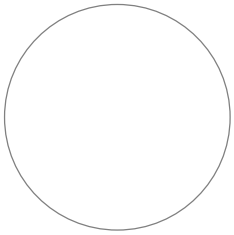
AIRI

2025

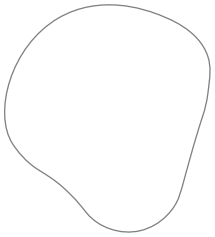
2024

Годовой отчёт



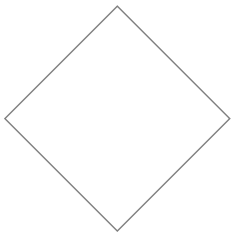
Природа

Постоянное движение и развитие, закономерность и системность, законы и ритмы, хаос и жизнь



Человек

Живой организм, часть природы, способный мыслить, чувствовать и говорить, несовершенный



Технологии

Творение человека и его продолжение, логичное и предсказуемое, утилитарное, создано помогать

AIRI

Единство природы, человека и технологий

Содержание

Четвертый год жизни AIRI

Иван Оселедец

Генеральный директор

В 2024 году наш Институт увеличился вдвое и насчитывает более 200 преданных своему делу специалистов, 9 из которых в течение года защитили кандидатские и докторские диссертации. Многие приехали к нам из-за границы, чтобы развивать науку об ИИ в России и мире с AIRI.

Вместе с тем, увеличилось и количество успешно проведенных исследований, 90 из которых было представлено на конференциях уровня A и A*. 17 статей — принято на ведущую международную конференцию по искусственному интеллекту NeurIPS.

Мы в AIRI продолжаем создавать универсальные системы ИИ для решения задач реального мира. В этому году мы сфокусировались на концепции прикладного AGI, в основе которой лежит научная обоснованность и польза для бизнеса. Совместно с индустриальными партнерами мы запустили ряд амбициозных проектов по генеративному проектированию и применению технологий искусственного интеллекта в медицине.

Появляются новые коллаборации с ведущими университетами и технологическими компаниями. К нам присоединились партнеры из ГИАП, СберМедИИ, ИСП РАН, KAMAZ Digital, Сибур Digital и многие другие. Обширная и вдохновляющая часть 2024 года — знакомство и сотрудничество с региональными научно-исследовательскими центрами. СВФУ, АГУ, ТГУ — всех не перечесть — демонстрируют, насколько разнообразные и сильные компетенции представлены в России. Я уверен, что благодаря стремлению AIRI соединять между собой коллег из самых разных точек мира и предметных областей, в будущем нас ждет очень много непредсказуемых, но приятных результатов.

Я выражаю благодарность сотрудникам AIRI и нашим партнерам за усердную работу. Вместе мы не только приближаем появление AGI, но и участвуем в подготовке нового поколения исследователей и вдохновляем друг друга на открытия.

Приглашаю вас узнать больше о результатах работы Института AIRI в 2024 году из этого отчета.

Миссия AIRI

Создание универсальных систем искусственного интеллекта, решающих задачи реального мира

Основная цель Института — найти возможности применения искусственного интеллекта для решения сложных социальных, экономических и научных задач. Научные сотрудники Института занимаются исследованиями в фундаментальных и прикладных областях. В работе над своими проектами они стремятся получить прорывные результаты в области искусственного интеллекта и его приложений, участвуя в формировании глобальной исследовательской повестки.

Ценности AIRI

Человекоцентричность

Свобода научного выражения

Ответственность и вклад в будущее

Открытость и прозрачность

Партнерство и коллаборация

Направления деятельности

Научные исследования

Проведение прорывных исследований в области искусственного интеллекта, формирование глобального центра экспертизы

Вклад в развитие AI

Участие в глобальном развитии искусственного интеллекта через создание, развитие и поддержку проектов с открытым кодом

Научно-технические партнерства

Развитие партнерств с научными организациями, промышленностью и государством, разработка и коммерциализация технологий в области искусственного интеллекта

Лаборатории

Сотрудничество с институтами, университетами и индустриальными партнерами по запуску совместных научно-исследовательских лабораторий в области искусственного интеллекта

Популяризация AI

Проведение профильных конференций и мероприятий, создание и поддержка соревнований, популяризация искусственного интеллекта в целом

Ключевые результаты

221
публикаций

65
статей
на конференциях A*

25
статей
на конференциях A

56
статей в журналах
с рейтингом Q1

Руководители научных групп

Андрей Кузнецов

Лаборатория FusionBrain

Владислав Шахуро

ИИ для роботов

Айбек Аланов

Контролируемый Генеративный ИИ

Константин Соболев

Генеративный ИИ для видео

Александр Панов

Лаборатория когнитивных систем искусственного интеллекта

Алексей Ковалёв

Воплощенные агенты

Юрий Куратов

Модели с памятью

Алексей Скрынник

RL агенты

Дмитрий Дылов

Лаборатория «Сильный AI в медицине»

Дмитрий Умеренков

Базовые модели

Ярослав Беспалов

Мультимодальные архитектуры ИИ

Назар Бузун

Обучение представлений

Дмитрий Крюков

Исследования биомаркеров

Руководители научных групп

Иван Оселедец

Вычислительный интеллект

Евгений Бурнаев

Обучаемый интеллект

Антон Конушин

Пространственный интеллект

Егор Ершов

Цветовая вычислительная фотография

Ольга Кардымон

Биоинформатика

Марина Мунхоева

Самообучение и обучение представлений

Илья Макаров

ИИ в индустрии

Александр Панченко

Вычислительная семантика

Александр Тюрин

Методы оптимизации в машинном обучении

Евгений Фролов

Технологии персонализации

Александр Горбань

Лаборатория ИИ, анализа данных и моделирования

Олег Рогов

Доверенные и безопасные интеллектуальные системы

Семен Буденный

Дизайн новых материалов

Елена Тутубалина

Прикладное NLP

Алексей Осадчий

Нейроинтерфейсы

Артур Кадурин

Глубокое обучение в науках о жизни

Владислав Куренков

Адаптивные агенты

Артем Шелманов

Обучение на слабо размеченных данных

Руководители функциональных направлений

Иван Оселедец

Генеральный директор

Максим Кузнецов

Директор управления стратегического развития и партнерств

Ольга Суровегина

Директор по научно-техническим партнерствам

Степан Мамонтов

Руководитель отдела научно-технической разработки

Ольга Попова

Руководитель отдела проектного управления

Антон Ризаев

Финансовый директор

Екатерина Сарафанова

Главный бухгалтер

Николь-Мария Кук

Руководитель отдела закупок

Александра Бройтман

Директор по маркетингу и коммуникациям

Мария Звонарева

HR Директор

Юлия Никитина

Директор по правовому обеспечению

Константин Катанов

IT Директор

Научные результаты

На пути к прикладному AGI

AGI =

Инженерный подход

- → Научная обоснованность: Публикации на A/A* и Q1
- → Трансфер идей в продуктовый пайплайн / тестирование в бизнесе

+

Польза для текущего бизнеса

- → Функциональный заказчик на проведение научного исследования
- → Проведенный на стороне заказчика предварительный анализ бизнес-эффекта и prerequisite research

Направления прикладного AGI

Compute and Data	Multimodality	Agency	Embodiment	World Model
Effective training & inference	Modality	Self-learning	Hardware	Knowledge
Effective data storage	Omnimodality	Self-reflection/Reasoning	Software	Forecasting
Synthetic data	Perception Augmentation	Self-alignment	Interface	Causality
		Goal-setting & planning		
		Multiagency		

Главный результат

Дмитрий Дылов
Директор лаборатории «Сильный AI в медицине»

AGI Med

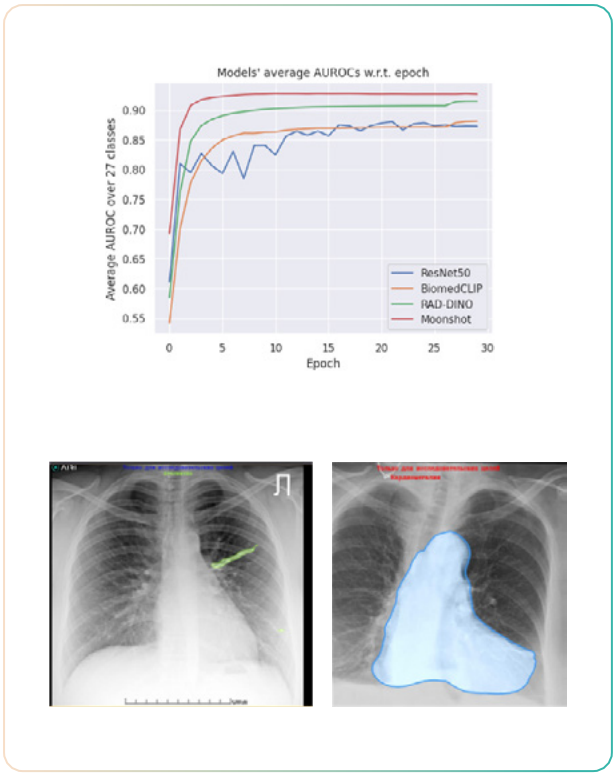
«Цифровой Помощник 2.0» и «Муншот», предназначенные для первичного приема пациентов и для содействия врачам лучевой диагностики

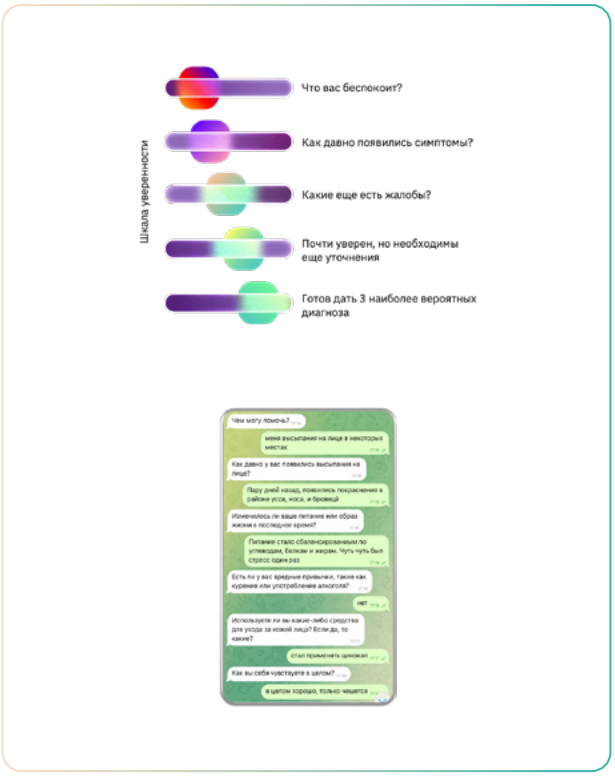
В мобильном приложении медицинской компании «СберЗдоровье» (входит в индустрию здоровья Сбербанка) появился ИИ-помощник, который поможет россиянам заботиться о здоровье. Технологию разработали ученые Института AIRI и компания «СберМедИИ» на базе нейросетевой модели Сбербанка GigaChat.

- → Постановка предварительного диагноза и маршрутизация к целевому специалисту
- → Расшифровка анализов и заключения врача
- → Реализован набор интеллектуальных фильтров безопасности

В то же время, базовая модель «Муншот» показала себя как мощная модель анализа радиологических снимков. Будучи обучена на крупном датасете Москвы, модель научилась генерировать разделы «Находки» и «Заключение» в радиологических отчетах по рентгенографии грудной клетки, предсказывать класс патологии и автоматически контурировать их. Этот метод объединяет визуальный энкодер (контрастное обучение), обученный на медицинских изображениях, и специализированную биомедицинскую языковую модель. Подход эффективно преодолевает сложности, связанные с медицинской семантикой, и учитывает разнообразие находок в изображениях, обеспечивая автоматизацию процесса составления отчетов.

Модель обучена на крупных наборах данных, таких как PadChest, BIMCV-COVID19, CheXpert, OpenI и MIMIC-CXR, что обеспечивает ее устойчивость и адаптацию к разнообразным клиническим случаям. Оценочные метрики подчеркивают точность и способность модели учитывать тонкие медицинские детали, что демонстрирует большой потенциал для оптимизации работы радиологов. Это достижение не только автоматизирует важный





этап диагностики, но и способствует внедрению передовых ИИ-решений в медицинскую практику.

В рамках главной конференции по мед ИИ (MICCAI 2024), одним из ключевых достижений стало создание нового подхода к интерактивной сегментации сложных удлиненных объектов, таких как провода, катетеры или вены, без использования традиционных масок. Разработанная структура данных ориентируется на набор центральных линий контурируемых объектов и работает особенно эффктивно. Также значительные усилия были направлены на создание новых подходов к анализу медицинских изображений. В частности, предложены анатомические позиционные эмбеддинги для 3D-изображений, способные точно предсказывать анатомическое положение областей, что открывает новые возможности для локализации органов и патологий. Эти результаты были представлены на MICCAI 2024, 2 из 3 российских статей на конференции были сделаны сотрудниками лаборатории AGI Med.

Команда AIRI-AGI-med приняла участие в двух международных соревнованиях: RRG24 на BioNLP и MIDRC XAI Challenge, где заняла второе и пятое место, соответственно, используя первое поколение своей модели Муншот. Это стало возможным благодаря разработке методов классификации затемнений и сегментации легких на датасете с ограниченной разметкой. На основе CLIP-модели удалось выиграть приз в $5000, а также показать отличные результаты в задачах классификации и сегментации. Команда уступила только ведущим IT-гигантам, будучи на порядок меньше по составу.

Крупные результаты

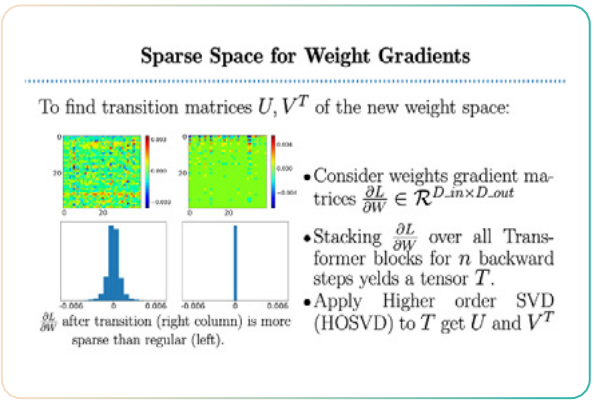
Разработан метод SparseGrad для параметро-эффективного обучения больших языковых моделей на базе тензорного разложения HOSVD

Александр Панченко
Ведущий научный сотрудник

Иван Оселедец
Генеральный директор

В методе SparseGrad с помощью HOSVD-декомпозиции мы находим пространство, где матрица весов линейных слоев сильно разрежена (остается ~1% от всех параметров). Создав новый класс линейного слоя и переписав автоград с учетом перехода в это пространство, авторы уменьшили количество параметров, используемых при тренировке Трансформеров, и, следовательно, используемую память. При одинаковом потреблении памяти на энкодерных архитектурах (BERT, RoBERTa) и декодерных архитектурах (LLaMa2 -7B) модель показывает результаты лучше LoRA и MeProp — бейзлайна, который отбирает наиболее значимые параметры в линейном слое и тренируют только их.

Тестирование подхода на LLaMa 2 7B показало, что наш метод дает лучший лосс по валидации, а также лучшую метрику на вопросно-ответном бенчмарке I-Bench. Качественная характеристика генерируемого текста подтвердила этот вывод.



SparseGrad: A Selective Method for Efficient Fine-tuning of MLP Layers
Viktoriia Chekalina, Anna Rudenko, Gleb Mezentsev, Alexander Mikhalev, Alexander Panchenko, Ivan Oseledets
EMNLP'24, A*

Разработан метод AMORE для оценки надежности и интерпретируемости больших языковых моделей в химии, основанный на аугментации молекулярных структур

Елена Тутубалина
Ведущий научный сотрудник

Артур Кадурин
Инженер-исследователь

Иван Оселедец
Генеральный директор

Андрей Кузнецов
Директор лаборатории

Денис Димитров
Научный консультант

Недавняя интеграция химии с обработкой естественного языка (NLP) продвинула вперед открытие лекарств. Сегодня представление молекул в языковых моделях (LM) имеет решающее значение для улучшения понимания химических процессов. Разработана система дополненного молекулярного поиска (Augmented MOlecular REtrieval, AMORE), гибкая zero-shot система оценки достоверности химических языковых моделей различной природы: обученных исключительно на молекулах для выполнения химических задач и на объединенном корпусе текстов на естественном языке и структур, представленных в виде строк. Эта система опирается на модификации молекул, сохраняющие их химические свойства, такие как кекулизация и замена циклов. Метрика основана на оценке сходства между распределенными представлениями молекул и их модификациями.

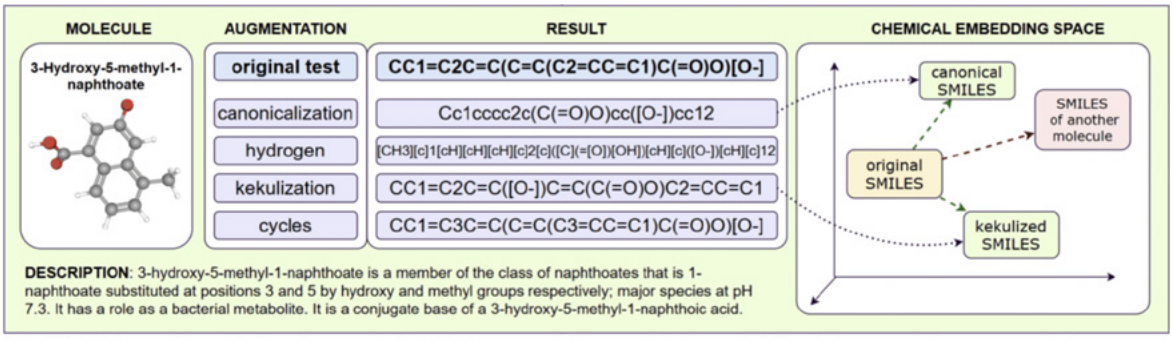


Fig.1 AMORE encodes original and augmented SMILES representations, calculates embedding distances, and assesses model performance based on top-1 accuracy, where the correct augmented SMILES is retrieved first

A*: Ganeeva, V., Sakhovskiy, A., Khrabrov, K., Savchenko, A., Kadurin, A. & Tutubalina, E. Lost in Translation: Chemical Language Models and the Misunderstanding of Molecule Structures. In Findings of the Association for Computational Linguistics: EMNLP 2024

A: The Shape of Learning: Anisotropy and Intrinsic Dimensions in Transformer-Based Models", Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, Andrey Kuznetsov, EACL 2024

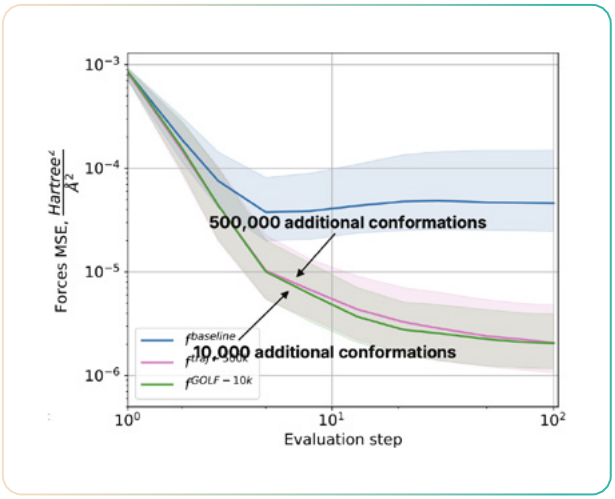
Разработан фреймворк GOLF для активного обучения нейронных потенциалов, ориентированный на локальную оптимизацию лекарственных молекул

Александр Панов
Директор лаборатории

Елена Тутубалина
Ведущий научный сотрудник

Артур Кадурин
Инженер-исследователь

Разработанный подход GOLF позволил в 50 раз сократить объем данных, необходимый для дообучения нейронных сетей на задачу оптимизации конформаций (пространственных представлений молекул). Используем суррогатный оракул, чтобы отобрать конформации для последующей разметки с помощью DFT. Оцениваем, уменьшилась ли энергия. Если энергия уменьшилась, продолжаем оптимизацию или рассматриваем предсказание NNP как некорректное и добавляем предыдущую конформацию в обучающую выборку.



A*: Tsypin, A., Ugadiarov, L. A., Khrabrov, K., Telepov, A., Rumiantsev, E., Skrynnik, A., Panov, A., Vetrov, D., Tutubalina, E., Kadurin, A. Gradual Optimization Learning for Conformational Energy Minimization. In The Twelfth International Conference on Learning Representations (ICLR).

A*: Tsypin, A., Ugadiarov, L. A., Khrabrov, K., Telepov, A., Rumiantsev, E., Skrynnik, A., Panov, A., Vetrov, D., Tutubalina, E., Kadurin, A. Gradual Optimization Learning for Conformational Energy Minimization. In The Twelfth International Conference on Learning Representations (ICLR).

Исследование внутренних свойств моделей трансформеров (внутренняя размерность и анизотропия), характеризующих наличие линейной зависимости ряда соседних слоёв эмбеддингов

Иван Оселедец
Генеральный директор

Денис Димитров
Научный консультант

Андрей Кузнецов
Директор лаборатории

В данной работе было продолжено исследование моделей-декодеров и выявлена линейная зависимость между последовательными слоями моделей, основанных на архитектуре Трансформер. Был предложен процесс прунинга моделей, который позволяет заменить часть слоев декодеров на линейное преобразование входного вектора. Замеры на бенчмарках показали, что замена 5-10% слоев на линейное преобразование не оказывает сильное влияние на результаты модели.

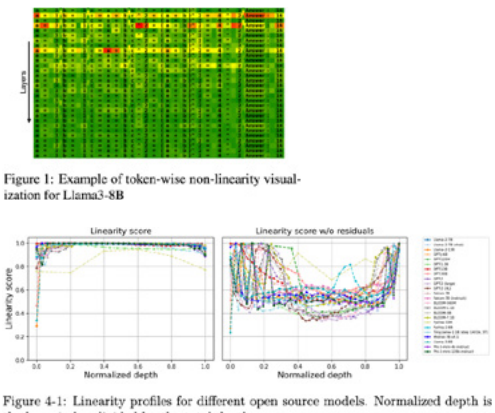


Figure 1: Example of token-wise non-linearity visualization for Llama3-8B

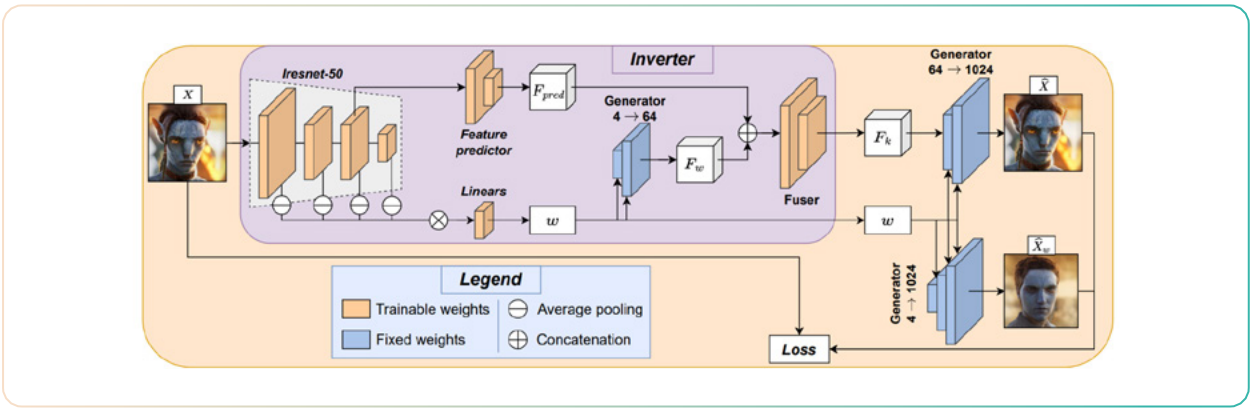
Figure 4-1: Linearity profiles for different open source models. Normalized depth is the layer index divided by the total depth.

A*: Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Nikolai Gerasimenko, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. 2024. Your Transformer is Secretly Linear. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)

Контролируемый генеративный ИИ

Айбек Аланов
Научный сотрудник

- Применение модели StyleGAN для манипуляции изображений. Мы разработали модель StyleFeatureEditor, которая позволяет редактировать изображения, при этом восстанавливая их с очень высокой точностью. Мы улучшили на тот момент sota подход в 4 раза по метрике LPIPS точность реконструкции[1].
- Разработка модели для переноса причесок для лиц людей. Мы предложили новую модель, которая показывает наилучшее качество переноса причесок и реалистичность результата по сравнению с бейзлайнами. Скорость применения на порядке быстрее[2].
- Разработка эффективной параметризации для дообучения нейросетевых моделей. Был предложен метод, который позволяет более эффективно с точки зрения используемой памяти дообучать нейросетевые модели, в том числе большие диффузионные модели[3].



[1] A*: "The Devil is in the Details: StyleFeatureEditor for Detail-Rich StyleGAN Inversion and High Quality Image Editing" на конференции CVPR 2024.
Авторы: Денис Бобков, Вадим Титов, Айбек Аланов, Дмитрий Ветров

[2] A*: "HairFastGAN: Realistic and Robust Hair Transfer with a Fast Encoder-Based Approach" на конференции NeurIPS 2024. Авторы: Максим Николаев, Михаил Кузнецов, Дмитрий Ветров, Айбек Аланов

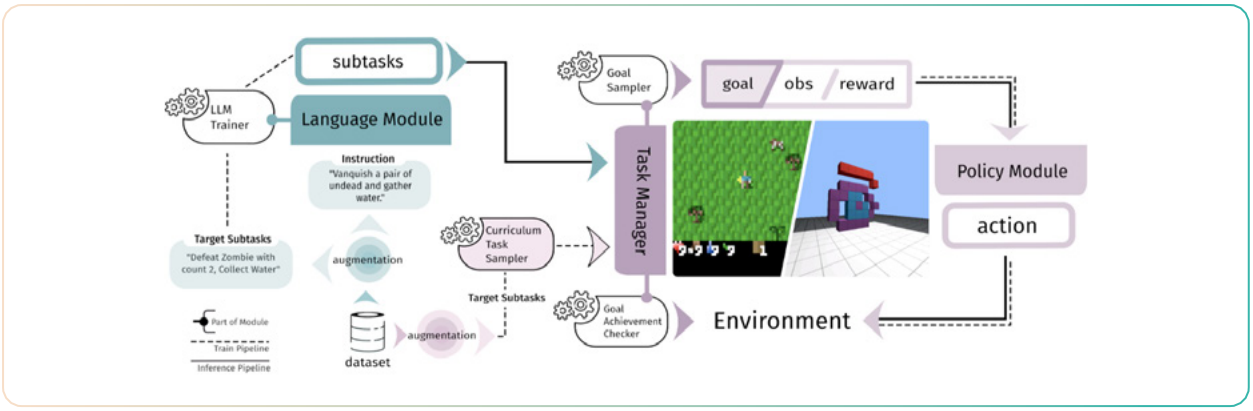
[3] A*: "Group and Shuffle: Efficient Structured Orthogonal Parametrization" на конференции NeurIPS 2024.
Авторы: Михаил Горбунов, Николай Юдин, Вера Соболева, Айбек Аланов, Алексей Наумов, Максим Рахуба

Разработана оригинальная нейросимвольная архитектура управления воплощенными агентами в сложных динамических средах с поддержкой планирования на основе языковых моделей и двумя типами модели мира для эффективного обучения в среде

Александр Панов
Директор лаборатории

Результаты работы изложены в докторской диссертации, защищенной на диссовете МФТИ по специальности 1.2.1 «Искусственный интеллект и машинное обучение».

Подход IGOR для построения стратегии в мультимодальных средах. Метод использует большую языковую модель для преобразования текстовых инструкций в высокоуровневый план, который затем воплощает RL-агент, выстраивая стратегию поведения в среде. Эффективность метода продемонстрирована в среде IGLU, где предложенный подход превзошел результаты первого места на соревновании NeurIPS, и в среде Crafter, опережая SOTA-решение Dynalang.



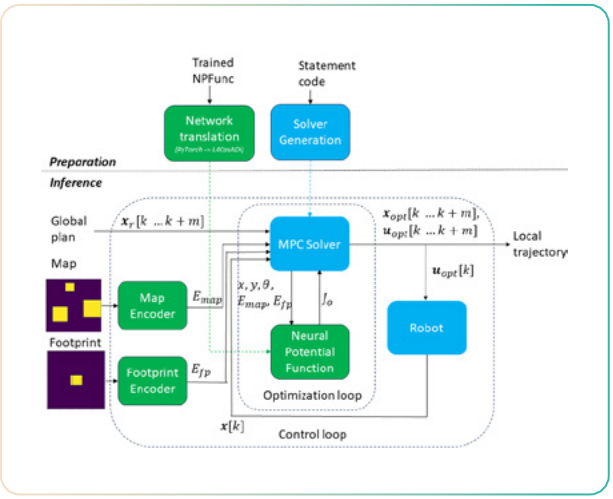
A: Volovikova, Z., Skrynnik, A., Kuderov, P. and Panov, A.I., 2024. Instruction Following with Goal-Conditioned Reinforcement Learning in Virtual Environments. In ECAI 2024 (pp. 650-657). IOS Press. [Core A]

Разработана модель для задачи локального планирования с учетом динамических препятствий

Александр Панов

Директор лаборатории

Разработана модель, вычисляющая нейронный потенциал, используемый в MPC для задачи локального планирования пути с учетом динамических препятствий. Используется трансформерная модель на основе GPT. Построенная система управления была протестирована на реальном мобильном роботе и показала преимущество по сравнению со всеми мировыми аналогами. Работа была представлена на ведущей конференции по робототехнике ICRA 2024.



A*: Alhaddad, M., Mironov, K., Staroverov, A., Panov, A. Neural Potential Field for Obstacle-Aware Local Motion Planning, IEEE International Conference on Robotics and Automation (ICRA) 2024, pp. 9313–9320.

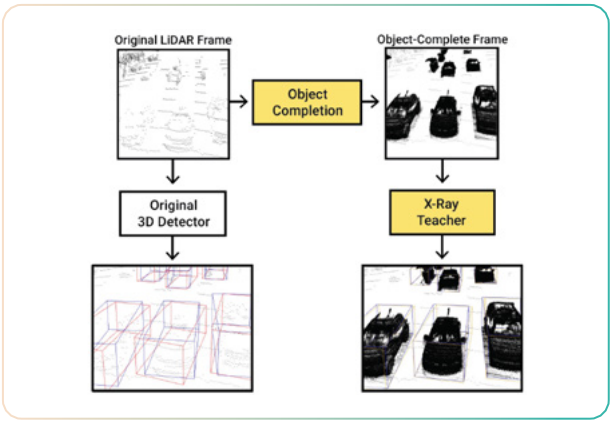
Разработан подход, объединяющий последовательности кадров LIDAR через алгоритмы регистрации облаков точек, достигающий SOTA-результатов в semi-supervised и улучшение в supervised сеттинге

CVPR 2024 (A*), IEEE Access (Q1), demo IJCAI 2024 (A*)

Илья Макаров
Ведущий научный сотрудник

Подход существенно улучшает точность и эффективность моделей 3D детекции, что критично для автономного транспорта, робототехники и других приложений, зависящих от высокоточного распознавания объектов.

Технология может быть использована в автономных транспортных системах, складской автоматизации и робототехнике.



Gambashidze, A., Dadukin, A., Golyadkin, M., Razzhivina, M., & Makarov, I. Weak-to-Strong 3D Object Detection with X-Ray Distillation. CVPR 2024

Исследование стратегических и этических решений LLM

Илья Макаров
Ведущий научный сотрудник

Разработан фреймворк для тестирования гипотез о соответствии эмоциональных реакций людей и решений LLM. Установлено, что эмоциональный алайнмент с людьми в большинстве случаев слабый, но эмоциональные промпты, особенно негативные, существенно влияют на поведение моделей, снижая кооперацию в играх и качество в этических задачах. NeurIPS (A*)

Понимание эмоционального воздействия на модели критично для разработки безопасных и надежных ИИ-систем в этических и стратегических приложениях.

Использование фреймворка возможно для оптимизации LLM в областях, где важно учитывать эмоциональный контекст, например, ассистентах или интерактивных системах.

Разработка эмпатичного LLM-ассистента:

Разработан подход InsideOut, где базовые эмоции по Экману участвуют в формулировании ответов. На базе GigaChat и GPT-3.5 достигнуто улучшение на 12–20% в распознавании эмоций и до 3.9% в качестве эмпатичных ответов.

ECAI demo (A)

Повышение эмпатии в LLM способствует созданию более человечных и эффективных ассистентов, полезных в психологии, обучении и клиентской поддержке. Подход может быть интегрирован в ассистентов для медицинских консультаций, обучения и сервисов поддержки клиентов.

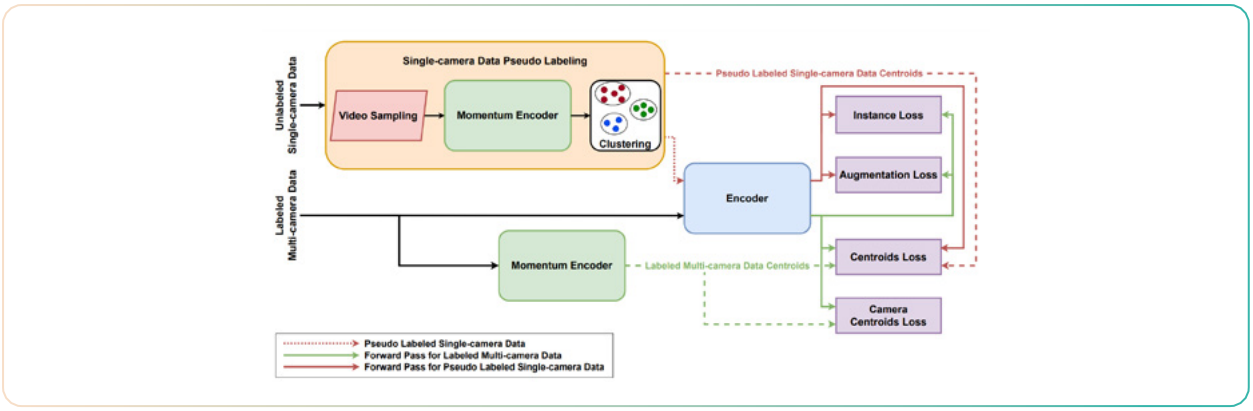
Mozikov, M., Severin, N., Bodishtianu, V., Glushanina, M., Nasonov, I., Orekhov, D., Pekhotin, V., Makovetskiy, I., Baklashkin, M., Lavrentyev, V., Tsvigun, A., Turdakov, D., Shavrina, T., Savchenko, A., & Makarov, I. EAI: Emotional Decision-Making of LLMs in Strategic Games and Ethical Dilemmas. NeurIPS 2024

Deeb, B. M., Savchenko, A., & Makarov, I. (2024). CA-SER: Cross-Attention Feature Fusion for Speech Emotion Recognition. In ECAI 2024 (pp. 4479-4482). IOS Press. ECAI 2024 Demo.

Умное обучение на смеси данных для распознавания людей

Антон Конушин
Старший научный сотрудник

Разработаны методы обучения моделей реидентификации человека на смесях неразмеченных однокамерных и размеченных многокамерных данных, позволившие существенно повысить обобщающую способность моделей, валидированную тестированием в кросс-датасет сценариях.



Mamedov T., Konushin V., Konushin A. ReMix: Training Generalized Person Re-identification on a Mixture of Data // arXiv preprint arXiv:2410.21938. — 2024 (принята на WACV 2025)

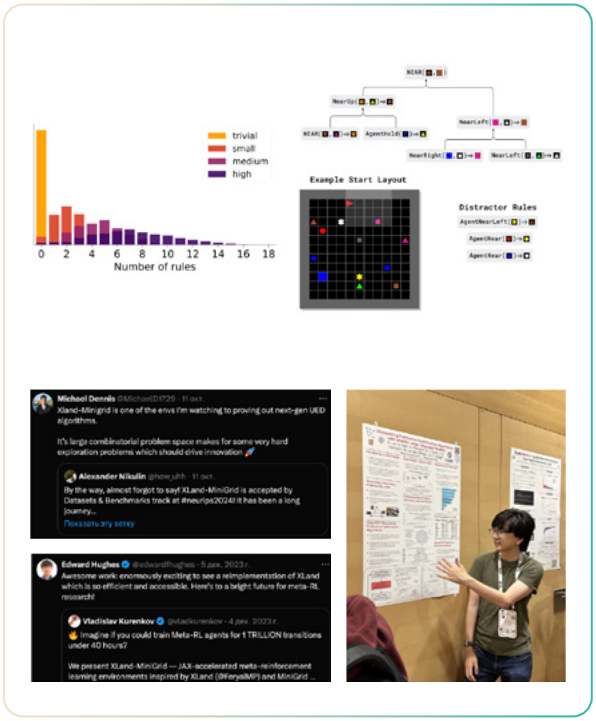
Первая открытая среда и самый большой датасет в мире для In-Context Reinforcement Learning

Владислав Куренков

Инженер-исследователь

Разработанная среда XLand-MiniGrid для контекстного обучения значительно ускорила скорость экспериментов, сделав область более доступной. За счет переноса вычислений среды на GPU обучение агентов для экспериментов и сбора данных ускорилось в десятки раз, с недель до минут. По сравнению с существующими аналогами, среда также предлагает большее разнообразие — миллионы уникальных задач разной сложности, пусть и абстрактных. Среда получила признание в сообществе и одобрение от множества исследователей, в том числе из Google DeepMind.

На основе среды был собран датасет XLand-100B, самый большой из существующих в области. Он позволит исследователям детальнее изучить scaling laws текущих и будущих подходов. Более того, текущие подходы в целом не справляются с решением представленных в датасете задач, поэтому он послужит хорошим бенчмарком и путеводной звездой в области.

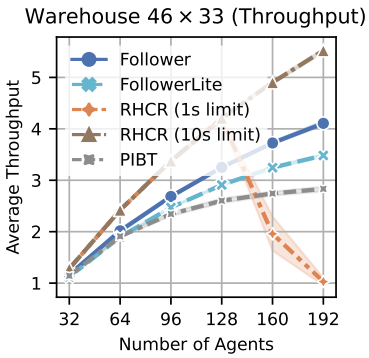


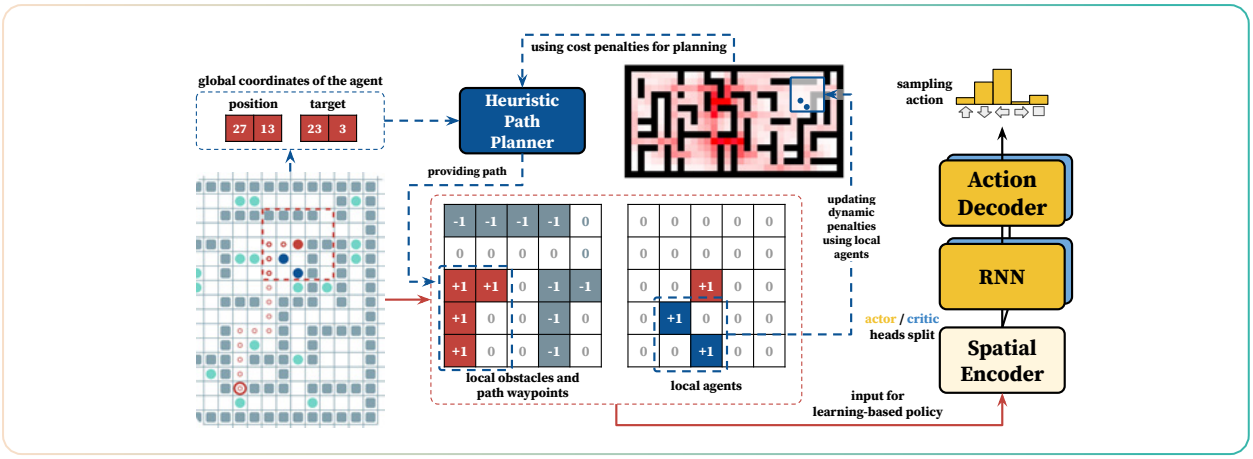
A*: XLand-MiniGrid: Scalable Meta-Reinforcement Learning Environments in JAX, NeurIPS 2024

Децентрализованные методы решения задачи многоагентного планирования

Александр Панов
Директор лаборатории

Предложены новые методы решения задачи много-агентного планирования траекторий в децентрализованной постановке, опирающихся на интеграцию классические техник решения этой задачи (эвристический поиск) с современными обучаемыми методами (model-based reinforcement learning, large-scale imitation learning with transformers и др.). Совместная работа с А.А. Андрейчуком, А.А. Скрынником, М. Нестеровой, К.С. Яковлевым.





Skrynnik A., Andreychuk A., Yakovlev K., Panov A. Decentralized Monte Carlo Tree Search for Partially Observable Multi-agent Pathfinding // In Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI 2024). pp. 17531-17540.

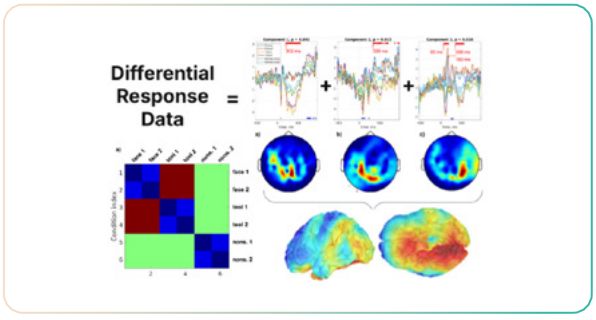
Skrynnik A, Andreychuk A, Nesterova M, Yakovlev K, Panov A. Learn to Follow: Decentralized Lifelong Multi-Agent Pathfinding via Planning and Learning. InProceedings of the AAAI Conference on Artificial Intelligence 2024 Mar 24 (Vol. 38, No. 16, pp. 17541-17549).

ReDisCA — метод быстрого анализа репрезентативного сходства по ЭЭГ и МЭГ данным активности головного мозга

Алексей Осадчий
Ведущий научный сотрудник

Поиск зон мозга, ответ на внешние стимулы в которых обладает заданной репрезентативной (геометрической) структурой. Например, реакции на изображения инструментов и изображения лиц диаметрально противоположны и лежат по «разные стороны» от реакции на бессмысленные зрительные стимулы.

Завершение разработки метода анализа компонентов репрезентативного сходства (ReDisCA) и применение его к анализу данных МЭГ и ЭЭГ с целью поиска нейрональных источников с желаемой репрезентативной структурой.
В отличие от традиционного анализа репрезентативного сходства (RSA), который требует исчерпывающего поиска по сетке нейронных источников, ReDisCA использует приблизительную стратегию оптимизации в закрытой форме для разложения данных ЭЭГ/МЭГ на пространственно-временные компоненты с заданными репрезентативными свойствами.
Это нововведение позволяет избежать проблемы множественных сравнений и значительно улучшает обнаруживаемость и локализацию источников, связанных с заданными пользователем профилями репрезентативного сходства.
ReDisCA предоставляет точную и интерпретируемую основу для объединения данных визуализации мозга с вычислительными моделями.
В 100+ раз быстрее традиционного RSA, в 2 раза точнее. Работа опубликована в Neuroimage, Q1, Top 1%.



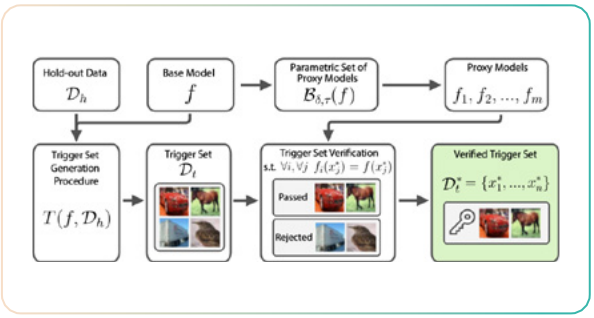
Q1: A. Ossadtchi, I. Semenkov, A. Zhuravleva, O. Serikov, E. Voloshina. Representational dissimilarity component analysis (ReDisCA). NeuroImage, 2024

Вероятностно-устойчивые водяные знаки для нейросетей

Олег Рогов
Старший научный сотрудник

Результат: метод нанесения цифровых водяных знаков, гарантированно устойчивых к наиболее сильным атакам кражи функциональности. Статья на IJCAI-2024 (A*)

С активной эксплуатациией моделей глубокого обучения на продуктовых платформах MLaaS («Машинное обучение как услуга») динамично возрастает и интерес к методам цифровых водяных знаков для таких моделей, которые можно использовать для подтверждения права собственности на определенную модель. К сожалению, ряд существующих методов формирует цифровые водяные знаки, уязвимые для атак, направленных на кражу функционала модели. В нашем решении мы предложили принципиально новый подход к водяным знакам на основе набора триггерных данных, который демонстрирует устойчивость к атакам с кражей функциональности, особенно к тем, которые включают дистилляцию знаний. Подход не требует дополнительного обучения модели и может быть применен к любой архитектуре. Основная идея метода заключается в поиске набора триггеров, который можно переносить между исходной моделью и набором прокси-моделей с высокой вероятностью. Мы в экспериментальном порядке показываем, что если вероятность переносимости набора достаточно высока, его можно эффективно использовать для проверки права собственности на украденную модель. Метод превосходит все аналогичные современные решения цифровой маркировки моделей машинного обучения.

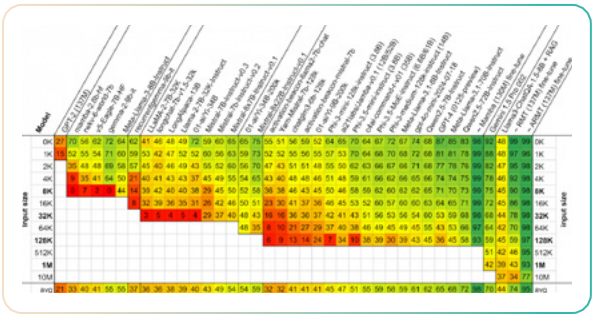


Pautov M, Bogdanov N, Pyatkin S, Rogov O, Oseledets I. Probabilistically Robust Watermarking of Neural Networks. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence 2024 August 3-9, (Vol. 33, pp. 4778-4787)

Новый бенчмарк для оценки реальной производительности языковых моделей при работе с большими объемами данных

Юрий Куратов
Старший научный сотрудник

С помощью разработанного бенчмарка BABILong продемонстрирована реальная эффективность open-source моделей и моделей семейства GigaChat на различных стадиях обучения: short-context pre-train, long-context pre-train, supervised fine-tuning. Показана высокая эффективность этапа SFT на контекстах менее 16к токенов и недостаточная на больших контекстах. Анализ подтвержден на задачах на русском и английском языке. BABILong/RuBABILong являются наиболее репрезентативными бенчмарками по оценке этапов обучения модели по эффективности работы с задачами на больших контекстах.



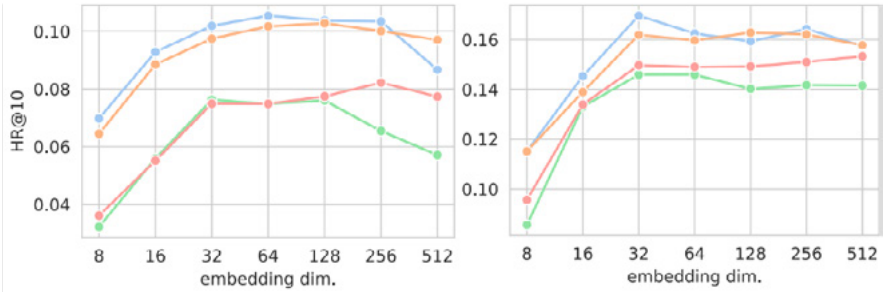
- С увеличением длины контекста качество у всех моделей падает, хотя интенсивность деградации варьируется.
- Большинство моделей эффективно работают лишь с 10–20% от заявленной длины контекста.
- Даже лучшие модели, такие как Gemini 1.5 и Qwen 2.5, показывают падение качества с 80-90% до 60% и меньше на более длинных задачах.
- BABILong демонстрирует, что работа даже с простыми задачами с длинными контекстами — всё ещё сложна даже для лидирующих моделей. Заявленный контекст в сотни тысяч токенов вовсе не означает, что модель способна его воспринять.

A*: BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack
Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Igorevich Sorokin, Artyom Sorokin, Mikhail Burtsev (NeurIPS).

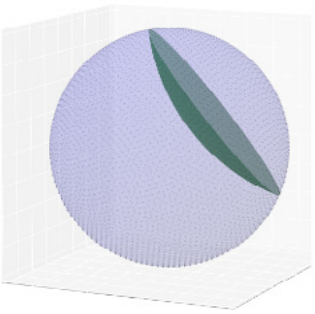
Разработано несколько новых алгоритмов, моделей и архитектур для решения практических задач в рекомендательных системах

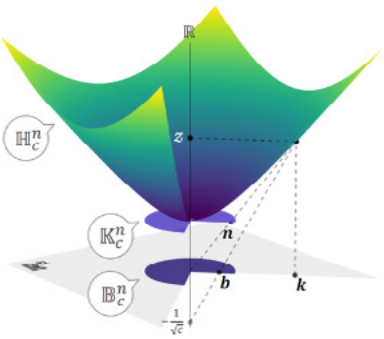
Евгений Фролов
Научный сотрудник

- Новый подход для масштабирования современных SOTA-моделей на огромные каталоги товаров [2,3]. Разработанные нами решения открывают новую веху в возможностях применения ранее зарекомендовавших себя трансформерных архитектур в реальных задачах, поскольку позволяют не только значительно сократить вычислительную нагрузку при обучении моделей, но и повышают качество моделей.
- Новый фреймворк для объединения различных источников поведенческой информации в рамках end-to-end архитектуры для обучения на последовательностях[1]. Универсальность предложенного решения открывает широкие возможности по моделированию гетерогенных источников информации о потребительском поведении в рамках рекомендательных систем. Ведутся работы по усовершенствованию подхода для работы в высокодинамичном окружении.
- Разработан эффективный алгоритм кросс-доменного обучения на основе метода ADMM [4], позволяющий улучшать качество работы рекомендательных систем за счет трансфера знаний между различными доменами. Разработка представляет особый практический интерес при работе в экосистеме маркетплейсов и онлайн-сервисов, т.к. не требует передачи персональной потребительской информации в явном виде, обеспечивая при этом эффективное обучение сразу на нескольких доменах. Ведутся работы по проверке и расширению возможностей алгоритма при обучении на более чем двух доменах.



→ Разработан геометрический подход на основе гиперболической геометрии для обучения на последовательностях действий пользователей[5], позволяющий получить более компактные модели без потери качества рекомендаций. В рамках проекта были впервые разработаны быстрые методы расчета кривизны пространства входных данных, масштабируемые на реальные датасеты. Полученные наработки уже используются в смежных исследованиях, требующих расчетов на больших данных. Планируется опробовать подход в смежной области больших языковых моделей.





[1] A*: Baikalov, Vladimir; Frolov, Evgeny; “End-to-End Graph-Sequential Representation Learning for Accurate Recommendations”, Proceedings of the ACM on Web Conference 2024, 501-504, 2024.

[2] A: Gusak, Danil; Mezentsev, Gleb; Oseledets, Ivan; Frolov, Evgeny; “RECE: Reduced Cross-Entropy Loss for Large-Catalogue Sequential Recommenders”, Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 3772-3776, 2024.

[3] A: Mezentsev, Gleb; Gusak, Danil; Oseledets, Ivan; Frolov, Evgeny; “Scalable Cross-Entropy Loss for Sequential Recommendations with Large Item Catalogs”, Proceedings of the 18th ACM Conference on Recommender Systems, 475-485, 2024.

[4] A: Samra, Abdulaziz; Frolov, Evgeny; Vasilev, Alexey; Grigorevskiy, Alexander; Vakhrushev, Anton; “Cross-Domain Latent Factors Sharing via Implicit Matrix Factorization”, Proceedings of the 18th ACM Conference on Recommender Systems, 309-317, 2024.

[5] A: Frolov, Evgeny; Matveeva, Tatyana; Mirvakhabova, Leyla; Oseledets, Ivan; “Self-Attentive Sequential Recommendations with Hyperbolic Representations”, Proceedings of the 18th ACM Conference on Recommender Systems, 981-986, 2024.

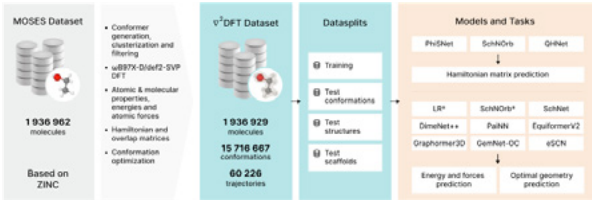
Прикладные результаты

Разработан бенчмарк для оценки качества LLM для русского языка

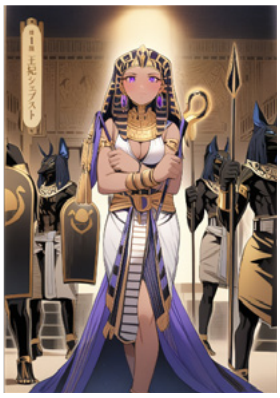
MERA: A Comprehensive LLM Evaluation in Russian. (ACL) Совместно с коллегами из SberDevices и SberAI был подготовлен и опубликован бенчмарк MERA для оценивания больших языковых моделей на задачах, сформулированных на русском языке. Нами были подготовлены задачи MathLogiQA и LogiQA, оценивающие способности моделей решать математические и логические задачи.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, et al.. 2024. MERA: A Comprehensive LLM Evaluation in Russian. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics

DFT: A Universal Quantum Chemistry Dataset of Drug-like Molecules and a Benchmark for Neural Network Potentials (NeurIPS D&B)



Предложен подход для создания парных датасетов, позволяющий впервые создать несинтетический датасет для задачи колоризации манги, что позволило достичь SOTA-результатов. Создан бенчмарк для задачи AR Manga Colorization с использованием синтетических и реальных данных. ISMAR 2024 (A*), WACV 2025 (A)



Golyadkin, M., Saraev, S., & Makarov, I. Benchmarking and Data Synthesis for Colorization of Manga Sequential Pages for Augmented Reality, ISMAR 2024.

Golyadkin, M., Plevokas, I., & Makarov, I. Closing the Domain Gap in Manga Colorization via Aligned Paired Dataset. WACV 2025

Оптимизация нейросетей для извлечения лицевых дескрипторов на мобильных устройствах. Фреймворк позволяет адаптировать архитектуры нейросетей под конкретные устройства с учётом их аппаратных ограничений. Методология включает использование эволюционных алгоритмов и суррогатных бинарных классификаторов для быстрого выбора субсетей. ECAI demo 2024 (A)

Разработка моделей, оптимизированных под конкретные устройства, необходима для повышения производительности и приватности в мобильных приложениях.

Savchenko, A., Maslov, D., & Makarov, I. (2024). Device-Specific Facial Descriptors: Winning a Lottery with a SuperNet. In ECAI 2024 (pp. 4439-4442)

ACL 2024 статьи “Fact-checking the output of large language models via token-level uncertainty quantification”, посвященной использованию нового метода оценки неопределенности для fact-checking-а генерации LLM.

Fadeeva, E., Rubashevskii, A., Shelmanov, A, Petrakov, S., Li, H., Mubarak, H., Tsymbalov, E., Kuzmin, G., Panchenko, A., Baldwin, T., Nakov, P., Panov, M. (2024): Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification. In Proceedings of the 62th Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand. Association for Computational Linguistics.

Предложены анатомические позиционные эмбеддинги для 3D-изображений, способные точно предсказывать анатомическое положение областей, что открывает новые возможности в ретривале изображений, локализации органов и работе генеративных моделей. Эти результаты были представлены на MICCAI 2024.

Трёхмерные медицинские изображения, например, компьютерная томография (КТ), получаются в результате сканирования некоторой части тела пациента. Чаще всего сканируется довольно большая область (вся грудная клетка, или вся брюшная полость, или и то, и другое). Если нарезать КТ изображение на трёхмерные патчи, то каждый патч будет являться мини-КТ снимком небольшой анатомической области.

Обычные позиционные эмбеддинги патчей кодируют положение патча относительно края изображения. Но так как область сканирования для разных КТ изображений может отличаться (у одного пациента край изображения — это шея, а у другого диафрагма), обычные позиционные эмбеддинги не содержат информацию об анатомическом положении патчей (у патча шеи и у патча диафграмы будет один и тот же обычный позиционный эмбеддинг).

В нашей работе “Anatomical Positional Embeddings” [1] (Core A) мы предложили модель для получения data-driven эмбеддингов, которые кодируют информацию об анатомическом положении патчей (или пикселей) трёхмерных медицинских изображений. Обучение модели основано на том, что у большинства пациентов органы и анатомические области расположены относительно друг друга примерно одинаково. Наша модель обучается предсказывать трёхмерные позиционные эмбеддинги патчей на основе их визуальных признаков, так что расстояния между эмбеддингами согласованы с физическими расстояниями между патчами на исходном КТ изображении. Оказывается, что эмбеддинги, обученные таким образом, фактически являются 3D координатами в некоторой условной системе координат, связанной с туловищем усредненого пациента.

Приложения нашей модели:

- Retrieval изображений определенной анатомической области.
- Few-shot локализация органов.
- Label-efficient классификация анатомической области, в которой находятся находки, задетектированные другими моделями (например, классификация легочных узлов по долям и сегментам легких).
- Трэкинг находок на изображениях одного пациента с разницей во времени.
- Регистрация изображений.
- Conditioning в генеративных и дискриминативных моделях.

Goncharov, M., Samokhin, V., Soboleva, E., et al. Anatomical Positional Embeddings. MICCAI 2024. Core A

Разработан новый метод оценки полной неопределенности в применении к задачам классификации и сегментации медицинских изображений с учетом мнения экспертов (принято на WACV’25)

Разработан фреймворк для тестирования гипотез о соответствии эмоциональных реакций людей и решений LLM. Установлено, что эмоциональный алайнмент с людьми в большинстве случаев слабый, но эмоциональные промпты, особенно негативные, существенно влияют на поведение моделей, снижая кооперацию в играх и качество в этических задачах. [EAI: Emotional Decision-Making of LLMs in Strategic Games and Ethical Dilemmas]

Mozikov, M., Severin, N., Bodishtianu, V., Glushanina, M., Nasonov, I., Orekhov, D., Pekhotin, V.,

Makovetskiy, I., Baklashkin, M., Lavrentyev, V., Tsvigun, A., Turdakov, D., Shavrina, T., Savchenko, A., &

Makarov, I. EAI: Emotional Decision-Making of LLMs in Strategic Games and Ethical Dilemmas. NeurIPS

Разработаны новые подходы к применению оптимального транспорта для задач машинного обучения.

- Метод обучения с подкреплением для построения автономных агентов. Данный походод позволять осуществлять эффективное обучение на неполных или зашумленных данных.
- Расширение алгоритма нейронного оптимального транспорта на задачи переноса знаний с сохранением по-классовой структуры.
- Метод для быстрого и эффективного решения задачи оптимального транспорта при дисбалансе данных.

Rethinking Optimal Transport in Offline Reinforcement Learning
A Asadulaev, R Korst, A Korotin, V Egiazarian, A Filchenkov, E Burnaev
Neural Information Processing Systems 2024 (NeurIPS 2024)

Neural Optimal Transport with General Cost Functionals
A Asadulaev, A Korotin, V Egiazarian, P Mokrov, E Burnaev
The 12th International Conference on Learning Representations (ICLR 2024)

Light Unbalanced Optimal transport
M Gazdieva, A Asadulaev, E Burnaev, A Korotin
Neural Information Processing Systems 2024 (NeurIPS 2024)

Incomplete Reinforcement Learning
A Asadulaev, R Korst, A Korotin, V Egiazarian, A Filchenkov, E Burnaev
The 12th International Conference on Learning Representations (ICLR 2024 Workshop)

Научные публикации

Статьи на конференциях A*

5 AAAI	4 ACL	2 CVPR	6 ICLR
7 ICML	2 ISMAR	17 NeurIPS	1 ICRA
1 WWW	3 ACM KDD	2 ECCV	4 ICDM
4 IJCAI	7 EMNLP		

AAAI

Beyond Attention: Breaking the Limits of Transformer Context Length with Recurrent Memory

Aydar Bulatov, Yuri Kuratov, Yermek Kapushev, Mikhail Burtsev

A major limitation for the broader scope of problems solvable by transformers is the quadratic scaling of computational complexity with input size. In this study, we investigate the recurrent memory augmentation of pre-trained transformer models to extend input context length while linearly scaling compute. Our approach demonstrates the capability to store information in memory for sequences of up to an unprecedented two million tokens while maintaining high retrieval accuracy. Experiments with language modeling tasks show perplexity improvement as the number of processed input segments increases. These results underscore the effectiveness of our method, which has significant potential to enhance long-term dependency handling in natural language understanding and generation tasks, as well as enable large-scale context processing for memory-intensive applications.

A*

Ссылка на источник

Topological and Node Noise Filtering on 3D Meshes Using Graph Neural Networks

Vladimir Mashurov, Natalia Semenova

Topological and node noise filtration are typically considered separately. Graph Neural Networks (GNN) are commonly used for node noise filtration, as they offer high efficiency and low exploitation costs. This paper explores the solution of joint node and topological noise filtration through the use of graph neural networks. Since treating a 3D mesh as a graph is challenging, an indicator function grid representation is employed as input for GNNs to perform the joint filtering. The resulting machine learning model is inspired by point cloud to mesh reconstruction algorithms and demonstrates low computational requirements during inference, producing successful results for smooth, watertight 3D models.

A*

Ссылка на источник

AAAI

Improved Anonymous Multi-Agent Path Finding Algorithm

Zain Alabedeen Ali, Konstantin Yakovlev

We consider an Anonymous Multi-Agent Path-Finding (AMAPF) problem where the set of agents is confined to a graph, a set of goal vertices is given and each of these vertices has to be reached by some agent. The problem is to find an assignment of the goals to the agents as well as the collision-free paths, and we are interested in finding the solution with the optimal makespan. A well-established approach to solve this problem is to reduce it to a special type of a graph search problem, i.e. to the problem of finding a maximum flow on an auxiliary graph induced by the input one. The size of the former graph may be very large and the search on it may become a bottleneck. To this end, we suggest a specific search algorithm that leverages the idea of exploring the search space not through considering separate search states but rather bulks of them simultaneously. That is, we implicitly compress, store and expand bulks of the search states as single states, which results in high reduction in runtime and memory. Empirically, the resultant AMAPF solver...

Ссылка на источник

Learn to follow: Decentralized lifelong multi-agent pathfinding via planning and learning

Alexey Skrynnik, Anton Andreychuk, Maria Nesterova, Konstantin Yakovlev, Aleksandr Panov

Multi-agent Pathfinding (MAPF) problem generally asks to find a set of conflict-free paths for a set of agents confined to a graph and is typically solved in a centralized fashion. Conversely, in this work, we investigate the decentralized MAPF setting, when the central controller that posses all the information on the agents' locations and goals is absent and the agents have to sequientially decide the actions on their own without having access to a full state of the environment. We focus on the practically important lifelong variant of MAPF, which involves continuously assigning new goals to the agents upon arrival to the previous ones. To address this complex problem, we propose a method that integrates two complementary approaches: planning with heuristic search and reinforcement learning through policy optimization. Planning is utilized to construct and re-plan individual paths. We enhance our planning algorithm with a dedicated technique tailored to avoid congestion and increase the throughput of the system. We employ reinforcement learning to discover the collision avoidance policies that effectively guide the agents along the paths. The policy is implemented as a neural network and is effectively trained...

Ссылка на источник

AAAI

Decentralized Monte Carlo Tree Search for Partially Observable Multi-agent Pathfinding

Alexey Skrynnik, Anton Andreychuk, Konstantin Yakovlev, Aleksandr Panov

The Multi-Agent Pathfinding (MAPF) problem involves finding a set of conflict-free paths for a group of agents confined to a graph. In typical MAPF scenarios, the graph and the agents' starting and ending vertices are known beforehand, allowing the use of centralized planning algorithms. However, in this study, we focus on the decentralized MAPF setting, where the agents may observe the other agents only locally and are restricted in communications with each other. Specifically, we investigate the lifelong variant of MAPF, where new goals are continually assigned to the agents upon completion of previous ones. Drawing inspiration from the successful AlphaZero approach, we propose a decentralized multi-agent Monte Carlo Tree Search (MCTS) method for MAPF tasks. Our approach utilizes the agent's observations to recreate the intrinsic Markov decision process, which is then used for planning with a tailored for multi-agent tasks version of neural MCTS. The experimental results show that our approach outperforms state-of-the-art learnable MAPF solvers. The source code is available at this https URL: https://github.com/AIRI-Institute/mats-lp

A*

Ссылка на источник

ACL

Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov

Large language models (LLMs) are notorious for hallucinating, i.e., producing erroneousclaims in their output. Such hallucinations canbe dangerous, as occasional factual inaccuracies in the generated text might be obscuredby the rest of the output being generally factual, making it extremely hard for the usersto spot them. Current services that leverageLLMs usually do not provide any means fordetecting unreliable generations. Here, we aimto bridge this gap. In particular, we proposea novel fact-checking and hallucination detection pipeline based on token-level uncertaintyquantification. Uncertainty scores leverage information encapsulated in the output of a neuralnetwork or its layers to detect unreliable predictions, and we show that they can be usedto fact-check the atomic claims in the LLMoutput.

Ссылка на источник

Your Transformer is Secretly Linear

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Nikolai Gerasimenko, Ivan Oseledets, Denis Dimitrov, Andrey Kuznetsov

This paper reveals a novel linear characteristic exclusive to transformer decoders, including models such as GPT, LLaMA, OPT, BLOOM and others. We analyze embedding transformations between sequential layers, uncovering a near-perfect linear relationship (Procrustes similarity score of 0.99). However, linearity decreases when the residual component is removed due to a consistently low output norm of the transformer layer. Our experiments show that removing or linearly approximating some of the most linear blocks of transformers does not affect significantly the loss or model performance. Moreover, in our pretraining experiments on smaller models we introduce a cosine-similarity-based regularization, aimed at reducing layer linearity. This regularization improves performance metrics on benchmarks like Tiny Stories and SuperGLUE and as well successfully decreases the linearity of the models. This study challenges the existing understanding of transformer architectures, suggesting that their operation may be more linear than previously assumed.

Ссылка на источник

ACL

MERA: A Comprehensive LLM Evaluation in Russian

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton A. Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Denis Dimitrov, Alexander Panchenko, Sergey Markov

Over the past few years, one of the most notable002 advancements in AI research has been in foun003 dation models (FMs), headlined by the rise of004 language models (LMs). However, despite re005 searchers' attention and the rapid growth in LM006 application, the capabilities, limitations, and as007 sociated risks still need to be better understood.008 To address these issues, we introduce a new009 instruction benchmark, MERA, for evaluating010 foundation models oriented towards the Rus011 sian language. The benchmark encompasses 21012 evaluation tasks for generative models covering013 10 skills and is designed as a black-box test to014 ensure the exclusion of data leakage. The pa015 per introduces a methodology to evaluate FMs016 and LMs...

A*

Ссылка на источник

TaxoLLaMA: WordNet-based Model for Solving Multiple Lexical Sematic Tasks

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, Irina Nikishina

In this paper, we explore the capabilities of LLMs in capturing lexical-semantic knowledge from WordNet on the example of the LLaMA-2-7b model and test it on multiple lexical semantic tasks. As the outcome of our experiments, we present TaxoLLaMA, the everything-in-one model, lightweight due to 4-bit quantization and LoRA. It achieves 11 SotA results, 4 top-2 results out of 16 tasks for the Taxonomy Enrichment, Hypernym Discovery, Taxonomy Construction, and Lexical Entailment tasks. Moreover, it demonstrates very strong zero-shot performance on Lexical Entailment and Taxonomy Construction with no fine-tuning. We also explore its hidden multilingual and domain adaptation capabilities with a little tuning or few-shot learning. All datasets, code, and model are available online at this https URL

A*

Ссылка на источник

CVPR

The Devil is in the Details: StyleFeatureEditor for Detail-Rich StyleGAN Inversion and High Quality Image Editing

Denis Bobkov, Vadim Titov, Aibek Alanov, Dmitry Vetrov

The task of manipulating real image attributes through StyleGAN inversion has been extensively researched. This process involves searching latent variables from a welltrained StyleGAN generator that can synthesize a real image, modifying these latent variables, and then synthesizing an image with the desired edits. A balance must be struck between the quality of the reconstruction and the ability to edit. Earlier studies utilized the low-dimensional W-space for latent search, which facilitated effective editing but struggled with reconstructing intricate details. More recent research has turned to the high-dimensional feature space F, which successfully inverses the input image but loses much of the detail during editing. In this paper, we introduce StyleFeatureEditor– a novel method that enables editing in both w-latents and F-latents.

A*

Ссылка на источник

Weak-to-Strong 3D Object Detection with X-Ray Distillation

Alexander Gambashidze, Aleksandr Dadukin, Maksim Golyadkin, Maria Razzhivina, Ilya Makarov

This paper addresses the critical challenges of sparsity andocclusion in LiDAR-based 3Dobject detection. Current methods often rely on supplementary modules or specific architectural designs, potentially limiting their applicability to new and evolving architectures. To our knowledge, we are the first to propose a versatile technique that seamlessly integrates into any existing framework for 3D Object Detection, marking the first instance of Weak-to-Strong generalization in 3D computer vision. We introduce a novel framework, X-Ray Distillation with Object-Complete Frames, suitable for both supervised and semi-supervised settings, that leverages the temporal aspect of point cloud sequences. This methodextracts crucial information from both previous and subsequent LiDAR frames, creating Object-Complete frames that represent objects from multiple viewpoints, thus addressing occlusion and sparsity. Given the limitation of not being able to generate Object-Complete frames during online inference, we utilize Knowledge Distillation within a Teacher-Student framework.

A*

Ссылка на источник

ICLR

Gradual Optimization Learning for Conformational Energy Minimization

Artem Tsypin, Leonid Ugadiarov, Kuzma Khrabrov, Manvel Avetisian, Alexander Telepov, Egor Rumiantsev, Alexey Skrynnik, Aleksandr I Panov, Dmitry Vetrov, Elena Tutubalina, Artur Kadurin

Molecular conformation optimization is crucial to computer-aided drug discovery and materials design. Traditional energy minimization techniques rely on iterative optimization methods that use molecular forces calculated by a physical simulator (oracle) as anti-gradients. However, this is a computationally expensive approach that requires many interactions with a physical simulator. One way to accelerate this procedure is to replace the physical simulator with a neural network. Despite recent progress in neural networks for molecular conformation energy prediction, such models are prone to distribution shift, leading to inaccurate energy minimization. We find that the quality of energy minimization with neural networks can be improved by providing optimization trajectories as additional training data. Still, it takes around 5×105 additional conformations to match the physical simulator's optimization quality. In this work, we present the Gradual Optimization...

A*

Ссылка на источник

Light Schrödinger Bridge

Alexander Korotin, Nikita Gushchin, Evgeny Burnaev

Despite the recent advances in the field of computational Schrödinger Bridges (SB), most existing SB solvers are still heavy-weighted and require complex optimization of several neural networks. It turns out that there is no principal solver which plays the role of simple-yet-effective baseline for SB just like, e.g., -means method in clustering, logistic regression in classification or Sinkhorn algorithm in discrete optimal transport. We address this issue and propose a novel fast and simple SB solver. Our development is a smart combination of two ideas which recently appeared in the field: (a) parameterization of the Schrödinger potentials with sum-exp quadratic functions and (b) viewing the log-Schrödinger potentials as the energy functions. We show that combined together these ideas yield a lightweight, simulation-free and theoretically justified SB solver with a simple straightforward optimization objective. As a result, it allows solving SB in moderate dimensions in a matter of minutes on CPU without a painful hyperparameter selection. Our light solver resembles the Gaussian mixture model which is widely used for density estimation. Inspired by this similarity, we also prove an important theoretical result showing that our light solver is a universal approximator of SBs.

A*

Ссылка на источник

ICLR

Chemical Language Models Have Problems with Chemistry: A Case Study on Molecule Captioning Task

Veronika Ganeeva, Kuzma Khrabrov, Artur Kadurin, Andrey V. Savchenko, Elena Tutubalina

Drug discovery has been greatly enhanced through the recent fusion of molecular sciences and natural language processing, leading these research fields to significant advancements. Considering the crucial role of molecule representation in chemical understanding within these models, we introduce novel probing tests designed to evaluate chemical knowledge of molecular structure in state-of-the-art language models (LMs), specifically MolT5 and Chem+Text T5. These probing tests are conducted on a molecule captioning task to gather evidence and insights into the language models' comprehension of chemical information. By applying rules to transform molecular SMILES into equivalent variants, we have observed significant differences in the natural language descriptions generated by the LM for a given molecule depending on the exact transformation used.

A*

Ссылка на источник

Energy-guided Entropic Neural Optimal Transport

Petr Mokrov, Alexander Korotin, Alexander Kolesov, Nikita Gushchin, Evgeny Burnaev

Energy-based models (EBMs) are known in the Machine Learning community for decades. Since the seminal works devoted to EBMs dating back to the noughties, there have been a lot of efficient methods which solve the generative modelling problem by means of energy potentials (unnormalized likelihood functions). In contrast, the realm of Optimal Transport (OT) and, in particular, neural OT solvers is much less explored and limited by few recent works (excluding WGAN-based approaches which utilize OT as a loss function and do not model OT maps themselves). In our work, we bridge the gap between EBMs and Entropy-regularized OT. We present a novel methodology which allows utilizing the recent developments and technical improvements of the former in order to enrich the latter. From the theoretical perspective, we prove generalization bounds for our technique. In practice, we validate its applicability in toy 2D and image domains. To showcase the scalability, we empower our method with a pre-trained StyleGAN and apply it to high-res AFHQ unpaired I2I translation.

A*

Ссылка на источник

ICLR

Neural optimal transport with general cost functionals

Arip Asadulaev, Alexander Korotin, Vage Egiazarian, Petr Mokrov, Evgeny Burnaev

We introduce a novel neural network-based algorithm to compute optimal transport (OT) plans for general cost functionals. In contrast to common Euclidean costs, i.e., or , such functionals provide more flexibility and allow using auxiliary information, such as class labels, to construct the required transport map. Existing methods for general cost functionals are discrete and do not provide an out-of-sample estimation. We address the challenge of designing a continuous OT approach for general cost functionals in high-dimensional spaces, such as images. We construct two example functionals: one to map distributions while preserving the class-wise structure and the other one to preserve the given data pairs. Additionally, we provide the theoretical error analysis for our recovered transport plans. Our implementation is available at \url{https://github.com/machinestein/gnot}

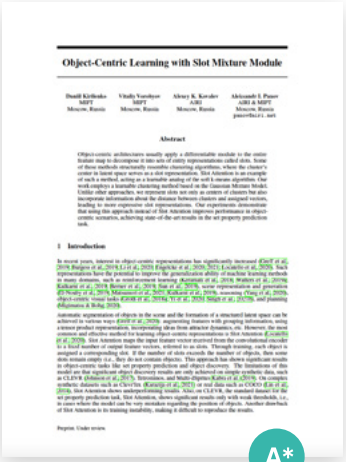
A*

Ссылка на источник

Object-Centric Learning with Slot Mixture Module

Daniil Kirilenko, Vitaliy Vorobyov, Alexey K. Kovalev, Aleksandr I. Panov

Object-centric architectures usually apply a differentiable module to the entire feature map to decompose it into sets of entity representations called slots. Some of these methods structurally resemble clustering algorithms, where the cluster's center in latent space serves as a slot representation. Slot Attention is an example of such a method, acting as a learnable analog of the soft k-means algorithm. Our work employs a learnable clustering method based on the Gaussian Mixture Model. Unlike other approaches, we represent slots not only as centers of clusters but also incorporate information about the distance between clusters and assigned vectors, leading to more expressive slot representations. Our experiments demonstrate that using this approach instead of Slot Attention improves performance in object-centric scenarios, achieving state-of-the-art results in the set property prediction task.



A*

Ссылка на источник

ICML

Estimating Barycenters of Distributions with Neural Optimal Transport

Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Evgeny Burnaev, Alexander Korotin

Given a collection of probability measures, a practitioner sometimes needs to find an «average» distribution which adequately aggregates reference distributions. A theoretically appealing notion of such an average is the Wasserstein barycenter, which is the primal focus of our work. By building upon the dual formulation of Optimal Transport (OT), we propose a new scalable approach for solving the Wasserstein barycenter problem. Our methodology is based on the recent Neural OT solver: it has bi-level adversarial learning objective and works for general cost functions. These are key advantages of our method, since the typical adversarial algorithms leveraging barycenter tasks utilize tri-level optimization and focus mostly on quadratic cost. We also establish theoretical error bounds for our proposed approach and showcase its applicability and effectiveness on illustrative scenarios and image data setups.

Ссылка на источник

Neural operators meet conjugate gradients: The FCG-NO method for efficient PDE solving

Alexander Rudikov, Vladimir Fanaskov, Ekaterina Muravleva, Yuri M Laevsky, Ivan Oseledets

Deep learning solvers for partial differential equations typically have limited accuracy. We propose to overcome this problem by using themas preconditioners. More specifically, we apply discretization-invariant neural operators tolearn preconditioners for the flexible conjugategradient method (FCG). Architecture paired withnovel loss function and training scheme allowsfor learning efficient preconditioners that can beused across different resolutions. On the theoretical side, FCG theory allows us to safely usenonlinear preconditioners that can be applied inO(N) operations without constraining the formof the preconditioners matrix. To justify learningscheme components (the loss function and theway training data is collected) we perform severalablation studies. Numerical results indicate thatour approach favorably compares with classicalpreconditioners and allows to reuse of preconditioners learned for lower resolution to the higherresolution data.

Ссылка на источник

ICML

In-Context Reinforcement Learning for Variable Action Spaces

Viacheslav Sinii, Alexander Nikulin, Vladislav Kurenkov, Ilya Zisman, Sergey Kolesnikov

Recently, it has been shown that transformers pre-trained on diverse datasets with multi-episode contexts can generalize to new reinforcement learning tasks in-context. A key limitation of previously proposed models is their reliance on a predefined action space size and structure. The introduction of a new action space often requires data re-collection and model re-training, which can be costly for some applications. In our work, we show that it is possible to mitigate this issue by proposing the Headless-AD model that, despite being trained only once, is capable of generalizing to discrete action spaces of variable size, semantic content and order. By experimenting with Bernoulli and contextual bandits, as well as a gridworld environment, we show that Headless-AD exhibits significant capability to generalize to action spaces it has never encountered, even outperforming specialized models trained for a specific set of actions on several environment configurations.

Ссылка на источник

Emergence of In-Context Reinforcement Learning from Noise Distillation

Ilya Zisman, Vladislav Kurenkov, Alexander Nikulin, Viacheslav Sinii, Sergey Kolesnikov

Recently, extensive studies in Reinforcement Learning have been carried out on the ability of transformers to adapt in-context to various environments and tasks. Current in-context RL methods are limited by their strict requirements for data, which needs to be generated by RL agents or labeled with actions from an optimal policy. In order to address this prevalent problem, we propose ADε, a new data acquisition approach that enables in-context Reinforcement Learning from noise-induced curriculum. We show that it is viable to construct a synthetic noise injection curriculum which helps to obtain learning histories. Moreover, we experimentally demonstrate that it is possible to alleviate the need for generation using optimal policies, with in-context RL still able to outperform the best suboptimal policy in a learning dataset by a 2x margin.

Ссылка на источник

ICML

Disentanglement Learning via Topology

Nikita Balabin, Daria Voronkova, Ilya Trofimov, Evgeny Burnaev, Serguei Barannikov

We propose TopDis (Topological Disentanglement), a method for learning disentangled representations via adding a multi-scale topological loss term. Disentanglement is a crucial property of data representations substantial for the explainability and robustness of deep learning models and a step towards high-level cognition. The state-of-the-art methods are based on VAE and encourage the joint distribution of latent variables to be factorized. We take a different perspective on disentanglement by analyzing topological properties of data manifolds. In particular, we optimize the topological similarity for data manifolds traversals. To the best of our knowledge, our paper is the first one to propose a differentiable topological loss for disentanglement learning. Our experiments have shown that the proposed TopDis loss improves disentanglement scores such as MIG, FactorVAE score, SAP score, and DCI disentanglement score with respect to state-of-the-art results...

A*

Ссылка на источник

Self-Supervised Coarsening of Unstructured Grid with Automatic Differentiation

Sergei Shumilin, Alexander Ryabov, Nikolay Yavich, Evgeny Burnaev, Vladimir Vanovskiy

Due to the high computational load of modern numerical simulation,there is a demand for approaches that would reduce the sizeof discrete problems while keeping the accuracy reasonable.In this work, we present an original algorithmto coarsen an unstructured grid basedon the concepts of differentiablephysics. We achieve thisby employing k-means clustering, autodifferentiation andstochastic minimization algorithms.We demonstrate performance of the designed algorithm on a linear parabolic equation which governsslightly compressible fluid flow in porous media.Our results show that in the considered scenarios, we reduced the number of grid points up to 10 times while preserving the modeled variable dynamics in the points of interest.The proposed approach can be applied to simulation of an arbitrary system described by evolutionary partial differential equations.

A*

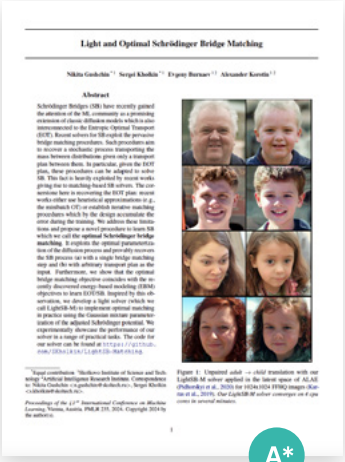
Ссылка на источник

ICML

Light and Optimal Schrödinger Bridge Matching

Nikita Gushchin, Sergei Kholkin, Evgeny Burnaev, Alexander Korotin

Schrödinger Bridges (SB) have recently gained the attention of the ML community as a promising extension of classic diffusion models whichis also interconnected to the Entropic OptimalTransport (EOT). Recent solvers for SB exploitthe pervasive bridge matching procedures. Suchprocedures aim to recover a stochastic processtransporting the mass between distributions givenonly a transport plan between them. In particular, given the EOT plan, these procedures can beadapted to solve SB. This fact is heavily exploitedby recent works giving rives to matching-basedSB solvers. The cornerstone here is recoveringthe EOT plan: recent works either use heuristical approximations (e.g., the minibatch OT) orestablish iterative matching procedures which bythe design accumulate the error during the training. We address these limitations and propose anovel procedure to learn SB which we call theoptimal Schrödinger bridge matching. It exploits the optimal parameterization of the diffusion process and provably recovers the SB process (a) with a single bridge matching step and(b) with arbitrary transport plan as the input.



Ссылка на источник

ISMAR

Benchmarking and Data Synthesis for Colorization of Manga SequentialPages for Augmented Reality

Maksim Golyadkin, Sergey Saraev, Ilya Makarov

This paper introduces an innovative approach to manga colorizationwithin augmented reality (AR) environments, focusing on theunique challenges posed by colorizing photos of manga books.We present a novel method using diffusion models to generatea synthetic dataset that accurately replicates photographed mangapages. Additionally, we have compiled a dataset of real manga photographs,capturing diverse environmental conditions. Integratingthese datasets, we established a comprehensive benchmark to evaluatecolorization models in scenarios that simulate AR applications.This benchmark was validated through a human study, confirmingthe accuracy of our metrics across both datasets. We also showedthat domain adaptation may improve model performance. Pavingthe way for practical applications, our framework enables...

Ссылка на источник

Pose Networks Unveiled: Bridging the Gap for Monocular Depth Perception

Yazan Dayoub, Anrey V. Savchenko, Ilya Makarov

Depth estimation is essential in Augmented Reality applications,enabling realistic object placement, scene understanding, spatialmapping, interaction, and environment awareness. This paper proposesa method to enhance depth model performance without increasinginference costs by improving the pose network in a selfsupervisedlearning setup. In particular, we enrich spatial informationin the pose network by incorporating features from differentscales and normalized coordinates. It is experimentally shown onthe KITTI dataset that our approach achieves a 2-7% improvementin the abs rel metric when compared to baseline techniques.

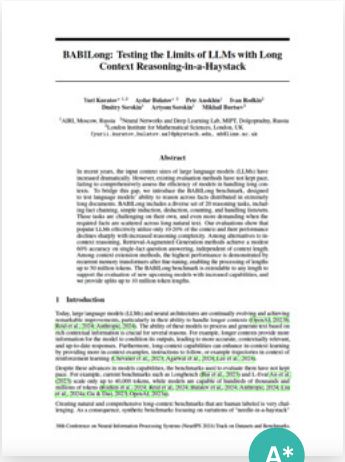
Ссылка на источник

NeurIPS

BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, Mikhail Burtsev

In recent years, the input context sizes of large language models (LLMs) have increased dramatically. However, existing evaluation methods have not kept pace, failing to comprehensively assess the efficiency of models in handling long contexts. To bridge this gap, we introduce the BABILong benchmark, designed to test language models' ability to reason across facts distributed in extremely long documents. BABILong includes a diverse set of 20 reasoning tasks, including fact chaining, simple induction, deduction, counting, and handling lists/sets. These tasks are challenging on their own, and even more demanding when the required facts are scattered across long natural text. Our evaluations show that popular LLMs effectively utilize only 10-20\% of the context and their performance declines sharply with increased reasoning complexity.



Ссылка на источник

ENOT: Expectile Regularization for Fast and Accurate Training of Neural Optimal Transport

Nazar Buzun, Maksim Bobrin, Dmitry V. Dylov

We present a new extension for Neural Optimal Transport (NOT) training procedure, capable of accurately and efficiently estimating optimal transportation plan via specific regularisation on conjugate potentials. The main bottleneck of existing NOT solvers is associated with the procedure of finding a near-exact approximation of the conjugate operator (i.e., the c-transform), which is done either by optimizing over maximin objectives or by the computationally-intensive fine-tuning of the initial approximated prediction. We resolve both issues by proposing a new, theoretically justified loss in the form of expectile regularization that enforces binding conditions on the learning dual potentials. Such a regularization provides the upper bound estimation over the distribution of possible conjugate potentials and makes the learning stable, eliminating the need for additional extensive finetuning. We formally justify the efficiency of our method, called Expectile-Regularised Neural Optimal Transport (ENOT). ENOT outperforms previous state-of-the-art approaches on the Wasserstein-2 benchmark tasks by a large margin (up to a 3-fold improvement in quality and up to a 10-fold improvement in runtime).

Ссылка на источник

NeurIPS

Light Unbalanced Optimal Transport

Milena Gazdieva, Arip Asadulaev, Alexander Korotin, Evgeny Burnaev

While the continuous Entropic Optimal Transport (EOT) field has been actively developing in recent years, it became evident that the classic EOT problem is prone to different issues like the sensitivity to outliers and imbalance of classes in the source and target measures. This fact inspired the development of solvers that deal with the unbalanced EOT (UEOT) problem – the generalization of EOT allowing for mitigating the mentioned issues by relaxing the marginal constraints. Surprisingly, it turns out that the existing solvers are either based on heuristic principles or heavy-weighted with complex optimization objectives involving several neural networks. We address this challenge and propose a novel theoretically-justified, lightweight, unbalanced EOT solver. Our advancement consists of developing a novel view on the optimization of the UEOT problem yielding tractable and a non-minimax optimization objective. We show that combined with a light parametrization recently proposed in the field our objective leads to a fast, simple, and effective solver which allows solving the continuous UEOT problem in minutes on CPU.

A*

Ссылка на источник

∇²DFT: A Universal Quantum Chemistry Dataset of Drug-Like Molecules and a Benchmark for Neural Network Potentials

Kuzma Khrabrov, Anton Ber, Artem Tsypin, Konstantin Ushenin, Egor Rumiantsev, Alexander Telepov, Dmitry Protasov, Ilya Shenbin, Anton Alekseev, Mikhail Shirokikh, Sergey Nikolenko, Elena Tutubalina, Artur Kadurin

Methods of computational quantum chemistry provide accurate approximations of molecular properties crucial for computer-aided drug discovery and other areas of chemical science. However, high computational complexity limits the scalability of their applications. Neural network potentials (NNPs) are a promising alternative to quantum chemistry methods, but they require large and diverse datasets for training. This work presents a new dataset and benchmark called ∇²DFT that is based on the nablaDFT. It contains twice as much molecular structures, three times more conformations, new data types and tasks, and state-of-the-art models. The dataset includes energies, forces, 17 molecular properties, Hamiltonian and overlap matrices, and a wavefunction object.

A*

Ссылка на источник

NeurIPS

Rethinking Optimal Transport in Offline Reinforcement Learning

Arip Asadulaev, Alexander Korotin, Vage Egiazarian, Rostislav Korst, Andrey Filchenkov, Evgeny Burnaev

We present a novel approach for offline reinforcement learning that bridges the gap between recent advances in neural optimal transport and reinforcement learning algorithms. Our key idea is to compute the optimal transport between states and actions with an action-value cost function and implicitly recover an optimal map that can serve as a policy. Building on this concept, we develop a new algorithm called Extremal Monge Reinforcement Learning that treats offline reinforcement learning as an extremal optimal transport problem. Unlike previous transport-based offline reinforcement learning algorithms, our method focuses on improving the policy beyond the behavior policy, rather than addressing the distribution shift problem. We evaluated the performance of our method on various continuous control problems and demonstrated improvements over existing algorithms.

Ссылка на источник

RClicks: Realistic Click Simulation for Benchmarking Interactive Segmentation

Anton Antonov, Andrey Moskalenko, Denis Shepelev, Vlad Shakhuro, Alexander Krapukhin, Konstantin Soshin, Anton Konushin

The emergence of Segment Anything (SAM) sparked research interest in the field of interactive segmentation, especially in the context of image editing tasks and speeding up data annotation. Unlike common semantic segmentation, interactive segmentation methods allow users to directly influence their output through prompts (e.g. clicks). However, click patterns in real-world interactive segmentation scenarios remain largely unexplored. Most methods rely on the assumption that users would click in the center of the largest erroneous area. Nevertheless, recent studies show that this is not always the case. Thus, methods may have poor performance in real-world deployment despite high metrics in a baseline benchmark. To accurately simulate real-user clicks, we conducted a large crowdsourcing study of click patterns in an interactive segmentation scenario and collected 475K real-user clicks.

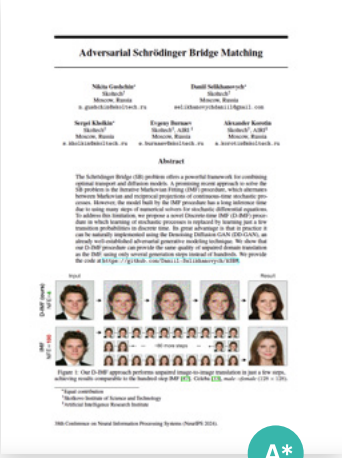
Ссылка на источник

NeurIPS

Adversarial Schrödinger Bridge Matching

Nikita Gushchin, Daniil Selikhanovych, Sergei Kholkin, Evgeny Burnaev, Alexander Korotin

The Schrödinger Bridge (SB) problem offers a powerful framework for combining optimal transport and diffusion models. A promising recent approach to solve the SB problem is the Iterative Markovian Fitting (IMF) procedure, which alternates between Markovian and reciprocal projections of continuous-time stochastic processes. However, the model built by the IMF procedure has a long inference time due to using many steps of numerical solvers for stochastic differential equations. To address this limitation, we propose a novel Discrete-time IMF (D-IMF) procedure in which learning of stochastic processes is replaced by learning just a few transition probabilities in discrete time. Its great advantage is that in practice it can be naturally implemented using the Denoising Diffusion GAN (DD-GAN), an already well-established adversarial generative modeling technique. We show that our D-IMF procedure can provide the same quality of unpaired domain translation as the IMF, using only several generation steps instead of hundreds.



Ссылка на источник

EAI: Emotional Decision-Making of LLMs in Strategic Games and Ethical Dilemmas

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Ivan Nasonov, Daniil Orekhov, Vladislav Pekhotin, Ivan Makovetskiy, Mikhail Baklashkin, Vasily Lavrentyev, Akim Tsvigun, Denis Turdakov, Tatiana Shavrina, Andrey Savchenko, Ilya Makarov

One of the urgent tasks of artificial intelligence is to assess the safety and alignment of large language models (LLMs) with human behavior. Conventional verification only in pure natural language processing benchmarks can be insufficient. Since emotions often influence human decisions, this paper examines LLM alignment in complex strategic and ethical environments, providing an in-depth analysis of the drawbacks of our psychology and the emotional impact on decision-making in humans and LLMs. We introduce the novel EAI framework for integratingemotion modeling into LLMs to examine the emotional impact on ethics and LLM-based decision-making in various strategic games, including bargaining and repeated games. Our experimental study with various LLMs demonstratedthat emotions can significantly alter the ethical decision-making...

Ссылка на источник

NeurIPS

Energy-Guided Continuous Entropic Barycenter Estimation for General Costs

Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Evgeny Burnaev, Alexander Korotin

Optimal transport (OT) barycenters are a mathematically grounded way of averaging probability distributions while capturing their geometric properties. In short, the barycenter task is to take the average of a collection of probability distributions w.r.t. given OT discrepancies. We propose a novel algorithm for approximating the continuous Entropic OT (EOT) barycenter for arbitrary OT cost functions. Our approach is built upon the dual reformulation of the EOT problem based on weak OT, which has recently gained the attention of the ML community. Beyond its novelty, our method enjoys several advantageous properties: (i) we establish quality bounds for the recovered solution; (ii) this approach seamlessly interconnects with the Energy-Based Models (EBMs) learning procedure enabling the use of well-tuned algorithms for the problem of interest; (iii) it provides an intuitive optimization scheme...

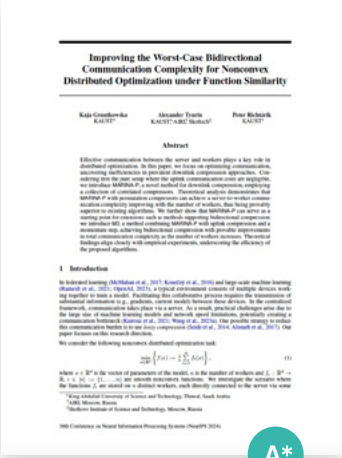
A*

Ссылка на источник

Improving the Worst-Case Bidirectional Communication Complexity for Nonconvex Distributed Optimization under Function Similarity

Kaja Gruntkowska, Alexander Tyurin, Peter Richtárik

Effective communication between the server and workers plays a key role in distributed optimization. In this paper, we focus on optimizing the server-to-worker communication, uncovering inefficiencies in prevalent downlink compression approaches. Considering first the pure setup where the uplink communication costs are negligible, we introduce MARINA-P, a novel method for downlink compression, employing a collection of correlated compressors. Theoretical analyses demonstrates that MARINA-P with permutation compressors can achieve a server-to-worker communication complexity improving with the number of workers, thus being provably superior to existing algorithms. We further show that MARINA-P can serve as a starting point for extensions such as methods supporting bidirectional compression. We introduce M3, a method combining MARINA-P with uplink compression and a momentum step, achieving bidirectional compression with provable improvements in total communication complexity as the number of workers increases. Theoretical findings align closely with empirical experiments, underscoring the efficiency of the proposed algorithms.



A*

Ссылка на источник

NeurIPS

XLand-MiniGrid: Scalable Meta-Reinforcement Learning Environments in JAX

Alexander Nikulin, Vladislav Kurenkov, Ilya Zisman, Artem Agarkov, Viacheslav Sinii, Sergey Kolesnikov

Inspired by the diversity and depth of XLand and the simplicity and minimalism of MiniGrid, we present XLand-MiniGrid, a suite of tools and grid-world environments for meta-reinforcement learning research. Written in JAX, XLand-MiniGrid is designed to be highly scalable and can potentially run on GPU or TPU accelerators, democratizing large-scale experimentation with limited resources. Along with the environments, XLand-MiniGrid provides pre-sampled benchmarks with millions of unique tasks of varying difficulty and easy-to-use baselines that allow users to quickly start training adaptive agents. In addition, we have conducted a preliminary analysis of scaling and generalization, showing that our baselines are capable of reaching millions of steps per second during training and validating that the proposed benchmarks are challenging.

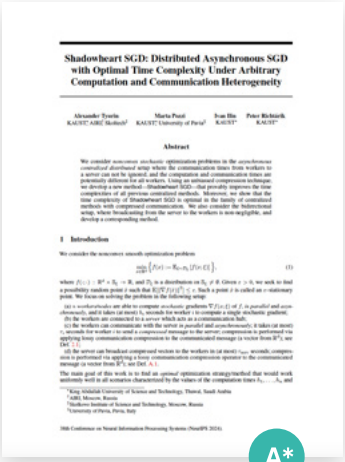
A*

Ссылка на источник

Shadowheart SGD: Distributed Asynchronous SGD with Optimal Time Complexity Under Arbitrary Computation and Communication Heterogeneity

Alexander Tyurin, Marta Pozzi, Ivan Ilin, Peter Richtárik

We consider nonconvex stochastic optimization problems in the asynchronous centralized distributed setup where the communication times from workers to a server can not be ignored, and the computation and communication times are potentially different for all workers. Using an unbiassed compression technique, we develop a new method-Shadowheart SGD-that provably improves the time complexities of all previous centralized methods. Moreover, we show that the time complexity of Shadowheart SGD is optimal in the family of centralized methods with compressed communication. We also consider the bidirectional setup, where broadcasting from the server to the workers is non-negligible, and develop a corresponding method.



A*

Ссылка на источник

NeurIPS

Optimal Flow Matching: Learning Straight Trajectories in Just One Step

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, Alexander Korotin

Over the several recent years, there has been a boom in development of Flow Matching (FM) methods for generative modeling. One intriguing property pursued by the community is the ability to learn flows with straight trajectories which realize the Optimal Transport (OT) displacements. Straightness is crucial for the fast integration (inference) of the learned flow's paths. Unfortunately, most existing flow straightening methods are based on non-trivial iterative FM procedures which accumulate the error during training or exploit heuristics based on minibatch OT. To address these issues, we develop and theoretically justify the novel Optimal Flow Matching approach which allows recovering the straight OT displacement for the quadratic transport in just one FM step. The main idea of our approach is the employment of vector field for FM which are parameterized by convex functions.

Ссылка на источник

On the Optimal Time Complexities in Decentralized Stochastic Asynchronous Optimization

Alexander Tyurin, Peter Richtárik

We consider the decentralized stochastic asynchronous optimization setup, where many workers asynchronously calculate stochastic gradients and asynchronously communicate with each other using edges in a multigraph. For both homogeneous and heterogeneous setups, we prove new time complexity lower bounds under the assumption that computation and communication speeds are bounded. We develop a new nearly optimal method, Fragile SGD, and a new optimal method, Amelie SGD, that converge under arbitrary heterogeneous computation and communication speeds and match our lower bounds (up to a logarithmic factor in the homogeneous setting). Our time complexities are new, nearly optimal, and provably improve all previous asynchronous/synchronous stochastic methods in the decentralized setup.

Ссылка на источник

NeurIPS

Freya PAGE: First Optimal Time Complexity for Large-Scale Nonconvex Finite-Sum Optimization with Heterogeneous Asynchronous Computations

Alexander Tyurin, Kaja Gruntkowska, Peter Richtárik

In practical distributed systems, workers are typically not homogeneous, and due to differences in hardware configurations and network conditions, can have highly varying processing times. We consider smooth nonconvex finite-sum (empirical risk minimization) problems in this setup and introduce a new parallel method, Freya PAGE, designed to handle arbitrarily heterogeneous and asynchronous computations. By being robust to «stragglers» and adaptively ignoring slow computations, Freya PAGE offers significantly improved time complexity guarantees compared to all previous methods, including Asynchronous SGD, Rennala SGD, SPIDER, and PAGE, while requiring weaker assumptions. The algorithm relies on novel generic stochastic

A*

Ссылка на источник

HairFastGAN: Realistic and Robust Hair Transfer with a Fast Encoder-Based Approach

Maxim Nikolaev, Mikhail Kuznetsov, Dmitry P. Vetrov, Aibek Alanov

Our paper addresses the complex task of transferring a hairstylefrom a reference image to an input photo for virtual hair try-on. This taskis challenging due to the need to adapt to various photo poses, the sensitivity of hairstyles, and the lack of objective metrics. The current stateof the art hairstyle transfer methods use an optimization process for different parts of the approach, making them inexcusably slow. At the sametime, faster encoder-based models are of very low quality because theyeither operate in StyleGAN's W+ space or use other low-dimensionalimage generators. Additionally, both approaches have a problem withhairstyle transfer when the source pose is very different from the targetpose, because they either don't consider the pose at all or deal with it inefficiently. In our paper, we present the HairFast model, which uniquelysolves these problems and achieves high resolution, near real-time performance, and superior reconstruction compared to optimization problembased methods.

A*

Ссылка на источник

NeurIPS

Group and Shuffle: Efficient Structured Orthogonal Parametrization

Mikhail Gorbunov, Nikolay Yudin, Vera Soboleva, Aibek Alanov, Alexey Naumov, Maxim Rakhuba

The increasing size of neural networks has led to a growing demand for methods of efficient fine-tuning. Recently, an orthogonal fine-tuning paradigm was introduced that uses orthogonal matrices for adapting the weights of a pretrained model. In this paper, we introduce a new class of structured matrices, which unifies and generalizes structured classes from previous works. We examine properties of this class and build a structured orthogonal parametrization upon it. We then use this parametrization to modify the orthogonal fine-tuning framework, improving parameter and computational efficiency. We empirically validate our method on different domains, including adapting of text-to-image diffusion models and downstream task fine-tuning in language modeling. Additionally, we adapt our construction for orthogonal convolutions and conduct experiments with 1-Lipschitz neural networks.


Group and Shuffle: Efficient Structured Orthogonal Parametrization

Abstract

1 Introduction

A*

Ссылка на источник

ICRA

Neural Potential Field for Obstacle-Aware Local Motion Planning

Muhammad Alhaddad, Konstantin Mironov, Aleksey Staroverov, Aleksandr Panov

Model predictive control (MPC) may provide local motion planning for mobile robotic platforms. The challenging aspect is the analytic representation of collision cost for the case when both the obstacle map and robot footprint are arbitrary. We propose a Neural Potential Field: a neural network model that returns a differentiable collision cost based on robot pose, obstacle map, and robot footprint. The differentiability of our model allows its usage within the MPC solver.
It is computationally hard to solve problems with a very high number of parameters. Therefore, our architecture includes neural image encoders, which transform obstacle maps and robot footprints into embeddings, which reduce problem dimensionality by two orders of magnitude. The reference data for network training are generated based on algorithmic calculation of a signed distance function. Comparative experiments showed that the proposed approach is comparable with existing local planners: it provides trajectories with outperforming smoothness, comparable path length, and safe distance from obstacles. Experiment on Husky UGV mobile robot showed that our approach allows real-time and safe local planning. The code for our approach is presented at https://github.com/cog-isa/NPField together with demo video.

A*

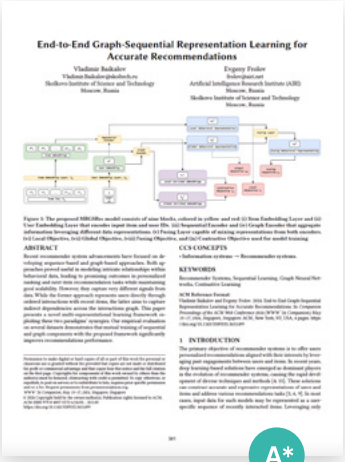
Ссылка на источник

WWW

End-to-End Graph-Sequential Representation Learning for Accurate Recommendations

Vladimir Baikalov, Evgeny Frolov

Recent recommender system advancements have focused on developing sequence-based and graph-based approaches. Both approaches proved useful in modeling intricate relationships within behavioral data, leading to promising outcomes in personalized ranking and next-item recommendation tasks while maintaining good scalability. However, they capture very different signals from data. While the former approach represents users directly through ordered interactions with recent items, the latter aims to capture indirect dependencies across the interactions graph. This paper presents a novel multi-representational learning framework exploiting these two paradigms' synergies. Our empirical evaluation on several datasets demonstrates that mutual training of sequential and graph components with the proposed framework significantly improves recommendations performance.



Ссылка на источник

ACM KDD

Uplift Modelling via Gradient Boosting

Bulat Ibragimov, Anton Vakhrushev

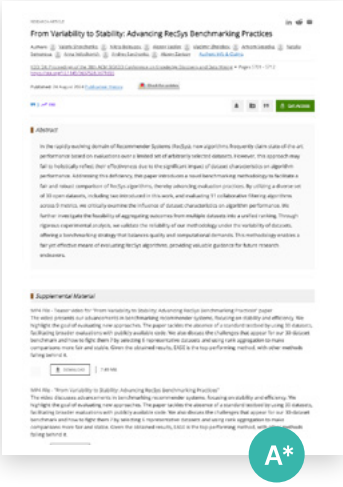
The Gradient Boosting machine learning ensemble algorithm, well-known for its proficiency and superior performance in intricate machine learning tasks, has encountered limited success in the realm of uplift modeling. Uplift modeling is a challenging task that necessitates a known target for the precise computation of the training gradient. The prevailing two-model strategies, which separately model treatment and control outcomes, are encumbered with limitations as they fail to directly tackle the uplift problem.This paper presents an innovative approach to uplift modeling that employs Gradient Boosting. Unlike previous works, our algorithm utilizes multioutput boosting model and calculates the uplift gradient based on intermediate surrogate predictions and directly models the concealed target. This method circumvents the requirement for a known target and addresses the uplift problem more effectively than existing solutions.Moreover, we broaden the scope of this solution to encompass multitreatment settings, thereby enhancing its applicability. This novel approach not only overcomes the limitations of the traditional two-model strategies...

Ссылка на источник

From Variability to Stability: Advancing RecSys Benchmarking Practices

Valeriy Shevchenko, Nikita Belousov, Alexey Vasilev, Vladimir Zholobov, Artyom Sosedka, Natalia Semenova, Anna Volodkevich, Andrey Savchenko, Alexey Zaytsev

In the rapidly evolving domain of Recommender Systems (RecSys),new algorithms frequently claim state-of-the-art performance basedon evaluations over a limited set of arbitrarily selected datasets.However, this approach may fail to holistically reflect their effectiveness due to the significant impact of dataset characteristicson algorithm performance. Addressing this deficiency, this paperintroduces a novel benchmarking methodology to facilitate a fairand robust comparison of RecSys algorithms, thereby advancingevaluation practices. By utilizing a diverse set of 30 open datasets,including two introduced in this work, and evaluating 11 collaborative filtering algorithms across 9 metrics, we critically examinethe influence of dataset characteristics on algorithm performance. We further investigate the feasibility of aggregating outcomes frommultiple datasets into a unified ranking. Through rigorous experimental analysis, we validate the reliability of our methodology under the variability of datasets, offering a benchmarking strategy thatbalances quality and computational demands. This methodologyenables a fair yet effective means of evaluating RecSys algorithms,providing valuable guidance for future research endeavors.



Ссылка на источник

ACM KDD

Learn Together Stop Apart: an Inclusive Approach to Ensemble Pruning

Bulat Ibragimov; Gleb Gusev

Gradient Boosting is a leading learning method that builds ensembles and adapts their sizes to particular tasks, consistently delivering top-tier results across various applications. However, determining the optimal number of models in the ensemble remains a critical yet underexplored aspect. Traditional approaches assume a universal ensemble size effective for all data points, which may not always hold true due to data heterogeneity.This paper introduces an adaptive approach to early stopping in Gradient Boosting, addressing data heterogeneity by assigning different stop moments to different data regions at inference time while still training a common ensemble on the entire dataset. We propose two methods: Direct Supervised Partition (DSP) and Indirect Supervised Partition (ISP). The DSP method uses a decision tree to partition the data based on learning curves, while ISP leverages the dataset's geometric and target distribution characteristics.An effective validation protocol is developed to determine the optimal number of early stopping regions or detect when the heterogeneity assumption does not hold. Experiments using state-of-the-art implementations of Gradient Boosting, LightGBM, and CatBoost, on standard benchmarks demonstrate that our methods enhance model precision by up to 2%, underscoring the significance of this research direction. This approach does not increase computational complexity and can be easily integrated into existing learning pipelines.

Ссылка на источник

ECCV

Scalar Function Topology Divergence: Comparing Topology of 3D Objects

Ilya Trofimov, Daria Voronkova, Eduard Tulchinskii, Evgeny Burnaev, Serguei Barannikov

We propose a new topological tool for computer visionScalar Function Topology Divergence (SFTD), which measures the dissimilarity of multi-scale topology between sublevel sets of two functions having a common domain. Functions can be defined on an undirected graph or Euclidean space of any dimensionality. Most of the existing methods for comparing topology are based on Wasserstein distance between persistence barcodes and they don’t take into account the localization of topological features. On the other hand, the minimization of SFTD ensures that the corresponding topological features of scalar functions are located in the same places. The proposed tool provides useful visualizations depicting areas where functions have topological dissimilarities. We provide applications of the proposed method to 3D computer vision. In particular, experiments demonstrate that SFTD improves the reconstruction of cellular 3D shapes from 2D fluorescence microscopy images, and helps to identify topological errors in 3D segmentation.

Ссылка на источник

Guide-and-Rescale: Self-Guidance Mechanism for Effective Tuning-Free Real Image Editing

Vadim Titov, Madina Khalmatova, Alexandra Ivanova, Dmitry Vetrov, and Aibek Alanov

Despite recent advances in large-scale text-to-image generative models, manipulating real images with these models remains a challenging problem. The main limitations of existing editing methods are that they either fail to perform with consistent quality on a wide range of image edits or require time-consuming hyperparameter tuning or fine-tuning of the diffusion model to preserve the image-specific appearance of the input image. We propose a novel approach that is built upon a modified diffusion sampling process via the guidance mechanism. In this work, we explore the self-guidance technique to preserve the overall structure of the input image and its local regions appearance that should not be edited. In particular, we explicitly introduce layout-preserving energy functions that are aimed to save local and global structures of the source image. Additionally, we propose a noise rescaling mechanism...

Ссылка на источник

ICDM

LLM-KT: A Versatile Framework for Knowledge Transfer from Large Language Models to Collaborative Filtering

Nikita Severin, Aleksei Ziablitsev, Yulia Savelyeva, Valeriy Tashchilin, Ivan Bulychev, Mikhail Yushkov, Artem Kushneruk, Amaliya Zaryvnykh, Dmitrii Kiselev, Andrey Savchenko, Ilya Makarov

We present LLM-KT, a flexible framework designed to enhance collaborative filtering (CF) models by seamlessly integrating LLM (Large Language Model)-generated features. Unlike existing methods that rely on passing LLM-generated features as direct inputs, our framework injects these features into an intermediate layer of any CF model, allowing the model to reconstruct and leverage the embeddings internally. This model-agnostic approach works with a wide range of CF models without requiring architectural changes, making it adaptable to various recommendation scenarios. Our framework is built for easy integration and modification, providing researchers and developers with a powerful tool for extending...

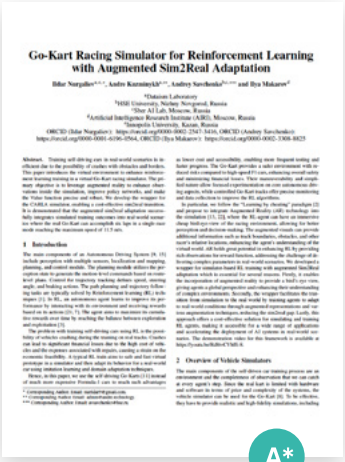
A*

Ссылка на источник

Go-Kart Racing Simulator for Reinforcement Learning with Augmented Sim2Real Adaptation

Ildar Nurgaliev, Andre Kuzminykh, Andrey Savchenko, and Ilya Makarov

Training self-driving cars in real-world scenarios is inefficient due to the possibility of crashes with obstacles and borders. This paper introduces the virtual environment to enhance reinforcement learning training in a virtual Go-Kart racing simulator. The primary objective is to leverage augmented reality to enhance observations inside the simulation, improve policy networks, and make the Value function precise and robust. We develop the wrapper for the CARLA simulator, enabling a cost-effective sim2real transition. It is demonstrated that the augmented sim2real adaptation successfully integrates simulated training outcomes into real-world scenarios where the real Go-Kart can accomplish six laps in a single-race mode reaching the maximum speed of 11.5 m/s.



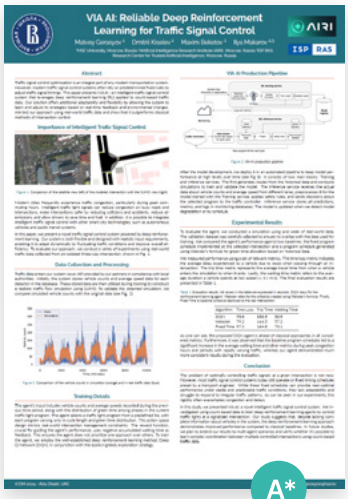
A*

ICDM

VIA AI: Reliable Deep Reinforcement Learning for Traffic Signal Control

Matvey Gerasyov, Dmitrii Kiselev, Maxim Beketov, and Ilya Makarov

Traffic signal control optimization is an integral part of any modern transportation system. However, modern traffic signal control systems often rely on predetermined fixed rules to adjust traffic signal timings. This paper presents VIA AI — an intelligent traffic signal control system that leverages deep reinforcement learning (RL) applied to count-based traffic data. Our solution offers additional adaptability and flexibility by allowing the system to learn and adjust its strategies based on real-time feedback and environmental changes. We test our approach using real-world traffic data and show that it outperforms classical methods of intersection control.



Ссылка на источник

Video-based learning of sign languages: one pre-train to fit them all

Maxim Novopoltsev, Aleksandr Tulenkov, Ruslan Murtazin, Roman Akhidov, Iuliia Zemtsova, Emilia Bojarskaja, Daria Bondarenko, Andrey Savchenko, and Ilya Makarov

This paper presents a novel system for recognizing sign language from video, addressing a critical need for improved communication accessibility for the deaf and hard-of-hearing communities. We developed a foundation model for Isolated Sign Language Recognition that uses self-supervised pre-training to address data scarcity issues. By applying the VideoMAE algorithm and a specially prepared dataset of American Sign Language videos, we created a vision transformer for video classification that performs exceptionally well. Our model achieves state-of-the-art results for Greek (GSL) and Russian (Slovo) sign languages, and comparable results for American (WLASL) and Turkish (AUTSL) sign languages. The fine-tuning process was efficient, with optimal performance in under forty epochs for each language. We also built a sign language learning tool integrated with the sign language recognition algorithm, showcasing its practical use in educational settings.

IJCAI

Do You Remember the Future? Weak-to-Strong Generalization in 3D Object Detection

Alexander Gambashidze, Aleksandr Dadukin, Maxim Golyadkin, Maria Razzhivina, Ilya Makarov

This paper demonstrates a novel method for LiDAR-based 3D object detection, addressing major field challenges: sparsity and occlusion. Our approach leverages temporal point cloud sequences to generate frames that provide comprehensive views of objects from multiple angles. To address the challenge of generating these frames in real-time, we employ Knowledge Distillation within a Teacher-Student framework, allowing the Student model to emulate the Teacher's advanced perception. We pioneered the application of weak-to-strong generalization in computer vision by training our Teacher model on enriched, object-complete data. In this demo, we showcase the exceptional quality of labels produced by the X-Ray Teacher on object-complete frames, showing our method distilling its knowledge to enhance object 3D detection models.

A*

Ссылка на источник

Plug-and-Play Unsupervised Fault Detection and Diagnosis for Complex Industrial Monitoring

Maksim Golyadkin, Maria Shtark, Petr Ivanov, Alexander Kozhevnikov, Leonid Zhukov, Ilya Makarov

Today industrial facilities are equipped with lots of sensors throughout all the production line for monitoring means. Gathered data can be used to detect and predict failures; however, manual labeling of large amounts of data for supervised learning is complicated. This paper introduces an innovative approach to unsupervised fault detection and diagnosis tailored for monitoring industrial chemical processes. We showcase the efficacy of our model using two publicly accessible datasets from the Tennessee Eastman Process, each containing various faults. Furthermore, we illustrate that by fine-tuning the model on a limited amount of labeled data, it achieves performance close to that of a state-of-the-art model trained on the entire dataset.In our experiments, we show through human evaluation and quantitative analysis that the proposed method allows to produce desired editing which is more preferable by humans and also achieves a better trade-off between editing quality and preservation of the original image. Our code is available at this https URL.

A*

Ссылка на источник

IJCAI

AADMIP: Adversarial Attacks and Defenses Modeling in Industrial Processes

Vitaliy Pozdnyakov, Aleksandr Kovalenko, Ilya Makarov, Mikhail Drobyshevskiy, Kirill Lukyanov

The development of the smart manufacturing trend includes the integration of Artificial Intelligence technologies into industrial processes. One example of such implementation is deep learning models that diagnose the current state of a technological process. Recent studies have demonstrated that small data perturbations, named adversarial attacks, can significantly affect the correct predictions of such models. This fact is critical in industrial systems, where AI-based decisions can be made to manage physical equipment. In this work, we present a system which can help to evaluate the robustness of technological process diagnosis models to adversarial attacks, as well as consider protection options. We briefly review the system's modules and also consider some useful applications. Our demo video is available at: http://tinyurl.com/3by9zcj5

Ссылка на источник

Probabilistically Robust Watermarking of Neural Networks

Mikhail Pautov, Nikita Bogdanov, Stanislav Pyatkin, Oleg Rogov, Ivan Oseledets

As deep learning (DL) models are widely and effectively used in Machine Learning as a Service(MLaaS) platforms, there is a rapidly growinginterest in DL watermarking techniques that canbe used to confirm the ownership of a particularmodel. Unfortunately, these methods usually produce watermarks susceptible to model stealing attacks. In our research, we introduce a novel trigger set-based watermarking approach that demonstrates resilience against functionality stealing attacks, particularly those involving extraction anddistillation. Our approach does not require additional model training and can be applied to anymodel architecture. The key idea of our method isto compute the trigger set, which is transferable between the source model and the set of proxy models with a high probability. In our experimentalstudy, we show that if the probability of the set being transferable is reasonably high, it can be effectively used for ownership verification of the stolenmodel. We evaluate our method on multiple benchmarks and show that our approach outperforms current state-of-the-art watermarking techniques in allconsidered experimental setups.

Ссылка на источник

EMNLP

Lost in Translation: Chemical Language Models and the Misunderstanding of Molecule Structures

Veronika Ganeeva, Andrey Sakhovskiy, Kuzma Khrabrov, Andrey Savchenko, Artur Kadurin, Elena Tutubalina

The recent integration of chemistry with natural language processing (NLP) has advanced drug discovery. Molecule representation in language models (LMs) is crucial in enhancing chemical understanding. We propose Augmented Molecular Retrieval (AMORE), a flexible zero-shot framework for assessment of Chemistry LMs of different natures: trained solely on molecules for chemical tasks and on a combined corpus of natural language texts and string-based structures. The framework relies on molecule augmentations that preserve an underlying chemical, such as kekulization and cycle replacements. We evaluate encoder-only and generative LMs by calculating a metric based on the similarity score between distributed representations of molecules...

Ссылка на источник

DeepPavlov 1.0: Your Gateway to Advanced NLP Models Backed by Transformers and Transfer Learning

Maksim Savkin, Anastasia Voznyuk, Fedor Ignatov, Anna Korzanova, Dmitry Karpov, Alexander Popov, Vasily Konovalov

Weintroduce DeepPavlov 1.0, an open-source framework designed for seamless use of Natural Language Processing (NLP) models, leveraging advanced transfer learning techniques. This frameworkoffers a modular, configurationbased approach, making it suitable for a wide range of NLP applications without requiring in-depth knowledge of machine learning or NLP. Built on PyTorch and supporting Hugging Face transformers, DeepPavlov 1.0 provides ready-to-use solutions for various NLP tasks. It is publicly available1 under the Apache 2.0 license and includes access to an interactive online demo2.

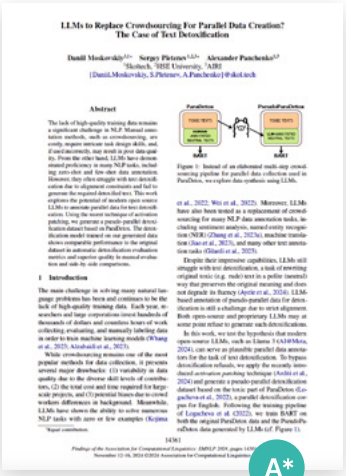
Ссылка на источник

EMNLP

LLMs to Replace Crowdsourcing For Parallel Data Creation: The Case of Text Detoxification

Daniil Moskovskiy, Sergey Pletenev, Alexander Panchenko

The lack of high-quality training data remains a significant challenge in NLP. Manual annotation methods, such as crowdsourcing, are costly, require intricate task design skills, and, if used incorrectly, may result in poor data quality. From the other hand, LLMs have demonstrated proficiency in many NLP tasks, including zero-shot and few-shot data annotation. However, they often struggle with text detoxification due to alignment constraints and fail to generate the required detoxified text. This work explores the potential of modern open source LLMs to annotate parallel data for text detoxification. Using the recent technique of activation patching, we generate a pseudo-parallel detoxification dataset based on ParaDetox. The detoxification model trained on our generated data shows comparable performance to the original dataset in automatic detoxification evaluation metrics and superior quality in manual evaluation and side-by-side comparisons.



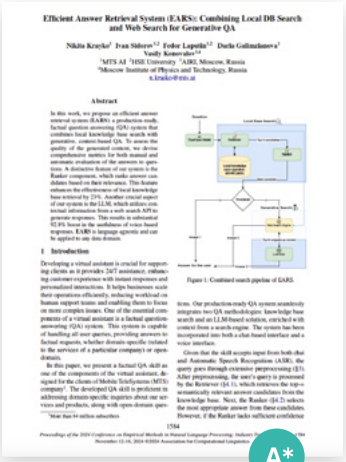
A*

Ссылка на источник

Efficient Answer Retrieval System (EARS): Combining Local DB Search and Web Search for Generative QA

Nikita Krayko, Ivan Sidorov, Fedor Laputin, Daria Galimzianova, Vasily Konovalov

In this work, we propose an efficient answer retrieval system **EARS**: a production-ready, factual question answering (QA) system that combines local knowledge base search with generative, context-based QA. To assess the quality of the generated content, we devise comprehensive metrics for both manual and automatic evaluation of the answers to questions. A distinctive feature of our system is the Ranker component, which ranks answer candidates based on their relevance. This feature enhances the effectiveness of local knowledge base retrieval by 23%. Another crucial aspect of our system is the LLM, which utilizes contextual information from a web search API to generate responses. This results in substantial 92.8% boost in the usefulness of voice-based responses. **EARS** is language-agnostic and can be applied to any data domain.



A*

Ссылка на источник

EMNLP

SparseGrad: A Selective Method for Efficient Fine-tuning of MLP Layers

Victoriia A. Chekalina, Anna Rudenko, Gleb Mezentsev, Aleksandr Mikhalev, Alexander Panchenko, Ivan Oseledets

The performance of Transformer models has been enhanced by increasing the number of parameters and the length of the processed text. Consequently, fine-tuning the entire model becomes a memory-intensive process. High-performance methods for parameter-efficient fine-tuning (PEFT) typically work with Attention blocks and often overlook MLP blocks, which contain about half of the model parameters. We propose a new selective PEFT method, namely SparseGrad, that performs well on MLP blocks. We transfer layer gradients to a space where only about 1\% of the layer's elements remain significant. By converting gradients into a sparse structure, we reduce the number of updated parameters. We apply SparseGrad to fine-tune BERT and RoBERTa for the NLU task and LLaMa-2 for the Question-Answering task. In these experiments, with identical memory requirements, our method outperforms LoRA and MeProp, robust popular state-of-the-art PEFT approaches.

A*

Ссылка на источник

Kandinsky 3: Text-to-Image Synthesis for Multifunctional Generative Framework

Arkhipkin Vladimir, Viacheslav Vasilev, Andrei Filatov, Igor Pavlov, Julia Agafonova, Nikolai Gerasimenko, Anna Averchenkova, Evelina Mironova, Bukashkin Anton, Konstantin Kulikov, Andrey Kuznetsov, Denis Dimitrov

Text-to-image (T2I) diffusion models are popular for introducing image manipulation methods, such as editing, image fusion, inpainting, etc. At the same time, image-to-video (I2V) and text-to-video (T2V) models are also built on top of T2I models. We present Kandinsky 3, a novel T2I model based on latent diffusion, achieving a high level of quality and photorealism. The key feature of the new architecture is the simplicity and efficiency of its adaptation for many types of generation tasks. We extend the base T2I model for various applications and create a multifunctional generation system that includes text-guided inpainting/outpainting, image fusion, text-image fusion, image variations generation, I2V and T2V generation. We also present a distilled version of the T2I model, evaluating inference in 4 steps of the reverse process without reducing image quality...

A*

Ссылка на источник

EMNLP

xCOMET-lite: Bridging the Gap Between Efficiency and Quality in Learned MT Evaluation Metrics

Daniil Larionov, Mikhail Seleznyov, Vasiliy Viskov, Alexander Panchenko, Steffen Eger

State-of-the-art trainable machine translation evaluation metrics like xCOMET achieve high correlation with human judgment but rely on large encoders (up to 10.7B parameters), making them computationally expensive and inaccessible to researchers with limited resources. To address this issue, we investigate whether the knowledge stored in these large encoders can be compressed while maintaining quality. We employ distillation, quantization, and pruning techniques to create efficient xCOMET alternatives and introduce a novel data collection pipeline for efficient black-box distillation. Our experiments show that, using quantization, xCOMET can be compressed up to three times with no quality degradation. Additionally, through distillation, we create an xCOMET-lite metric, which has only 2.6% of xCOMET-XXL parameters, but retains 92.1% of its quality. Besides, it surpasses strong small-scale metrics like COMET-22 and BLEURT-20 on the WMT22 metrics challenge dataset by 6.4%, despite using 50% fewer parameters. All code, dataset, and models are available online.

Ссылка на источник

Публикации AIRI

airi.net/articles/

Open source репозитории AIRI

Hugging Face

GitHub

Награды

MIDRC XAI Challenge

Исследователи из лаборатории «Сильный ИИ в медицине» и Лаборатории FusionBrain вошли в топ-5 соревнования MIDRC XAI Challenge. Конкурс был направлен создание интерпретируемых и надежных моделей искусственного интеллекта. По итогам команда представила 3 варианта решения задачи и вышла в топ-5 лучших наряду с учёными из Университета Джона Хопкинса, Университета Берна, команды Женского госпиталя в Бирмингеме, а также исследователями из Стенфорда и Университета Тюбингена.

Eedi — Mining Misconceptions in Mathematics

Команда исследователей из группы «Доверенные и безопасные интеллектуальные системы» AIRI, компании VeinCV и Сколтеха получила серебряную медаль международного соревнования по созданию алгоритма, который поможет выявлять причины появления неправильного ответа на задачи по математике в автоматическом режиме.

Concordia Challenge

Коллектив учёных из AIRI, ИСП РАН, ИТМО и стартапа Coframe вошел в топ-5 финалистов соревнования Concordia Challenge, проведенного в рамках ведущей конференции по ИИ — NeurIPS 2024. Работа получила почетное упоминание организаторов — Google DeepMind и Cooperative AI Foundation.

ASVspoof 2024 Challenge

Команда «Доверенные и безопасные интеллектуальные системы» AIRI и команда НИО «Интеллектуальные решения» МТУСИ при участии аспиранта Сколтеха создала модель определения синтетических голосов, которая вошла в топ-10 лучших решений международного соревнования ASVspoof 2024 Challenge.

Награды

YandexML Prize

Илья
Зисман

Младший научный сотрудник группы «Адаптивные агенты» Илья Зисман в номинации «Первая публикация».

Александр
Коротин

Научный сотрудник AIRI и руководитель исследовательской группы в Сколтехе Александр Коротин победил в номинации «Молодые научные руководители».

Алексей
Скрынник

Александр
Тюрин

В номинации «Исследователи» сразу два сотрудника AIRI: кандидат физико-математических наук, руководитель группы «RL агенты» лаборатории когнитивных систем искусственного интеллекта Алексей Скрынник и кандидат компьютерных наук, руководитель группы «Методы оптимизации в машинном обучении» Александр Тюрин.

Александр
Панов

Елена
Тутубалина

Сразу два исследователя AIRI победили в номинации «Научные руководители»: кандидат физико-математических наук, директор лаборатории когнитивных систем искусственного интеллекта Александр Панов и доктор компьютерных наук, руководитель группы «Прикладное NLP» AIRI, старший научный сотрудник ИСП РАН Елена Тутубалина.

Антон
Конушин

В номинации «Преподаватели ML» — кандидат физико-математических наук, руководитель группы «Пространственный интеллект» AIRI Антон Конушин.

Награды

Международная олимпиада по ИИ (IOAI)

Андрей Громыко

Стажер–исследователь лаборатории FusionBrain AIRI Андрей Громыко вошел в состав Российской команды, которая показала лучший результат в научном туре олимпиады, завоевала золотые медали, а также получила серебро в практическом этапе и стала первой по сумме баллов за оба этапа конкурса.

ImageCLEFmed MEDVQA-GI

Михаил Чайчук

Инженер-исследователь Михаил Чайчук из команды «Прикладное NLP» создал лучшее решение конкурса по генерации медицинских изображений, имитирующих результаты эндоскопических исследований желудка и кишечника, таких как гастроскопия и колоноскопия.

Золотые имена Высшей Школы

Антон Конушин

Антон Конушин победил в номинации «За вклад в науку и высшее образование» Всероссийского конкурса «Золотые имена высшей школы»

Национальная премия «Лидеры ИИ»

Елена Тутубалина

Одним из трех победителей национальной премии за вклад в развитие технологий искусственного интеллекта в категории «Премия учёным» стала Елена Тутубалина, доктор компьютерных наук, руководитель группы «Прикладное NLP» в AIRI и старший научный сотрудник ИСП РАН. Среди номинантов в этой категории — Александр Тюрин, кандидат компьютерных наук, руководитель группы «Методы оптимизации в машинном обучении» в AIRI и старший преподаватель в Сколтехе.

Мероприятия и выступления

Плейлист
AIRI Seminars

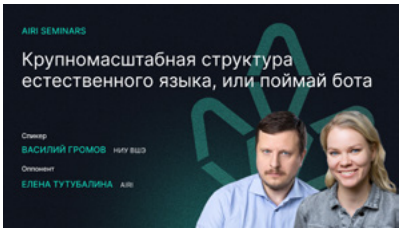
AIRI Seminars

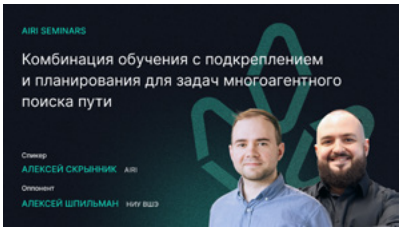
AIRI Seminars — это научный диалог на равных и знакомство профессионального сообщества с достижениями в области искусственного интеллекта.

Семинар призван популяризировать и распространять в профессиональной среде принципы и ценности, которых придерживается Институт, а также продвигать идеи, реализующие миссию AIRI: создание универсальных систем искусственного интеллекта, решающих задачи реального мира.

В качестве докладчиков или оппонентов на семинар приглашаются ведущие специалисты в области искусственного интеллекта из России и из-за рубежа, которые рассказывают и конструктивно критикуют исследовательские работы. В этом году семинары проводились не только онлайн, но и офлайн.

В 2024 году было проведено 19 семинаров.





Научные семинары AIRI доступны для всех желающих в VK Видео

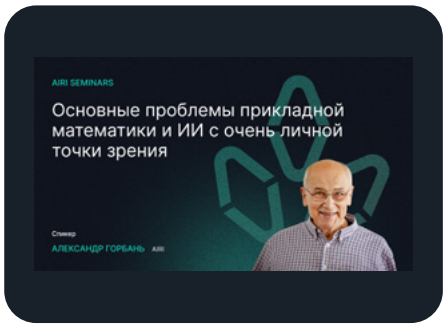
Статистика

19
семинаров

22 000
просмотров

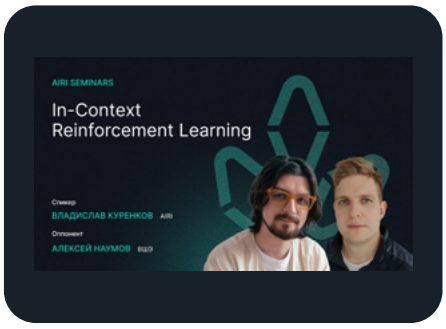
Самые просматриваемые семинары

Основные проблемы прикладной математики и ИИ с очень личной точки зрения



Ссылка на семинар

In-Context Reinforcement Learning



Ссылка на семинар

Команда


Александр Панов


Алексей Скрынник


Александра Бройтман

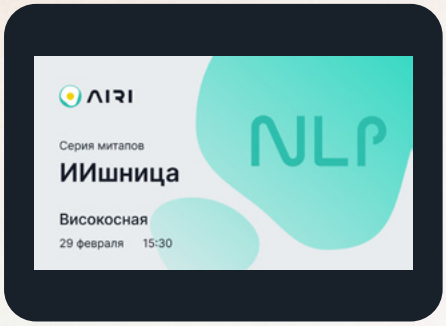

Екатерина Мамонтова


Юлия Трехлетова


Кристина Денисова

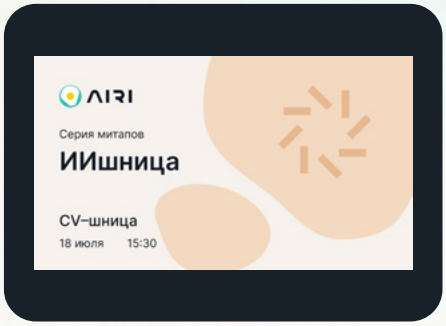
Серия митапов ИИшница

В 2024 году мы продолжили развивать спецпроект «ИИшница» — это серия онлайн-митапов, где ученые рассказывают про искусственный интеллект в рамках 20-ти минутного научного доклада.



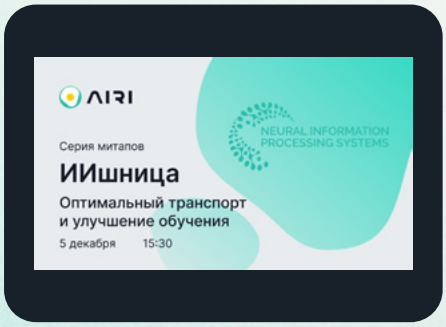
29 февраля

Високосная ИИшница



18 июля

CV-шница



5 декабря

NeurIPS 2024: Оптимальный транспорт и улучшение обучения

Ссылка на митап

6 декабря

NeurIPS 2024: Датасеты, бенчмарки и обзоры

Ссылка на митап

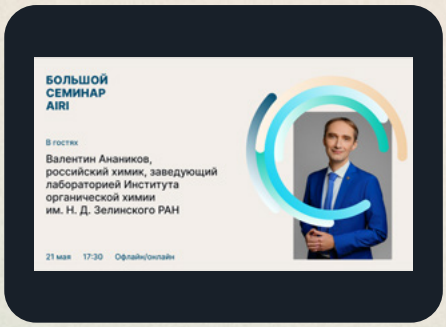
Большой Семинар AIRI

Большой Семинар AIRI — пространство для знакомства исследователей из разных областей и всех, кто интересуется наукой.

Руководитель Большого Семинара AIRI — доктор физико-математических наук, профессор РАН, CEO Института AIRI, профессор Сколтеха Иван Оселедец.

21 мая

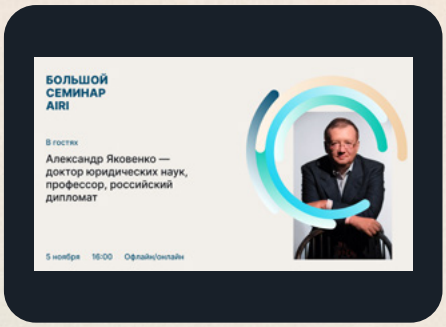
Искусственный интеллект в химии



Ссылка на семинар

5 ноября

Тенденции мирового развития



Ссылка на семинар

Ученые AIRI приняли участие в подкастах

Александр Панов в гостях у подкаста «Биг Дата».

Ссылка на подкаст

Андрей Кузнецов в гостях у подкаста «ОСНОВА».

Ссылка на подкаст

Ольга Кардымон и Вениамин Фишман в гостях у подкаста Creative Science Lab.

Ссылка на подкаст

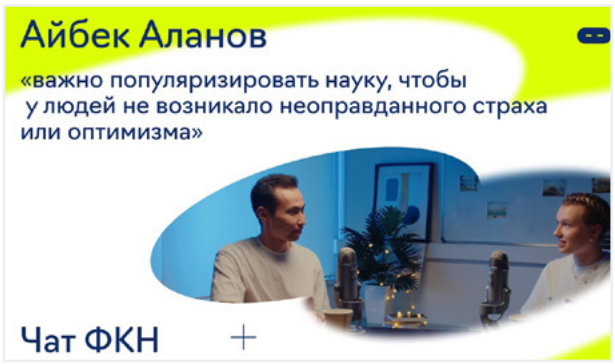
Иван Оселедец в гостях у подкаста «Деньги любят техно».

Ссылка на подкаст

Ученые AIRI приняли участие в подкастах

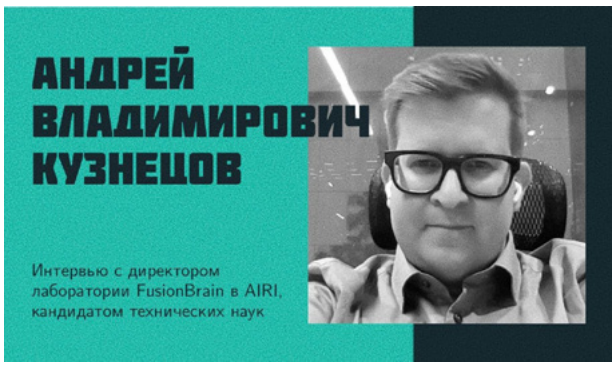
Иван Оселедец в гостях
у подкаста «Ноосфера»

Ссылка
на подкаст



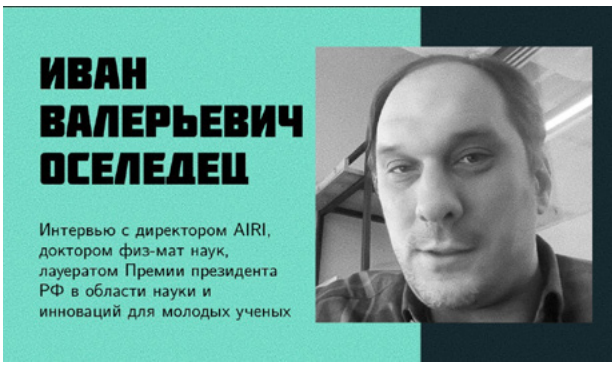
Айбек Аланов в гостях
у подкаста «ЧатФКН»

Ссылка
на подкаст



Андрей Кузнецов в гостях
у Егора Бугаенко

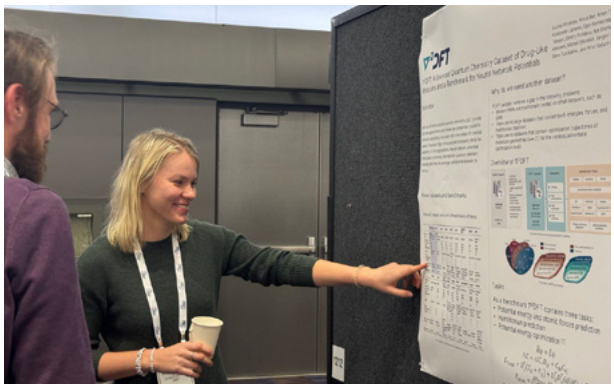
Ссылка
на подкаст



Иван Оселедец в гостях
у Егора Бугаенко

Ссылка
на подкаст

Исследователи AIRI представили свои результаты на более чем 70-ти научных конференциях



NeurIPS 2024 в Канаде

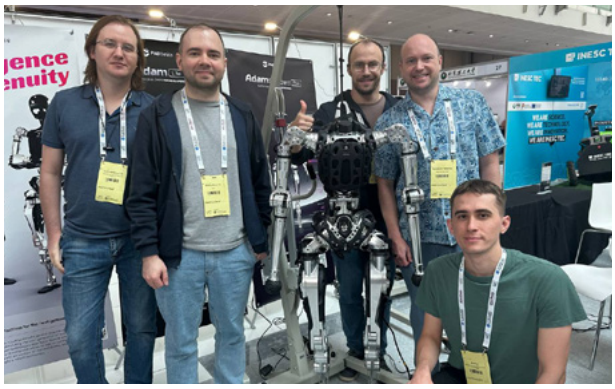
EMNLP 2024 в США

ECAI–2024 в Испании

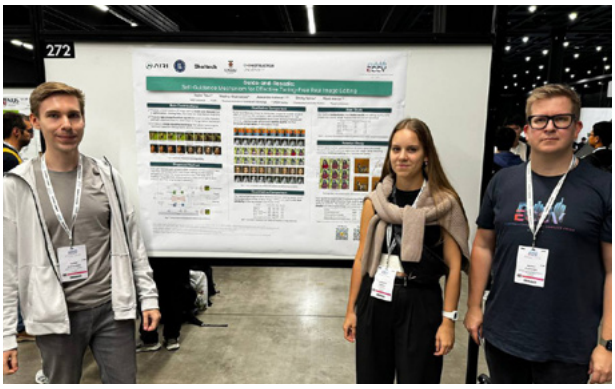
AIST 2024 в Кыргызстане

Fall into ML 2024 в Москве

RecSys 2024 в Италии



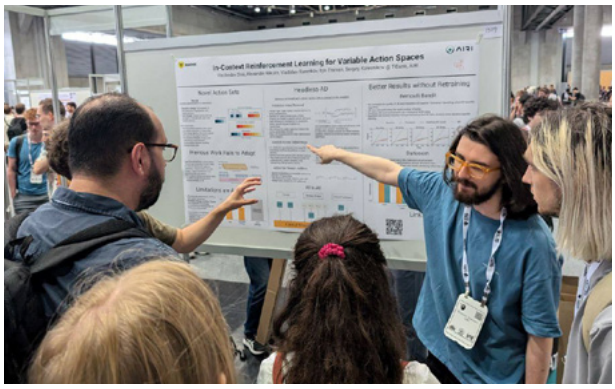
IROS 2024 в ОАЭ



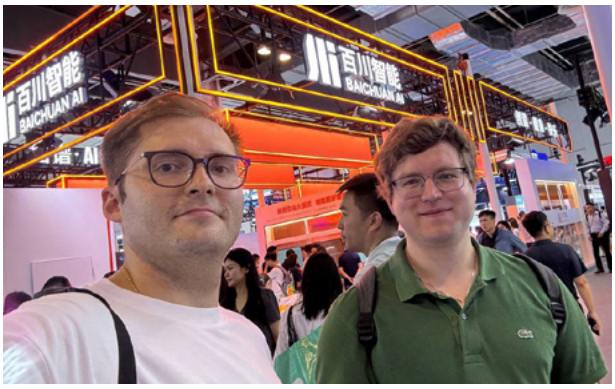
ECCV 2024 в Италии

ACL-2024 в Тайланде

IJCAI 2024 в Южной Корее



ICML 2024 в Австрии

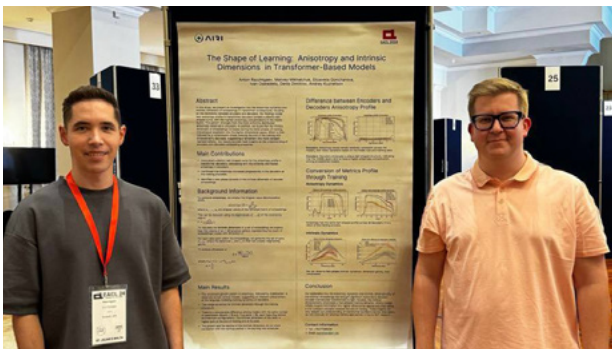


WAIC 2024 в Китае

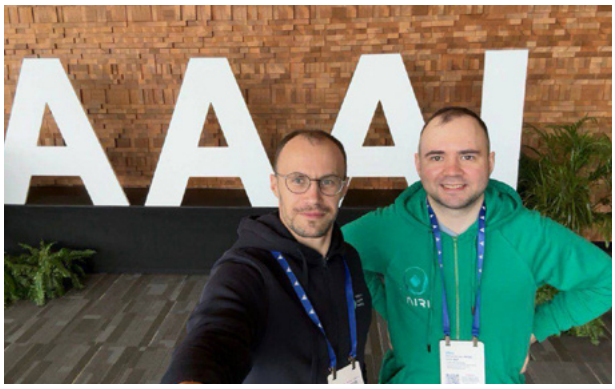
NAACL 2024 в Мексике

LREC-COLING 2024 в Италии

ICLR 2024 в Австрии

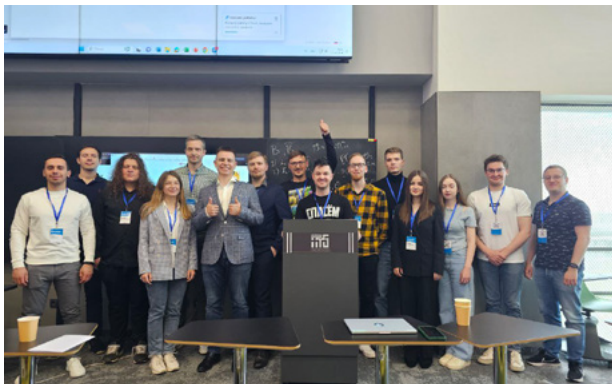


EACL 2024 на Мальте



AAAI 2024 в Канаде

IEEE International Conference on Robotics and Automation в Японии

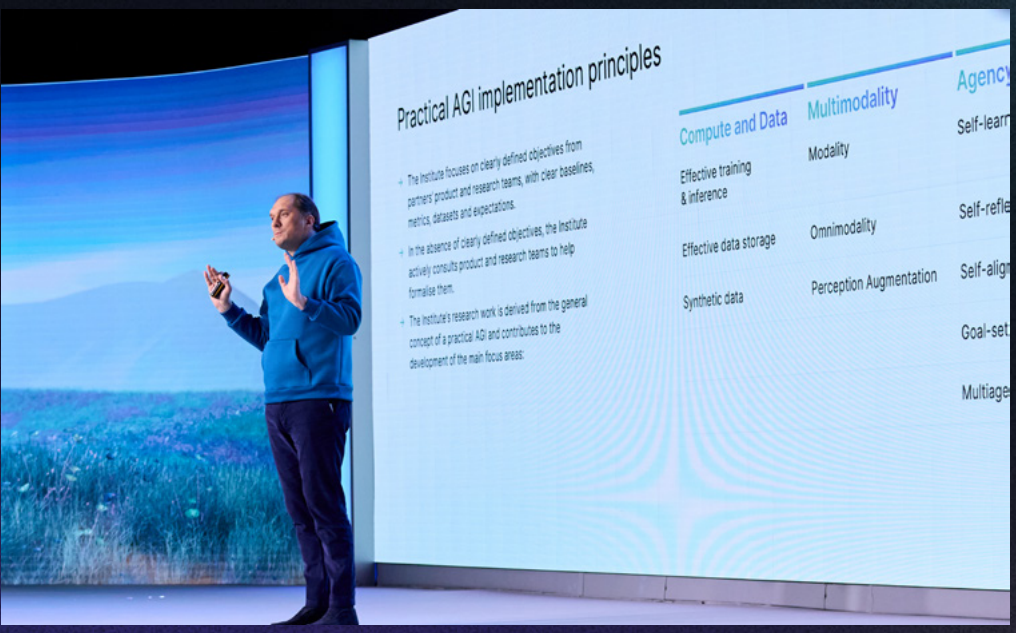


Воркшоп по доверенному искусственному интеллекту WAIT в Казахстане

AI Journey 2024

18

исследователей
выступили
на конференции
с докладами



AI Journey 2024

7

ученых
представили
свои постеры

Лето с AIRI 2024

Лето с AIRI

Летняя школа Института искусственного интеллекта AIRI для студентов и аспирантов — это глубокое погружение в работу с широким спектром современных методов искусственного интеллекта и машинного обучения.

Участники программы провели 10 дней с учеными из AIRI, ИТМО, МФТИ, ВШЭ, Сколтеха, Лаборатории искусственного интеллекта Сбера и других авторитетных научно-исследовательских организаций и вузов.

Исследователи, которые прошли отбор на Школу, смогли не только познакомиться с потенциальными научными руководителями, но и получить карьерные консультации. Также применили полученные знания в рамках практических семинаров. В финале программы они защитили 25 исследовательских проектов по поисковым и практикооорентирвоанным тематикам. Серди них: выявление аномалий электрических показаний промышленных электродвигателей, изучение состязательных атак на защищенные модели, дообучение мультимодальных моделей работе в узком домене, предсказание частот поглощения и испускания молекул с использованием большой языковой модели.

941
заявка

37
преподавателей

90
студентов

10
дней

25
проектов

45
лекций
и семинаров

80
научных
постеров

наука меняет мир

лето открытий

Партнеры проекта

Научные парнеры

Плейлист Reading Club | Embodied AI

Журнальные клубы

Reading Clubs AIRI — это пространство для увлекательных дискуссий и обмена идеями. На них исследователи обсуждают новые, важные или просто любопытные научные статьи. На встречах участники глубже вникают в суть материалов, оценивают их значимость для своей области и расширяют собственный исследовательский горизонт.

Reading Club | ControlGenA

Лидирует:
Айбек Аланов

32
встречи
за 2024 год

Reading Club | Embodied AI

Лидирует:
Алексей Ковалёв

10
встреч
за 2024 год

Reading Club | CV and Robotics

Лидирует:
Влад Шахуро

31
встреча
за 2024 год

AIRI LEGO клуб

В этом году неформальный клуб по интересам сотрудников AIRI продолжил развиваться. Коллеги собирали Lego, знакомились, приглашали друзей и играли в квизы.

Хакатон SafeSpeak-2024

Обнаружение дипфейков для обеспечения безопасности голосовой связи

SafeSpeak2024 — хакатон, посвященный разработке технологий обнаружения аудио-спуфинга, нацеленный на решение актуальных проблем безопасной голосовой аутентификации и защиту биометрических систем от атак.

>240

человек подали заявки на участие

Заявки на участие в соревновании подали молодые ученые из России, Эфиопии, Казахстана, Вьетнама, Бразилии, Китая, Узбекистана и Индии.

Хакатон «Типовой ИИ vs Базовые модели»

14 декабря состоялся финал хакатона «Типовой ИИ vs Базовые модели». За звание лучших сражались 23 команды из ведущих вузов России. Их главной задачей стало создание модели машинного обучения, способной автоматически обнаруживать патологию на рентгенограммах органов грудной клетки.

AI Journey Contest 2024

Исследователи AIRI подготовили 3 задачи на хакатон AI Journey Contest 2024:

- Emotional FusionBrain 4.0: создание мультимодальных моделей работы с видео, аудио и текстом.
- Multiagent AI: создание мультиагентной RL-системы, агенты которой смогут решать задачи, объединяясь в различные схемы кооперации.
- Embodied AI: создание роботов, которые смогут решать сложные задачи, требующие взаимодействия с окружающей средой и пользователем, а также общения с ними на естественном языке.

Night Photography Rendering Challenge 2024, NTIRE Workshop. CVPR 2024

Ученые AIRI и Института проблем передачи информации (ИППИ РАН) провели приуроченное к одной из наиболее престижных конференций по компьютерному зрению CVPR 2024 научное соревнование по рендерингу ночных фотографий. Его целью было с помощью ИИ-алгоритмов обработать ночное изображение с камеры смартфона и получить кадр фотографического качества.

58 команд приняли участие в соревновании, включая представителей академии (Гонконг, Сингапур, Вашингтон, Милан и др.) и индустрии (Samsung, XiaoMi, Honor, Doshua и др.).

О нас пишут и говорят

ТАСС, РИА, РБК, Forbes, Ведомости, КоммерсантЪ, Российская Газета, Газета. Ru, Lenta.ru, N+1, Известия, Хайтек, Первый Канал, Россия 24, BFM, Код Дурова, Радио «Маяк», Радио России, телеканал «Культура» и многие другие.

Прочесть экспертные материалы от сотрудников AIRI

Forbes
Иван Оселедец о том, как оценивать работу DeepTech-исследователей

Forbes
Антон Конушин о том, кого брать в штат ИИ-стартапа и на какие навыки смотреть

РБК Тренды
Александр Панов о том, как ученые используют игры, чтобы обучать системы искусственного интеллекта решать реальные задачи.

РБК Тренды
Юрий Куратов о том, как устроена память нейросетей

RB
Максим Кузнецов о том, как выбрать нейросеть для бизнеса

RB
Александра Бройтман о том, как организовывать летние школы

О нас пишут и говорят

N+1

Партнерский материал
о том, что производит наука
и как она помогает бизнесу

КоммерсантЪ

Евгений Фролов о том,
как исследователи
повысили точность
рекомендательных систем

КоммерсантЪ

Иван Оселедец для КоммерсантЪ
о том, как команда российских
математиков доработала выводы
нобелевского лауреата

ТАСС

Александр Панов о том,
что бизнес и потребители в
России выигрывают
от внедрения ИИ?

КоммерсантЪ

Илья Макаров о том,
что малые нейросети
способны обучать большие
ИИ-модели лучше человека

КоммерсантЪ

Айбек Аланов о том,
как редактирование
изображений поможет
науке

Новые партнерства за 2024 год

Ученые AIRI в социальных сетях

Ivan Oseledets' Channel

Telegram-канал Ивана Оселедца

The Oleg

Telegram-канал Олега Рогова

Kardymonchik Channel

Telegram-канал Ольги Кардымон

Complete AI

Telegram-канал Андрея Кузнецова

AbstractDL

Telegram-канал Антона Разжигаева

Grounding Knowledge

Telegram-канал Александра Панова

Dendi Math&AI

Telegram-канал Дениса Димитрова

Чердачок Ктошика

Telegram-канал Антона Конушина

AI Brain

Telegram-канал
Айбека Аланова

Causality links

Telegram-канал
Владислава Куренкова

Labrats

Telegram-канал
Дмитрия Пензара

ComputAgeChannel

Telegram-канал
Дмитрия Крюкова

Embrace the data

Telegram-канал
Аллы Чепуровой

Гречневые мысли

Telegram-канал
Никиты Сушко

BaseLine

Telegram-канал
Алексея Ковалева

Nygma's Diaries

Telegram-канал
Максима Жданова

Вениамин Фишман

Telegram-канал
Вениамина Фишмана

Марат пишет про науку

Telegram-канал
Марата Хамадеева

Контакты

Сайт

airi.net

Соцсети

airi_research_institute

Airi_institute

habr

artificial-intelligence-research-institute

AIRIInstitute

AIRI_inst

Адрес

Москва, Пресненская набережная, д. 6, стр. 2