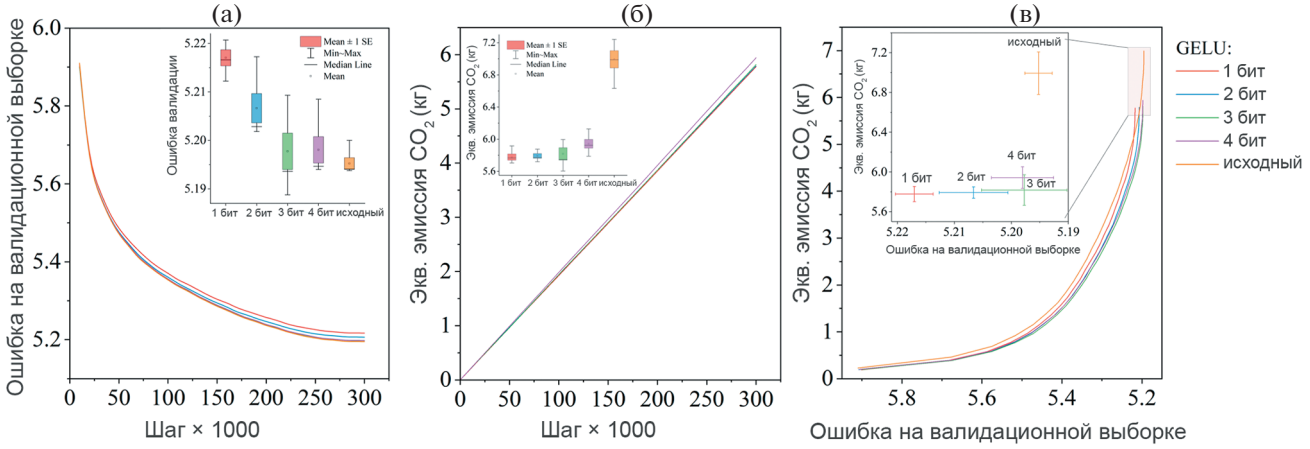

Рис. 5. Сравнение квантованных функций активации GELU и исходного GELU, использованных при обучении модели Malevich: (а) ошибка на валидационной выборке на каждом шаге обучения (ящичковая диаграмма на вставке показывает статистику ошибки при валидации каждой модели на 300000 шаге), (б) Эмиссия CO_2 на каждом шаге обучения моделей (диаграмма на вставке показывает статистику эмиссии CO_2 каждой модели на шаге 300000), (в) Эмиссия CO_2 для ошибки на валидационной выборке для каждой модели. (Для более наглядной демонстрации различия между моделями на вставке показана увеличенная область графика с пиковой эмиссией CO_2).

исходным кодом, рассчитывающая эквивалентную эмиссию углерода при обучении или инференсе моделей ИИ на Python исходя из энергопотребления GPU, CPU и RAM. В *eco2AI* делается акцент на точности расчета энергопотребления, что достигается с помощью учета региональных коэффициентов интенсивности эмиссии CO_2 и точным определением загрузки CPU.

В работе приведены примеры использования *eco2AI* для контроля эмиссии CO_2 при дообучении больших моделей text2image, таких как Malevich и Kandinsky, а также для оптимизации функции активации GELU, используемой при обучении модели Malevich. С помощью *eco2AI* было продемонстрировано, что использование 3-битного GELU позволяет уменьшить эквивалентную эмиссию CO_2 при обучении модели примерно на 17%. Авторы ожидают, что *eco2AI* может помочь сообществу перейти к “Зеленому ИИ”, в рамках предложенной концепции циклического снижения выбросов парниковых газов.

ПРИЛОЖЕНИЕ

Примеры использования библиотеки eco2AI

eco2AI – библиотека на Python с открытым исходным кодом, распространяется под лицензией Apache 2.0 [3]. Она доступна для установки с PyPI [32], а также ее исходный код можно найти на репозитории в GitHub [10]. Ниже представлены различные способы интеграции библиотеки *eco2AI* в код своего Python проекта.

После установки библиотеки, ее нужно импортировать в свой код, создать объект класса *eco2ai.Tracker()*, затем вызвать до основного кода метод *.start()* класса *eco2ai.Tracker()* и после него метод *.stop()* (см. Листинг 1).

Таблица 4. Эквивалентная эмиссия CO_2 и энергопотребление предобученной модели Malevich на наборе данных, состоящем из 300000 изображений за одну эпоху (A100 Graphics, AMD EPYC 7742 64-Core)

Модель	Время обучения, ч	Энергия, кВт · ч	CO_2, кг	Ошибка валидации
Malevich, GELU исходная	75.2 ± 1.3	29.1 ± 1	7.0 ± 0.24	5.195 ± 0.002
Malevich, GELU 4-бит	67.2 ± 1.5	24.7 ± 0.5	5.94 ± 0.12	5.198 ± 0.005
Malevich, GELU 3-бит	67.2 ± 1.5	24.2 ± 0.7	5.81 ± 0.17	5.198 ± 0.008
Malevich, GELU 2-бит	66.5 ± 0.3	24.0 ± 0.3	5.79 ± 0.06	5.207 ± 0.006
Malevich, GELU 1-бит	67.7 ± 0.73	24 ± 0.36	5.77 ± 0.09	5.217 ± 0.003

Листинг 1: Базовый пример использования

```
import eco2ai
tracker = eco2ai.Tracker(project_name="YourProjectName",
        experiment_description="training_the_<your_model>_model")

tracker.start()
<your gpu & (or) cpu calculations>
tracker.stop()
```

Еще одним способом встроить трекер в свой является использование декораторов (см. Листинг 2). Декоратор *track* нашей библиотеки позволяет модифицировать любую функцию так, чтобы при ее выполнении система вела расчет косвенно выделившегося CO_2.

Листинг 2: Пример использования декораторов

```
from eco2ai import track


@track
def train_func(model, dataset, optimizer, epochs, *args, **kwargs):
  ...
train_func(model, dataset, optimizer, epochs, *args, **kwargs)
```

После каждого вызова метода *.stop()* трекер завершит работу, и все результаты будут записаны в локальный файл *"emission.csv"*, название которого также можно задавать параметром *file_name* класса *eco2ai.Tracker()*. Результирующий файл представляет собой таблицу со следующими столбцами:

- *id* – уникальный id эксперимента;
- *project_name* – название проекта, задаваемое пользователем;
- *experiment_description* – описание эксперимента, также задаваемое пользователем;
- *epoch* – информация об эпохе обучения, в случае, если используется данная функция;
- *start_time* – дата начала эксперимента в формате yyyy-mm-dd hh:mm:ss;
- *duration(s)* – длительность эксперимента в секундах;
- *power_consumption(kWh)* – энергия, косвенно выделившаяся за время эксперимента, выраженная в кВт·ч;
- *CO_2_emissions(kg)* – масса косвенно выделившегося углекислого газа за время эксперимента, выраженная в кг;
- *CPU_name* – информация о наименовании и количестве CPU;
- *GPU_name* – информация о наименовании и количестве GPU;
- *OS* – название ОС устройства, на котором проводился эксперимент;
- *region/country* – регион и страна, определяющиеся автоматически по IP или задаваемые пользователем вручную с помощью параметров *alpha_2_code* и *region* класса *eco2ai.Tracker()*;
- *cost* – стоимость затраченной электроэнергии в условных единицах. Расчитывается в случае, если пользователь задал параметр *electricity_pricing*;

eco2AI позволяет пользователю записывать информацию об обучении в зашифрованном виде (см. Листинг 3). Эта функция может быть полезна в тех случаях, когда требуется защитить файл с результатами от ручной модификации пользователем. Для кодировки результатов необходимо задать параметр *"encode_file"* класса *eco2ai.Tracker()*, тогда в файл эксперимента будут записаны защифрованные данные *"encoded_emission.csv"*.

Листинг 3: Функция кодирования результатов

```
import eco2ai

tracker = eco2ai.Tracker(
        file_name='encoded_emissions.csv',
        project_name="Test_1",
        experiment_description="testing_Eco2AI_in_encoding_mode",
        encode_file=True,
)
```

В *eco2AI* также реализована возможность контролировать эквивалентную эмиссию CO_2 в каждой эпохе обучения моделей машинного и глубокого обучения. В Листинге 4 разобран простейший пример использования данной возможности.

Листинг 4: Эксперимент для глубокого обучения

```
tracker = eco2ai.Tracker(
        project_name="CIFAR10_for_ESG",
        experiment_description="ML_tracking",
        file_name="emission.csv",
        emission_level=None,
        alpha_2_code="AU",
        region="Queensland",
        )
tracker.start_training()

net = Net()
criterion = nn.CrossEntropyLoss()
optimizer = optim.SGD(net.parameters(), lr=0.001, momentum=0.9)
parameters_to_save = dict()
for epoch in range(epochs):
parameters_to_save["loss"] = train_epoch(net, optimizer, trainloader)
        parameters_to_save["train_accuracy"] = get_accuracy(net, trainloader)
        parameters_to_save["test_accuracy"] = get_accuracy(net, testloader)

        tracker.new_epoch(parameters_to_save)
tracker.stop_training()
```

Подробную документацию для каждого метода класса *eco2ai.Tracker()* и его параметров можно вызвать с помощью функции *help(eco2ai.Tracker())*.

СПИСОК ЛИТЕРАТУРЫ

1. Paris Agreement. Paris agreement. In Report of the Conference of the Parties to the United Nations Framework Convention on Climate Change (21st Session, 2015: Paris). Retrived December, volume 4, page 2017. HeinOnline, 2015. Open AI. URL https://openai.com/blog/ai-and-compute/
2. *Lasse F Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan.* Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. arXiv preprint arXiv:2007.03051, 2020.
3. Apache licence 2.0. URL https://www.apache.org/licenses/LICENSE-2.0.
4. Carbontracker. URL https://github.com/lfwa/carbontracker.
5. Cloud Carbon Footprint. URL https://github.com/cloud-carbon-footprint/cloud-carbon-footprint.
6. Codecarbon. URL https://github.com/mlco2/codecarbon.
7. *Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer.* 8-bit optimizers via blockwise quantization. arXiv preprint arXiv:2110.02861, 2021.
8. *Payal Dhar.* The carbon impact of artificial intelligence. Nat. Mach. Intell., 2(8):423–425, 2020.
9. *Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith.* Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. arXiv preprint arXiv:2002.06305, 2020.
10. eco2ai github. URL https://github.com/sb-ai-lab/eco2AI.
11. Ember. Global electricity review 2022, Mar 2022. URL https://ember-climate.org/insights/research/global-electricity-review-2022/.
12. EMISS. The unified interdepartmental information and statistical systems (emiss). URL https://fedstat.ru/indicator/58506.
13. Emissions factors sources, 2021. URL https://www.carbonfootprint.com/docs/2022_01_emissions_factors_-sources_for_2021_electricity_v10.pdf.
14. emojich ruDALL-E. URL https://www.kaggle.com/shonenkov/emojich-rudall-e.

15. *Patrick Esser, Robin Rombach, and Björn Ommer.* Taming transformers for high-resolution image synthesis, 2020.
16. Experiment impact tracker. URL https://github.com/Breakend/experiment-impact-tracker.
17. Green algorithms tool. URL https://github.com/GreenAlgorithms/green-algorithms-tool.
18. *Julia Gusak, Daria Cherniuk, Alena Shilova, Alexander Katrutsa, Daniel Bershatsky, Xunyi Zhao, Lionel Eyraud-Dubois, Oleg Shlyazhko, Denis Dimitrov, Ivan Oseledets, and Olivier Beaumont.* Survey on large scale neural network training, 2022.
19. *Song Han, Huizi Mao, and William J Dally.* Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149, 2015.
20. *Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau*. Towards the systematic reporting of the energy and carbon footprints of machine learning. Journal of Machine Learning Research, 21(248):1–43, 2020.
21. *Dan Hendrycks and Kevin Gimpel.* Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016.
22. *Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al.* Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2(7), 2015.
23. *Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi*. The curious case of neural text degeneration. CoRR, abs/1904.09751, 2019. URL http://arxiv.org/abs/1904.09751.
24. *Vladan Joler Kate Crawford*. Anatomy of an ai system, 2018. URL http://www.anatomyof.ai.
25. *Matt J. Kusner and José Miguel Hernández-Lobato*. Gans for sequences of discrete elements with the gumbel-softmax distribution, 2016.
26. *Loc Lannelongue, Jason Grealey, and Michael Inouye.* Green algorithms: quantifying the carbon footprint of computation. Advanced science, 8(12):2100707, 2021.
27. *DA Maevsky, EJ Maevskaya, and ED Stetsuyk*. Evaluating the ram energy consumption at the stage of software development. In Green IT Engineering: Concepts, Models, Complex Systems Architectures, pages 101–121. Springer, 2017.
28. Minprirody (Russia). URL https://xn–d1ahaoghbejbc5k.xn–p1ai/documents/active/664/.
29. *Georgii Novikov, Daniel Bershatsky, Julia Gusak, Alex Shonenkov, Denis Dimitrov, and Ivan Oseledets.* Few-bit backward: Quantized gradients of activation functions for memory footprint reduction, 2022.
30. *David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean*. Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350, 2021.
31. *Pesce M*. Cloud computing's coming energy crisis. IEEE Spectrum, 2021.
32. PyPi Eco2AI. URL https://pypi.org/project/eco2AI/.
33. *Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.* Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, pages 8821–8831. PMLR, 2021. URL http://proceedings.mlr.press/v139/ramesh21a.html.
34. Rosstat. Russian federal state statistics service. URL https://rosstat.gov.ru/enterprise_industrial.
35. russian-emoji. URL https://www.kaggle.com/datasets/shonenkov/russian-emoji.
36. Science and Resources. National greenhouse accounts factors. URL https: //www.industry.gov.au/sites/default/files/August%202021/document/national-greenhouse-accounts-factors-2021.pdf.
37. *Alex Shonenkov, Daria Bakshandaeva, Denis Dimitrov, and Aleksandr Nikolich*. Emojich zero-shot emoji generation using russian language: a technical report. arXiv preprint arXiv:2112.02448, 2021.
38. *Emma Strubell, Ananya Ganesh, and Andrew McCallum.* Energy and policy considerations for deep learning in nlp. arXiv preprint arXiv:1906.02243, 2019.
39. Tracarbon. URL https://github.com/fvaleye/tracarbon.
40. UNFCCC. Canada, national inventory report, 2021. URL https://unfccc.int/sites/default/files/resource/can-2021-nir-12apr21.zip. Part 3, page 60.
41. USA EPA. egrid2020, May 2020. URL https://www.epa.gov/system/files/documents/2022-01/egrid2020_data.xlsx.
42. *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.* Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M.Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008, 2017. URL https://proceedings.neurips. cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
43. *Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini*. The role of artificial intelligence in achieving the sustainable development goals. Nature communications, 11(1):1–10, 2020.
44. *YouTokenToMe*. URL https://github.com/VKCOM/YouTokenToMe.
45. ZeRO-3. URL https://www.deepspeed.ai/2021/03/07/zero3-offload.html.

EDN: KOMTUA
ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 146–148

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

FusionBrain: ИССЛЕДОВАТЕЛЬСКИЙ ПРОЕКТ ПО МУЛЬТИМОДАЛЬНОМУ И МУЛЬТИЗАДАЧНОМУ ОБУЧЕНИЮ

© 2022 г. Д. В. Димитров[1,2,*], А. В. Кузнецов[1,2], А. А. Мальцева[1], Е. Ф. Гончарова[2]

Представлено академиком РАН С.С. Гончаровым
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

FusionBrain – это исследовательский проект, основными задачами которого являются разработка эффективных мультизадачных и мультимодальных моделей и применение их для решения широкого круга практических задач. Общая цель и идея проекта – научиться создавать модели, которые смогут как можно более эффективно извлекать дополнительные важные знания из большого количества модальностей и задач при обучении, и за счет этого лучше решать разные другие задачи. Исследования проводятся во многих модальностях: тексты, изображения, аудио, видео, языки программирования, графы (например, молекулярные структуры), временные ряды и так далее. Список решаемых задач очень большой: от классических задач CV и NLP до задач, вовлекающих разные модальности: VideoQA, Visual Commonsense Reasoning, IQ tests (эти задачи сложны даже для человека). Изучается также способность моделей решать задачи, сформулированные на естественном или визуальном языках, и даже справляться со скрытыми задачами (для которых в обучающей выборке отсутствовали примеры). Исследования сосредоточены в том числе на сокращении данных, человеческих и вычислительных ресурсов, необходимых для обучения и инференса различных моделей. В рамках данного материала мы поделимся полученными результатами в рамках исследования и разработки некоторых мультимодальных и мультизадачных архитектур.

Ключевые слова: мультимодальность, мультизадачность, компьютерное зрение, обработка естественного языка, нейронные сети, трансформеры, фундаментальные модели, FusionBrain

DOI: 10.31857/S2686954322070244

Информация, обрабатываемая мозгом человека в каждый момент времени и необходимая для принятия даже самых простых повседневных решений, имеет разную природу и представлена в самом разном виде (по-другому – представлена в разных модальностях). Для восприятия такой разнородной информации человек использует свои органы чувств, а для ее анализа – специальные зоны головного мозга (и, как следствие, специализированные знания, полученные в течение жизни): так, визуальная информация требует зрительного восприятия, слуховая информация предполагает восприятие и анализ звука, обработка текстов на естественном языке предполагает знание языка, и так далее. Почти всегда при этом для успешного решения возникающих задач и проблем в реальном мире необходимо использовать одновременно информацию, поступающую из разных модальностей, так как сами задачи по своей природе вовлекают несколько таких модальностей (например, вождение автомобиля, просмотр фильма, ответы на разные вопросы, и так далее).

Тем не менее в науке о данных и машинном обучении исторически сложилось так, что изучению и способам обработки каждой из основных модальностей посвящены отдельные области, часто не сильно пересекающиеся: например, в рамках CV (computer vision или компьютерного зрения) разрабатываются модели, которые решают задачи, связанные с анализом изображений, 3D-объектов или видео, в рамках NLP (natural language processing или обработки естественного языка) изучаются архитектуры, которые умеют работать с текстовыми данными на разных языках, в рамках PLP (program language processing) – с кодом на разных языках программирования, от-

[1] *ПАО Сбербанк, Москва, Россия*
[2] *AIRI, Москва, Россия*
*****E-mail: Dimitrov.D.V@sberbank.ru*

дельно стоят модели, работающие с временными рядами разной природы и табличными данными. Из-за этого разрабатываемые модели (особенно те, которые работают и используются в реальных бизнес-процессах) в большинстве своем умеют работать строго с одним типом данных и решать ровно одну узкоспециализированную задачу, на которую и были обучены.

Но последние несколько лет все больше исследований ведется в области разработки мультимодальных и мультизадачных архитектур. Помимо того, что это современный тренд, это еще и большой научный и инженерный вызов, еще один шаг на пути к созданию сильного искусственного интеллекта. В настоящее время ведется большее количество исследований в области мультимодальных и мультизадачных моделей. Все разработки ведутся в нескольких направлениях, исследуются либо трансформеры с архитектурами энкодер-декодер [1, 2], либо только декодерные трансформеры [3]. В своих исследованиях авторы экспериментально подбирают задачи, которые используются на претрейне, чтобы затем обученная мультимодальная модель могла решать множество задач в режиме zero-shot.

Мы предлагаем модель RUDOLPH – мультизадачную модель-декодер, способную решать ряд задач на стыке двух модальностей: текст и изображение. На претрейне RUDOLPH обучался на 3 типах задач – text2image, image2text и text2text. На вход в модель подается следующая последовательность токенов: текстовые, картиночные, текстовые. Такая комбинация позволяет обучать текстово-визуальные и визуально-текстовые задачи. Наряду с текстовыми и визуальными токенами, используемыми в трансформере, вводятся спецтокены, отражающие конкретную задачу на обучение. Эти спецтокены явно подсказывают модели, какая конкретно задача пришла на обучение в текущий момент. Благодаря такому обучению, на инференсе модель способна сама определить задачу, при этом качество генерации становится выше. Существует три версии модели RUDOLPH: 350M, 1.3B, 2.7B.

Мы стремимся способствовать развитию такой перспективной и сложной области, как мультимодальные исследования, и проводим соревнование Fusion Brain Challenge 2.0. В рамках данной задачи предлагается построить единую multitask-модель, которая бы успешно решала подзадачи в двух модальностях (визуальной и текстовой), принимая на вход описания подзадач, выраженные на естественном русском языке, например: “сгенерируй изображение”, “опиши изображение”, “ответь на вопрос” и т.д. В состав входит 12 подзадач, из которых 6 известны участникам с момента начала Конкурса (открытые подзадачи), а 6 неизвестны (скрытые подзадачи) и представляют собой частные случаи открытых задач (имеют некоторые отличительные особенности в постановке). Основная задача участников заключается в построении и обучении единой мультимодальной мультизадачной архитектуры, которая позволила бы получить максимальные значения метрик для каждой отдельной подзадачи и, как следствие, достичь максимального значения интегральной метрики на 12 подзадачах.

К открытым подзадачам относятся Text QA, Mathematical QA, Image Generation, Image Captioning, Visual QA, Text Recognition in the Wild. Подзадача Text QA – задание на понимание прочитанного текста. Для успешного решения модель должна уметь устанавливать причинно-следственные связи, разрешать кореференции, а также определять правильную последовательность действий, учитывая временную информацию. Подзадача Mathematical QA проверяет способность модели выполнять простейшие арифметические действия, необходимые для решения линейных уравнений или систем линейных уравнений, а также производить операции сравнения. Подзадача Image Generation подразумевает генерацию изображений на основе текстовых описаний на русском языке. Ответом на подзадачу является изображение, чье содержание соответствует входному текстовому описанию. Подзадача Image Captioning подразумевает генерацию текстовых описаний на русском языке к изображениям. Ответом на подзадачу является текстовая строка, содержащая текстовое описание входного изображения. Подзадача Visual QA предполагает, что обученная модель способна формировать ответ на вопрос по изображению. В этой подзадаче на вход модели подается пара вида “текстовый вопрос – картинка”, а выходом является соответствующий текстовый ответ. Подзадача Text Recognition in the Wild – задание на распознавание текста в городской или иной подобной местности (вывески, дорожные знаки, рекламные объявления и т.п.). Данные представляют собой фотографии объектов с изображенным на них текстом. Ответом на подзадачу является текстовая строка. Скрытые задачи относятся к этим же модальностям и позволяют оценить обобщающую способность модели. То есть модель, обученная на открытых задачах, должна использовать свои знания в решение скрытых задач.

Baseline решение для соревнования FusionBrain основано на модели RUDOLPH. В качестве базовой модели мы использовали модель RUDOLPH 2.7B, дообученную для решения шести открытых задач FBC2.

В заключение отметим, что наш вклад заключался в следующем – были подготовлены данные

как тестовые, так и для обучения, была определена постановка задачи и подготовлена платформа для соревнования Fusion Brain Challenge 2.0. Также был разработан baseline, обученный на 6 открытых задачах и сочетающий мультимодальный и мультизадачный подход. Помимо этого, были разработаны специализированные метрики под каждую задачу и общая метрика для оценки моделей.

СПИСОК ЛИТЕРАТУРЫ

1. *Wang P. et al.* Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework//CoRR. 2022.
2. *Wang W. et al.* Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks//arXiv preprint arXiv:2208.10442. 2022.
3. *Reed S. et al.* A Generalist Agent // arXiv preprint arXiv:2205.06175. 2022.