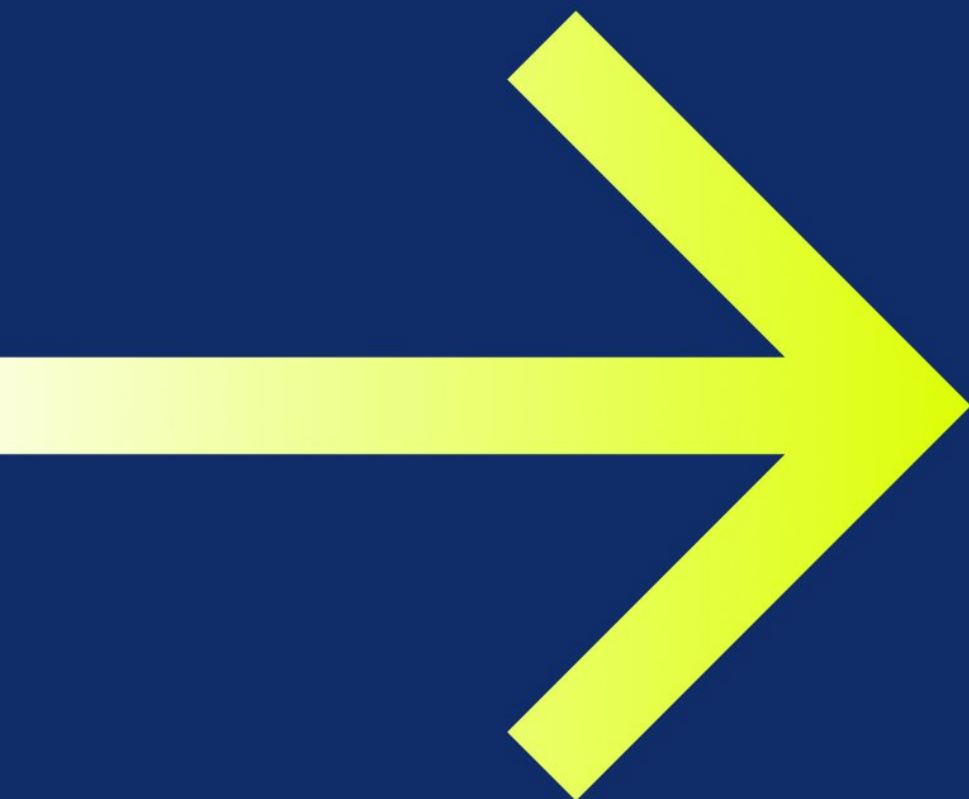


II Научная
конференция ФКН

27-30 октября
2024



ФКН

УЦ Вороново

Оглавление

Оглавление.....	2
Полезная информация.....	7
Контакты для оперативной связи.....	7
Место проведения.....	7
Схема проезда.....	8
Размещение.....	9
Программа конференции.....	10
Участники конференции.....	15
Приглашённые лекции.....	23
Рождение программной инженерии (Шилов В. В.).....	23
История цифровой информации от ткацкого станка до компьютера (Басов А. И.).....	23
Круглые столы и дискуссии.....	24
Круглый стол по программной инженерии «MLOps-платформа для развертывания и мониторинга интеллектуальных сервисов НИУ ВШЭ».....	24
Круглый стол по прикладной математике и информатике на ФКН.....	26
Дискуссия «Исследовательские работы: что мы хотим от студентов, что студенты хотят от нас?».....	28
Доклады основной программы.....	29
Сложность данных и простота классификаторов (<i>Кузнецов С. О.</i>).....	29
Возможности динамического полиморфизма при процедурно-параметрическом программировании (<i>Легалов А. И.</i>).....	30
Использование дискурсивного анализа для выявления манипулятивных приемов и умышленных искажений в текстовых данных (<i>Ильвовский Д. А.</i>).....	31
Выявление коннектома у глухих посредством решения обратной задачи (<i>Харламов А. А.</i>).....	32
Diffusion models for controllable image generation (<i>Аланов А.</i>).....	33
Подводные камни в бенчмарках табличного ML и новый бенчмарк TabReD (<i>Рубачёв И. В.</i>).....	34

Symbolic regression for properties prediction and defect structure generation in MoS ₂ and WSe ₂ (Лазарев М. В.).....	35
Инструмент на основе методов NLP для анализа журналов событий среды .NET CLR (Ставцев М. А.).....	36
Проект программно аппаратного комплекса для ухода за растениями с элементами ИИ (Прозорский М. А.).....	37
Анализ автоматического распознавания, генерации и интерпретации юмора (Баранов А. М.).....	38
Синтез иерархических моделей процессов: подход на основе разбиения событий на множества (Бегичева А. К.).....	39
Решение задачи непарного переноса стиля на основе дистилляции диффузионных моделей (Ракитин Д. Р.).....	40
Декодирование BOLD сигнала по данным ЭЭГ (Семенов И. В.).....	41
Genetic engineering algorithm (GEA) and genetic speciation in solving combinatorial optimization problems (Сохраби М.).....	42
Майнинг ациклических DFG моделей для ациклических процессов (Шамов Н. Д.).....	43
Анализ применения Web-фреймворков: React, Angular, VUE.JS и Svelte в проектах с различным уровнем вложенности компонентов (Караваяева Е. А.).....	44
MiSTi: метод оценки времени разделения и скорости миграции при переменном размере популяции (Щур В. Л.).....	45
ИИ и генетический компьютер (Полцова М. С.).....	46
Нейросетевое определение популяционной принадлежности человека по графу общих сегментов генома (Глызин Д. С.).....	47
Инновационные медицинские приложения: Проект EYAS для анализа фондус-изображений (Рябцев Д. И.).....	48
Исследование механизмов лекарственной устойчивости возбудителя туберкулеза к лекарственным препаратам с помощью моделей машинного обучения (Неверов А. Д.).....	49
Гибридизация методов машинного обучения и классических моделей (Деркач Д. А.).....	51
Применение диффузионных моделей для генерации текстур трёхмерных моделей (Струминский К. А.).....	52
Нормальная аппроксимация и мультипликативный бутстреп для алгоритмов стохастической аппроксимации (Самсонов С. В.).....	53
Автоматизированное обучение сверточных нейронных сетей (Болдырев А. С.).....	54
Эффективная структурированная параметризация ортогональных матриц (Рахуба М. В.).....	55
О преимуществах использования сходства данных в задачах обучения для сокращения числа коммуникаций в распределённых алгоритмах (Двинских Д. М.).....	56

Обнаружение разладки во временных рядах с помощью отслеживания лучших экспертных стратегий (<i>Пучкин Н. А.</i>).....	57
Формирование ленты новостей на основании технических событий с помощью технологий больших языковых моделей (<i>Пашигорев К. И.</i>).....	58
Архитектурный подход с одной точкой входа и выхода: использование Istio для управления микросервисами (<i>Кензин И. И.</i>).....	59
Типизированные сети Петри с переменными весами дуг (<i>Ломазова И. А.</i>).....	60
Двусторонний вывод с режимом применения для зависимых типов с неявными аргументами (<i>Соколов П. П.</i>).....	61
Дизъюнктная сложность (<i>Вялый М. Н.</i>).....	62
Моделирование асинхронного взаимодействия в мультиагентных системах на основе журналов событий (<i>Нестеров Р. А.</i>).....	63
Using LLMs for materials design (<i>Аль-Маини А.</i>).....	64
HiFi-Stream: решение для потокового усиления речи с помощью генеративных состязательных сетей (<i>Каледин М. Л.</i>).....	65
Heterogeneous GNN-based neutron reconstruction in the HGND at the BM@N experiment (<i>Бочарников В. О.</i>).....	66
Density of sphere packings (<i>Ферник Т.</i>).....	67
Introduction to visual analytics for explainable deep learning (<i>Джин С.</i>).....	68
PAC-bayesian learning meets synthetic Ricci curvature bounds (<i>Пари К.</i>).....	69
Lost in Translation: chemical language models and the misunderstanding of molecule structures (<i>Тутубалина Е. В.</i>).....	70
Поймай бота, или широкомасштабная структура естественного языка (<i>Громов В. А.</i>).....	71
Построение доверенного ИИ: проблемы, стратегии и перспективы (<i>Силаев Ю. В.</i>).....	72
О перспективных направлениях исследований методов изучения человеко-машинного взаимодействия в области разработки компьютерных игр (<i>Поздняков Д. А.</i>).....	73
Пучки на графах и рекомендательные системы (<i>Тараканов А. А.</i>).....	74
Сегментация и идентификация кузовных деталей легкового автомобиля POLO VW 5 (<i>Пантюхин Д. В.</i>).....	75
Towards reliable false discovery rate control in classification problems under distribution shift (<i>Боревский А. О.</i>).....	77
Improving generative flow networks through reinforcement learning (<i>Морозов Н. В.</i>).....	78
Особенности применения архитектурных паттернов в мобильной разработке (<i>Абросов С. А.</i>).....	79

Плакаты постерной сессии	80
Sequence variation identification in high resolution MS/MS data with accurate FDR control (<i>Бхимани К.</i>)	81
Гомотопические группы некоторых торических многообразий (<i>Вылегжанин Ф. Е.</i>)	82
Tight and efficient upper bound on spectral norm of convolutional layers (<i>Гришина Е. Р.</i>)	83
Применение Apache Kafka в высоконагруженных приложениях (<i>Данилов А. А.</i>)	84
Large-scale evaluation of the generalization performance of deep neural networks in tandem mass spectrometry data analysis (<i>Джоши К.</i>)	85
Guide-and-Rescale: self-guidance mechanism for effective tuning-free real image editing (<i>Иванова А. А.</i>)	86
Spectrum annotation with transformer models (<i>Кривулец Н. В.</i>)	87
Подход к созданию сервиса генерации программного кода мобильных приложений (<i>Резуник Л.</i>)	88
False discovery rate control in classification problems with knock-off classes (<i>Ромашкина А. Д.</i>)	89
Принципы формирования трехмерных изображений или количественная фазовая визуализация (<i>Талайкова Н. А.</i>)	90
Group and Shuffle: efficient structured orthogonal parametrization (<i>Юдин Н. Е.</i>)	91
Knowledge graph completion with mixed geometry tensor factorization (<i>Юсупов В. А.</i>)	92
Организаторы	93

С 27 по 30 октября 2024 года в учебном центре «Вороново» НИУ ВШЭ прошла II-я выездная научная конференция Факультета компьютерных наук (сокращённо – НКФКН2024). На конференции исследователи факультета представили основные научные результаты и достижения подразделений ФКН за последний год работы. Среди тем конференции: машинное обучение и анализ данных, программная инженерия, теоретическая информатика, биоинформатика и медицинские информационные системы, применение методов машинного обучения в физике, химии, биологии, промышленности, обработка естественного языка, проектирование человеко-машинного взаимодействия и разработка компьютерных игр, системное программирование, анализ процессов, управление сложными системами и другие.

К участию в работе конференции были приглашены преподаватели и научные сотрудники факультета, а также аспиранты и студенты, готовые представить результаты своей исследовательской работы.

Цель научной конференции ФКН – собрать вместе коллег, которые заняты весьма разными темами. Задача конференции скорее не глубоко погрузиться в конкретную узкую область информатики, но рассказать, исследования по каким темам проводятся на факультете, показать коллегам значение и красоту каждого из направлений. По этой причине программа конференции состоит из одинаковых коротких докладов продолжительностью по 20 минут – 15 минут на рассказ и 5 минут на вопросы и комментарии. Такой сжатый формат позволил сделать мероприятие динамичным и интересным для широкой аудитории.

Страница конференции: https://cs.hse.ru/sci_conf2024

Темы и аннотации докладов, круглых столов, плакаты приводятся в авторской редакции.

Полезная информация

Контакты для оперативной связи

Андрей Игнатов aignatov@hse.ru Тг: [@a_whynot](https://t.me/@a_whynot)

Ксения Кузнецова ktupikina@hse.ru

Место проведения

Конференция пройдёт в УЦ Вороново НИУ ВШЭ [\[📍\]](#)



Адрес: 108 830, г. Москва, поселение Вороновское, с. Вороново, ул. Канторовича, домовладение 1

Схема проезда

Общественный транспорт

Чтобы самостоятельно добраться до Учебного Центра вы можете воспользоваться маршрутным транспортом:

от м. Тёплый Стан (выход из последнего вагона из центра, в подземном переходе — 2-й выход налево) **Автобус № 503** или **маршрутное такси № 887** (остановки находятся в 30-40 метрах от подземного выхода метро возле ТРЦ «Принц Плаза»; ориентир: справа от остановок должен быть фонтан-водопад).
Время в пути: автобус — от 1 ч. до 1 ч. 15 мин., маршрутное такси — 50-60 мин. (без учёта пробок).

от м. Ольховая автобус № 508

Выйти на остановке **Вороново-2**, предварительно заказав микроавтобус из УЦ «Вороново» (это можно сделать к определенному времени, или по прибытию на остановку, позвонив дежурному администратору УЦ "Вороново" по телефону +7 (495) 916-89-16, +7 (495) 772-95-90 доб. 17010. Доставка микроавтобусом от автобусной остановки Вороново до УЦ по рабочим дням с 9.00 до 18.00.

На автомобиле

Координаты для GPS-навигатора:

широта: 55°18'54.13"N (55.315035)

долгота: 37°7'25.77"E (37.123825)

Вы едете со стороны МКАД по Калужскому шоссе в сторону области до села Вороново, поворачиваете на первом светофоре направо по дорожному указателю, двигаетесь 2,8 км до поворота налево по указателю.

Трек: <http://maps.yandex.ru/-/CVCgECOА>

Размещение

Карточка участника: Каждый участник при прибытии в УЦ Вороново получил набор участника, в котором, среди прочего, была карточка (бейдж) с именем, ручка и блокнот для записей.

Доступ в интернет: В УЦ Вороново имеется Wi-Fi. Данные для входа указаны в книгах постояльцев в каждом номере. Также работает корпоративная Wi-Fi сеть НИУ ВШЭ.

Еда: В ходе конференции организовано трёхразовое питание (завтрак, обед и ужин), а также перерыв на кофе в вечерней сессии докладов. Всё питание организовано в столовой УЦ Вороново. Таким образом, участникам не нужно думать про поиск еды. Имеет смысл иметь с собой ёмкость, чтобы набрать в неё питьевой воды. Системы выдачи воды присутствуют на каждом этаже УЦ.

Программа конференции

День 1 (27 октября, воскресенье)

15:30	Трансфер от кампуса на Покровском бульваре до УЦ Вороново	Покровский бул., 11 вход 3
18:00	Заезд в УЦ Вороново. Размещение.	
18:00	Ужин	Столовая
19:00	Приглашённая лекция профессора ДПИ к.т.н. Шилова В.В. «Рождение программной инженерии»	Большой зал
20:30	Просмотр фильма (по желанию)	Ауд. 1, 6 этаж
	Настольные игры (по желанию)	Ауд. 2, 6 этаж

День 2 (28 октября, понедельник)

9:00	Завтрак	Столовая
9:50	Открытие конференции	Большой зал
10:00	1 Сложность данных и простота классификаторов Кузнецов С.О. руководитель ДАДИИ, профессор, д.ф.-м.н.	
10:20	2 Возможности динамического полиморфизма при процедурно-параметрическом программировании Легалов А.И. профессор ДПИ, д.т.н.	
10:40	3 Использование дискурсивного анализа для выявления манипулятивных приемов и умышленных искажений в текстовых данных Ильвовский Д.А. доцент ДАДИИ, к.т.н.	
11:00	4 Выявление коннектома у глухих посредством решения обратной задачи Харламов А.А. профессор ДПИ, д.т.н.	
11:20	5 Diffusion models for controllable image generation Аланов А. руководитель ЦБОГ М	
11:40	6 Подводные камни в бенчмарках табличного ML и новый бенчмарк TabReD Рубачёв И.В. м.н.с. НУЛ Яндекс	
12:00	7 Symbolic regression for properties prediction and defect structure generation in MoS2 and WSe2 Лазарев М.В. доцент ДБДИП, PhD	
12:20	8 Инструмент на основе методов NLP для анализа журналов событий среды .NET CLR Ставцев М.А. стажёр-исследователь НУЛ ПОИС, Шершаков С.А. доцент ДПИ, к.комп.н.	
12:40	9 Проект программно аппаратного комплекса для ухода за растениями с элементами ИИ Прозорский М.А. стажёр-исследователь НУЛ ОиМТ, Александров Д.В. профессор ДПИ, д.т.н., Резунки Л. стажёр-исследователь НУЛ ОиМТ	

13:00	Обед	Столовая
14:00	Анализ автоматического распознавания, генерации и интерпретации юмора Баранов А.М. стажёр-исследователь НУЛ ММВП	Большой зал
14:20	Синтез иерархических моделей процессов: подход на основе разбиения событий на множества Бегичева А.К. стажёр-исследователь НУЛ ПОИС	
14:40	Решение задачи непарного переноса стиля на основе дистилляции диффузионных моделей Ракитин Д.Р. м.н.с. ЦГОБМ	
15:00	Декодирование BOLD сигнала по данным ЭЭГ Семенов И.В. аспирант ДАДИИ	
15:20	Genetic engineering algorithm (GEA) and genetic speciation in solving combinatorial optimization problems Сохраби М. стажёр-исследователь НУЛ ММВП	
15:40	Майнинг ациклических DFG моделей для ациклических процессов Шаимов Н.Д. стажёр-исследователь НУЛ ПОИС	
16:00	Анализ применения Web-фреймворков: React, Angular, VUE.JS и Svelte в проектах с различным уровнем вложенности компонентов Караваева Е.А., Думкин Н.А. стажёры-исследователи НУЛ ОиМТ	
16:30	Перерыв на кофе	Столовая
17:00	MiSTi: метод оценки времени разделения и скорости миграции при переменном размере популяции Щур В.Л. профессор ДБДИП, д.ф.-м.н.	Большой зал
17:20	ИИ и генетический компьютер Попцова М.С. доцент ДБДИП, к.ф.-м.н.	
17:40	Нейросетевое определение популяционной принадлежности человека по графу общих сегментов генома Глызин Д.С. н.с. МЛ СВГ, к.ф.-м.н.	
18:00	Инновационные медицинские приложения: Проект EYAS для анализа фондус-изображений Рябцев Д.И. доцент ДПИ, к.м.н.	
18:20	Исследование механизмов лекарственной устойчивости возбудителя туберкулеза к лекарственным препаратам с помощью моделей машинного обучения Неверов А.Д. в.н.с. МЛ СВГ, к.б.н.	
18:40	Ужин	Столовая
19:30	Круглый стол по программной инженерии «MLOps-платформа для развертывания и мониторинга интеллектуальных сервисов НИУ ВШЭ» Модератор: Лебедев С.А., к.э.н., руководитель ДПИ	Большой зал
21:00	Просмотр фильма (по желанию)	Ауд. 1, 6 этаж
	Дискуссия «Исследовательские работы: что мы хотим от студентов, что студенты хотят от нас?» Модератор: Ратников Ф.Д., к.ф.-м.н., доцент ДБДИП	Ауд. 2, 6 этаж

День 3 (29 октября, вторник)

9:00	Завтрак		Столовая
10:00	22	Гибридизация методов машинного обучения и классических моделей Деркач Д.А. директор по прикладным исследованиям и разработкам ИИИиЦН, доцент ДБДИП, Ph.D.	Большой зал
10:20	23	Применение диффузионных моделей для генерации текстур трёхмерных моделей Струминский К.А. стажёр-исследователь НУЛ компании Яндекс, к.комп.н.	
10:40	24	Нормальная аппроксимация и мультипликативный бутстреп для алгоритмов стохастической аппроксимации Самсонов С.В. н.с. МЛ СААМД	
11:00	25	Автоматизированное обучение сверточных нейронных сетей Болдырев А.С. доцент ДБДИП, к.ф.-м.н.	
11:20	26	Эффективная структурированная параметризация ортогональных матриц Рахуба М.В. доцент ДБДИП, к.ф.-м.н.	
11:40	27	О преимуществах использования сходства данных в задачах обучения для сокращения числа коммуникаций в распределённых алгоритмах Двинских Д.М. доцент БК ИППИ РАН, Ph.D.	
12:00	28	Обнаружение разладки во временных рядах с помощью отслеживания лучших экспертных стратегий Пучкин Н.А. доцент БК ИППИ РАН, к.комп.н.	
12:20	29	Формирование ленты новостей на основании технических событий с помощью технологий больших языковых моделей Пашигорев К. И. пригл. преп. БК ПАО Сбербанк, руководитель направления Сбер	
12:40	30	Архитектурный подход с одной точкой входа и выхода: использование Istio для управления микросервисами Кензин И. И. пригл. преп. БК ПАО Сбербанк, ведущий инженер Сбер	
13:00	Обед		Столовая
13:45	Общее фото участников конференции		Атриум
14:00	31	Типизированные сети Петри с переменными весами дуг Ломазова И.А. профессор ДПИ, д.ф.-м.н.	Большой зал
14:20	32	Двусторонний вывод с режимом применения для зависимых типов с неявными аргументами Соколов П.П. преподаватель ДБДИП	
14:40	33	Дизъюнктивная сложность Вялый М.Н. профессор ДБДИП, к.ф.-м.н.	
15:00	34	Моделирование асинхронного взаимодействия в мультиагентных системах на основе журналов событий Нестеров Р.А. доцент ДПИ, к.комп.н.	
15:20	35	Using LLMs for materials design Аль-Маини А. м.н.с. НУЛ МАБД	

15:40	36	HiFi-Stream: решение для потокового усиления речи с помощью генеративных состязательных сетей Каледин М.Л. доцент ДБДИП, к.комп.н.	
16:00	37	Heterogeneous GNN-based neutron reconstruction in the HGND at the BM@N experiment Бочарников В.О. м.н.с. НУЛ МАБД	
16:30		Перерыв на кофе	Столовая
16:30		Постерная сессия (список плакатов приводится далее)	Холл у большого зала
17:40	38	Density of sphere packings Ферник Т. доцент ДАДИИ, Ph.D.	Большой зал
18:00	39	Introduction to visual analytics for explainable deep learning Джин С. доцент ДБИ ВШБ, к.комп.н.	
18:20	40	PAC-bayesian learning meets synthetic Ricci curvature bounds Пари К. доцент ДАДИИ, Ph.D.	
18:40		Ужин	Столовая
19:30		Круглый стол по прикладной математике и информатике на ФКН Модератор: Соколов Е.А., руководитель ДБДИП	Большой зал
21:00		Просмотр фильма (по желанию)	Ауд. 1, 6 этаж
		Приглашённая лекция редактора ФКН и популяризатора истории техники Басова А.И. «История цифровой информации от ткацкого станка до компьютера»	Ауд. 2, 6 этаж
День 4 (30 октября, среда)			
9:00		Завтрак	Столовая
10:00	41	Lost in Translation: chemical language models and the misunderstanding of molecule structures Тутубалина Е.В. н.с. НУЛ ММВП, д.комп.н., исследователь AI/RI	Большой зал
10:20	42	Поймай бота, или широкомасштабная структура естественного языка Громов В.А. профессор ДАДИИ, д.ф.-м.н.	
10:40	43	Построение доверенного ИИ: проблемы, стратегии и перспективы Силаев Ю.В. ст. преподаватель ДПИ	
11:00	44	О перспективных направлениях исследований методов изучения человеко-машинного взаимодействия в области разработки компьютерных игр Поздняков Д.А. преподаватель ДПИ, совладелец и продуктовый директор студии «Винторог»	
11:20	45	Пучки на графах и рекомендательные системы Тараканов А.А. доцент ДБДИП, Ph.D., Ермилов С.И. разработчик VK	

11:40	46	Сегментация и идентификация кузовных деталей легкового автомобиля POLO VW 5 Пантюхин Д.В. ст. преподаватель ДПИ	
12:00	47	Towards reliable false discovery rate control in classification problems under distribution shift Боревский А.О. ассистент ДАДИИ	
12:20	48	Improving generative flow networks through reinforcement learning Морозов Н.В. стажёр-исследователь ЦГОБМ	
12:40	49	Особенности применения архитектурных паттернов в мобильной разработке Абросов С.А. стажёр-исследователь НУЛ ОиМТ, Сосновский Г.М. ведущий разработчик компании Звук	
13:00		Обед	Столовая
14:00		Выезд из УЦ Вороново до 16:00	
14:30		Трансфер из УЦ Вороново до кампуса на Покровском бульваре	Вход в УЦ

Участники конференции

Полный список участников научной конференции ФКН в 2024 году:

Участник	Должность	Подразделение
1 Абдуллаева Сабина	администратор	Отдел сопровождения проектов
2 Абросов Сергей	стажер-исследователь	Научно-учебная лаборатория облачных и мобильных технологий
3 Агамирзян Игорь	профессор	Департамент программной инженерии
4 Аланов Айбек	заведующий центром	Центр глубинного обучения и байесовских методов
5 Аланова Ширин	студент	ФКН
6 Александров Дмитрий	профессор	Департамент программной инженерии
7 Алиев Мишан	стажер-исследователь	Центр глубинного обучения и байесовских методов
8 Аль-Маини Абдалазиз	младший научный сотрудник	Научно-учебная лаборатория методов анализа больших данных
9 Аржанцев Иван	профессор	Департамент больших данных и информационного поиска
10 Баранков Даниил	студент	Отдел коммуникаций , команда ЧатФКН
11 Баранов Александр	стажер-исследователь	Научно-учебная лаборатория моделей и методов вычислительной прагматики
12 Басов Антон	редактор	Отдел коммуникаций ФКН
13 Бауенс Бруно	доцент	Департамент больших данных и информационного поиска
14 Бегичева Антонина	преподаватель	Департамент программной инженерии
15 Бекетов Максим	стажер-исследователь	Международная лаборатория стохастических алгоритмов и анализа многомерных данных
16 Белова Наталья	доцент	Департамент программной инженерии
17 Бобков Денис	стажер-исследователь	Центр глубинного обучения и байесовских методов
18 Болдырев Алексей	научный сотрудник	Научно-учебная лаборатория методов анализа больших данных
19 Боровский Андрей	ассистент	Департамент анализа данных и искусственного интеллекта
20 Бочарников Владимир	младший научный сотрудник	Научно-учебная лаборатория методов анализа больших данных
21 Брейман Александр	доцент	Департамент программной инженерии
22 Бхимани Кишанкумар	стажер-исследователь	Научно-учебная лаборатория искусственного интеллекта для вычислительной биологии

Участник	Должность	Подразделение
23 Васильева Анастасия	студент	ФКН
24 Василева Анна	студент	ФКН
25 Виденин Сергей	доцент	Департамент программной инженерии
26 Власенко Галина	стажер-исследователь	Международная лаборатория статистической и вычислительной геномики
27 Воротников Игорь	стажер-исследователь	Международная лаборатория теоретической информатики
28 Вылегжанин Федор	стажер-исследователь	Международная лаборатория алгебраической топологии и приложений
29 Вялый Михаил	профессор	Международная лаборатория теоретической информатики
30 Гапонов Александр	студент	ФКН
31 Гаянов Никита	стажер-исследователь	Международная лаборатория статистической и вычислительной геномики
32 Глызин Дмитрий	научный сотрудник	Международная лаборатория статистической и вычислительной геномики
33 Голобородько Ирина	стажер-исследователь	Научно-учебная лаборатория матричных и тензорных методов в машинном обучении
34 Горбач Марина	студент	4 курс ПМИ ФКН
35 Горденко Мария	старший преподаватель	Департамент больших данных и информационного поиска
36 Грибов Филипп	преподаватель	Департамент больших данных и информационного поиска
37 Гришина Екатерина	стажер-исследователь	Научно-учебная лаборатория матричных и тензорных методов в машинном обучении
38 Громов Василий	профессор	Департамент анализа данных и искусственного интеллекта
39 Грузинцев Егор	студент	Научный куратор ФКН
40 Данилов Алексей	младший научный сотрудник	Научно-учебная лаборатория облачных и мобильных технологий
41 Двинских Дарина	доцент	Департамент больших данных и информационного поиска
42 Деркач Денис	доцент	Департамент больших данных и информационного поиска
43 Джин Сеунгмин	преподаватель	Департамент анализа данных и искусственного интеллекта
44 Джоши Картик	стажер-исследователь	Научно-учебная лаборатория искусственного интеллекта для вычислительной биологии
45 Дрибноход Евгений	студент	4 курс ПИ ФКН
46 Дробышевский Илья	студент	4 курс ПМИ ФКН

Участник	Должность	Подразделение
47 Дудковская Анастасия	стажер-исследователь	Международная лаборатория статистической и вычислительной геномики
48 Думкин Никита	студент	Студент ФКН
49 Дьячкова Наталия	стажер-исследователь	Международная лаборатория статистической и вычислительной геномики
50 Ермилов Сергей	разработчик	ВК
51 Жукова Галина	доцент	Департамент программной инженерии
52 Зароднюк Алёна	доцент	Департамент больших данных и информационного поиска
53 Ибрахим Анвар	стажер-исследователь	Научно-учебная лаборатория методов анализа больших данных
54 Иванова Александра	стажер-исследователь	Лаборатория теоретических основ моделей искусственного интеллекта
55 Игнатов Андрей	преподаватель	Департамент программной инженерии
56 Ильвовский Дмитрий	доцент	Департамент анализа данных и искусственного интеллекта
57 Каледин Максим	доцент	Департамент больших данных и информационного поиска
58 Каленбет Ульяна	студент	Научный куратор ФКН
59 Каравеева Екатерина	стажер-исследователь	Научно-учебная лаборатория облачных и мобильных технологий
60 Карпов Максим	старший преподаватель	Департамент больших данных и информационного поиска
61 Карпова Анна	студент	3 курс ПМИ ФКН
62 Качмазов Руслан	студент	3 курс ПИ ФКН
63 Кензин Игорь	приглашённый преподаватель	Базовая кафедра ПАО Сбербанк «Финансовые технологии и анализ данных» / ведущий инженер Сбер
64 Кертес-Фаркаш Аттила	доцент	Департамент анализа данных и искусственного интеллекта
65 Кирова Валерия	научный сотрудник	Научно-учебная лаборатория методов анализа больших данных
66 Кодрян Максим	младший научный сотрудник	Центр глубинного обучения и байесовских методов
67 Кожина Елена	заместитель проректора	НИУ ВШЭ
68 Козлова Мария	менеджер	Отдел сопровождения проектов
69 Коновалов Дмитрий	стажер-исследователь	Международная лаборатория биоинформатики
70 Красноженов Григорий	эксперт	Научно-учебная лаборатория методов анализа больших данных
71 Кривулец Нина	стажер	Научно-учебная лаборатория искусственного интеллекта для вычислительной биологии

Участник	Должность	Подразделение
72 Кузнецов Сергей	профессор	Департамент анализа данных и искусственного интеллекта
73 Кузнецова Ксения	менеджер	Научно-учебная лаборатория методов анализа больших данных
74 Курбатов Евгений	младший научный сотрудник	Научно-учебная лаборатория методов анализа больших данных
75 Куренков Владимир	старший преподаватель	Департамент больших данных и информационного поиска
76 Куцаков Александр	студент	3 курс ПМИ ФКН
77 Лазарев Михаил	доцент	Департамент больших данных и информационного поиска
78 Лебедев Сергей	доцент	Департамент программной инженерии
79 Легалов Александр	профессор	Департамент программной инженерии
80 Лесовская Ирина	доцент	Департамент программной инженерии
81 Ломазова Ирина	профессор	Департамент программной инженерии
82 Лялин Дмитрий	преподаватель	Базовая кафедра МТС
83 Максаев Артем	доцент	Департамент больших данных и информационного поиска
84 Максименкова Ольга	доцент	Департамент программной инженерии
85 Мамай Игорь	доцент	Департамент больших данных и информационного поиска
86 Мануйленко Никита	преподаватель	Базовая кафедра Т-банка
87 Масютин Алексей	директор	Институт искусственного интеллекта и цифровых наук
88 Матвеева Татьяна	стажер-исследователь	Научно-учебная лаборатория матричных и тензорных методов в машинном обучении
89 Мирный Игорь	стажер-исследователь	Международная лаборатория статистической и вычислительной геномики
90 Миронов Михаил	стажер-исследователь	Проектно-учебная лаборатория «Искусственный интеллект в математических финансах»
91 Мицюк Алексей	доцент	Департамент программной инженерии
92 Моложавенко Александр	стажер-исследователь	Научно-учебная лаборатория матричных и тензорных методов в машинном обучении
93 Мореано Брандон	стажер-исследователь	Международная лаборатория биоинформатики
94 Морозов Никита	стажер-исследователь	Центр глубинного обучения и байесовских методов
95 Неверов Алексей	ведущий научный сотрудник	Международная лаборатория статистической и вычислительной геномики
96 Нестеров Роман	доцент	Департамент программной инженерии

Участник	Должность	Подразделение
97 Неудачина Ева	стажер-исследователь	Центр глубинного обучения и байесовских методов
98 Николаев Максим	стажер-исследователь	Центр глубинного обучения и байесовских методов
99 Никулов Сергей	координатор образовательных проектов	Т-Банк
100 Нургаязова Диляра	начальник отдела	Отдел коммуникаций, команда ЧатФКН
101 Охотин Андрей	стажер-исследователь	Центр глубинного обучения и байесовских методов
102 Павлова Анна	стажер-исследователь	Проектно-учебная лаборатория «Искусственный интеллект в математических финансах»
103 Павловчев Николай	старший преподаватель	Департамент программной инженерии
104 Пантюхин Дмитрий	старший преподаватель	Департамент программной инженерии
105 Пари Кентан	доцент	Департамент анализа данных и искусственного интеллекта
106 Паринов Андрей	старший преподаватель	Департамент анализа данных и искусственного интеллекта
107 Пашигорев Кирилл	приглашённый преподаватель	Базовая кафедра ПАО Сбербанк «Финансовые технологии и анализ данных» / руководитель направления Сбер
108 Перевышина Татьяна	преподаватель	Департамент больших данных и информационного поиска
109 Першина Анна	стажер-исследователь	Научно-учебная лаборатория методов анализа больших данных
110 Поздняков Денис	старший преподаватель	Департамент программной инженерии
111 Попов Виктор	заведующий лабораторией	Научно-учебная лаборатория моделирования и управления сложными системами
112 Попова Анфиса	научный сотрудник	Международная лаборатория статистической и вычислительной геномики
113 Попцова Мария	заведующая лабораторией	Международная лаборатория биоинформатики
114 Прозорский Михаил	стажер-исследователь	Научно-учебная лаборатория облачных и мобильных технологий
115 Промыслов Валентин	старший преподаватель	Департамент больших данных и информационного поиска
116 Пучкин Никита	старший научный сотрудник	Международная лаборатория стохастических алгоритмов и анализа многомерных данных
117 Ракитин Денис	младший научный сотрудник	Центр глубинного обучения и байесовских методов
118 Ратников Федор	ведущий научный сотрудник	Научно-учебная лаборатория методов анализа больших данных

Участник	Должность	Подразделение
119 Рахуба Максим	доцент	Департамент больших данных и информационного поиска
120 Резников Андрей	главный разработчик	Сбер
121 Резуник Людмила	ассистент	Департамент программной инженерии
122 Ровняго Дмитрий	студент	Научный куратор ФКН
123 Розаева Мария	студент	Научно-учебная лаборатория матричных и тензорных методов в машинном обучении
124 Ромашкина Арина	стажер-исследователь	Научно-учебная лаборатория искусственного интеллекта для вычислительной биологии
125 Рубачёв Иван	младший научный сотрудник	Научно-учебная лаборатория компании Яндекс
126 Рябцев Дмитрий	доцент	Департамент программной инженерии
127 Ряполов Денис	стажер-исследователь	Научно-учебная лаборатория матричных и тензорных методов в машинном обучении
128 Салех Хади	доцент	Департамент программной инженерии
129 Самилык Анастасия	студент	Команда ЧатФКН
130 Самойлова Анастасия	сотрудник	Команда ЧатФКН
131 Самоненко Илья	доцент	Департамент больших данных и информационного поиска
132 Самсонов Сергей	научный сотрудник	Международная лаборатория стохастических алгоритмов и анализа многомерных данных
133 Саночкин Юрий	научный сотрудник	Научно-учебная лаборатория моделирования и управления сложными системами
134 Сарадва Хушбу	стажер-исследователь	Научно-учебная лаборатория искусственного интеллекта для вычислительной биологии
135 Сафонов Никита	администратор	Отдел внешних коммуникаций
136 Сvirщевский Юрий	стажер-исследователь	Международная лаборатория стохастических алгоритмов и анализа многомерных данных
137 Семенов Илья	аспирант	Департамент анализа данных и искусственного интеллекта
138 Силаев Юрий	старший преподаватель	Департамент программной инженерии
139 Скворцов Владимир	приглашённый преподаватель	Департамент программной инженерии / аналитик Сбер
140 Соколов Евгений	доцент	Департамент больших данных и информационного поиска
141 Соколов Павел	ассистент	Департамент больших данных и информационного поиска
142 Сорокин Дмитрий	стажер-исследователь	Лаборатория теоретических основ моделей искусственного интеллекта
143 Сосновский Григорий	приглашённый преподаватель	Департамент программной инженерии / ведущий разработчик компании Звук

Участник	Должность	Подразделение
144 Сохраби Маджид	ассистент	Департамент анализа данных и искусственного интеллекта
145 Спирин Вадим	стажер-исследователь	Международная лаборатория статистической и вычислительной геномики
146 Ставцев Максим	стажер-исследователь	Научно-учебная лаборатория процессно-ориентированных информационных систем
147 Степин Сергей	студент	3 курс ПМИ ФКН
148 Струминский Кирилл	стажер-исследователь	Научно-учебная лаборатория компании Яндекс
149 Судаков Илья	стажер-исследователь	Международная лаборатория стохастических алгоритмов и анализа многомерных данных
150 Сулейкин Александр	научный сотрудник	Научно-учебная лаборатория процессно-ориентированных информационных систем
151 Талайкова Наталья	менеджер	Научно-учебная лаборатория методов анализа больших данных
152 Тараканов Александр	доцент	Департамент больших данных и информационного поиска
153 Терлыч Никита	преподаватель	Базовая кафедра фирмы 1С
154 Толстиков Алексей	Академический руководитель	Школа анализа данных компании Яндекс
155 Тутубалина Елена	научный сотрудник	Научно-учебная лаборатория моделей и методов вычислительной прагматики
156 Ферник Тома	доцент	Департамент анализа данных и искусственного интеллекта
157 Хамраева Мария	сотрудник	Команда ЧатФКН
158 Харламов Александр	профессор	Департамент программной инженерии
159 Хендави Баян	стажер-исследователь	Научно-учебная лаборатория моделирования и управления сложными системами
160 Хританков Антон	доцент	Департамент программной инженерии
161 Чельшева Нина	студент	4 курс ПМИ ФКН
162 Шаимов Никита	стажер-исследователь	Научно-учебная лаборатория процессно-ориентированных информационных систем
163 Шаламова Алиса	студент	3 курс ПИ ФКН
164 Шершаков Сергей	доцент	Департамент программной инженерии
165 Шилов Валерий	профессор	Департамент программной инженерии
166 Шишкин Михаил	младший научный сотрудник	Международная лаборатория статистической и вычислительной геномики
167 Щеникова Анна	аналитик	Центр R&D MTS Digital
168 Щербакова Елизавета	разработчик	Яндекс / выпускница ПИ ФКН
169 Щукин Владислав	приглашённый преподаватель	Департамент программной инженерии

Участник	Должность	Подразделение
170 Щур Владимир	заведующий лабораторией	Международная лаборатория статистической и вычислительной геномики
171 Юдин Николай	студент	Научно-учебная лаборатория матричных и тензорных методов в машинном обучении
172 Юсупов Вячеслав	стажер-исследователь	Научно-учебная лаборатория матричных и тензорных методов в машинном обучении
173 Ядав Нанд	научный сотрудник	Научно-учебная лаборатория методов анализа больших данных

Версия списка: 27 октября 2024

Приглашённые лекции

Рождение программной инженерии



Приглашённая лекция профессора ДПИ
В. В. Шилова [✉], к.т.н.

В докладе прослеживается возникновение идей, которые привели к формированию в середине XIX в. концепции программного управления техническими устройствами, ее дальнейшее развитие

применительно к вычислительным машинам вплоть до появления первых релейных и электронных компьютеров, ранних средств и методов программирования. Показано, как последующий взрывной рост возможностей компьютеров привел к пониманию необходимости выработки новых подходов к созданию программного обеспечения и, в итоге, к формированию программной инженерии.

История цифровой информации от ткацкого станка до компьютера



Приглашённая лекция редактора ФКН и популяризатора истории техники **А. И. Басова** [✉]

Для нас понятия цифровой и компьютерный неотделимы друг от друга. Между тем, так было не всегда — цифровые технологии зародились еще два века назад. В рамках лекции узнаем:

- что общего у ткацкого станка и компьютера,
- можно ли записать эмодзи на перфоленту,
- почему цифровая звукозапись появилась раньше аналоговой.

Круглые столы и дискуссии

В рамках II научной конференции ФКН запланировано проведение двух круглых столов, а также свободная дискуссия о научной работе студентов и со студентами для заинтересованных и небезразличных участников конференции. Важно, что в рамках обоих круглых столов затрагиваются темы, связанные как с разработкой программных решений, так и с разработкой, использованием и внедрением алгоритмов и методов машинного обучения. Таким образом, круглые столы объединяют направления исследований факультета и доклады по различным проблемам науки, представленные на конференции.

Круглый стол по программной инженерии «MLOps-платформа для развертывания и мониторинга интеллектуальных сервисов НИУ ВШЭ»



Модератор:

Лебедев С. А. [✉] руководитель ДПИ, к.э.н.

На круглом столе предлагается рассказать широкому кругу представителей ФКН и гостям конференции о проекте «MLOps-платформа для развертывания и мониторинга интеллектуальных сервисов НИУ ВШЭ», который реализуется в рамках стратегического проекта НИУ ВШЭ «ИИ-технологии для человека».

В формате вопросов-ответов предлагается рассказать о цели проекта «MLOps-платформа для развертывания и мониторинга интеллектуальных сервисов НИУ ВШЭ», о ключевых свойствах и ценностях создаваемой системы, о запланированных и реализованных задачах, о возможностях создаваемой системы для администрации, исследователей, преподавателей, студентов, внешних пользователей,

об архитектурных особенностях реализуемой MLOps-платформы на базе фреймворка Центра ИИ, о планах по развитию и дальнейшей поддержке платформы.

Участники круглого стола:



Масютин А. А. [📧] директор института искусственного интеллекта и цифровых наук ФКН, научный руководитель стратегического проекта «ИИ-технологии для человека» [📧](СП4), к.комп.н.



Кожина Е. С. [📧] координатор стратегического проекта «ИИ-технологии для человека» (СП4), заместитель проректора НИУ ВШЭ



Салех Х. М. [📧] руководитель проекта СП4 «MLOps-платформа исполнения и мониторинга ИИ-моделей», доцент ДПИ, к.т.н.



Хританков А. С. [📧] архитектор MLOps-платформы, доцент ДПИ, доцент МФТИ, к.ф.-м.н.

Круглый стол по прикладной математике и информатике на ФКН



Модератор:

Соколов Е. А. [📧], руководитель ДБДИП, научный руководитель ЦНО ФКН

В рамках круглого стола мы поговорим о той части работы факультета, которую можно отнести к прикладной математике и информатике. Это наука и образование в сфере машинного обучения и искусственного интеллекта, аналитики, фундаментальной математики, инфраструктурной разработки, алгоритмов. Участники круглого стола поделятся своим видением о состоянии и сложностях по указанным направлениям. Вместе мы обсудим, какими могут быть следующие шаги, каким представляется развитие направления в ближайшие годы.

Участники круглого стола:



Аржанцев И. В. [📧] декан ФКН, заведующий НУЛ АГП, профессор ДБДИП, д.ф.-м.н.



Грибов Ф. Ю. [📧] преподаватель ДБДИП, член тренерского штаба по подготовке сборной России по информатике



Деркач Д. А. [✉] директор по прикладным исследованиям и разработкам ИИИиЦН, заведующий НУЛ МАБД, доцент ДБДИП, Ph.D.



Пучкин Н. А. [✉] заведующий Л ТОМИИ, доцент БК ИППИ РАН, с.н.с. МЛ СААМД, к.комп.н.



Толстикова А. А. [✉] академический руководитель Школы анализа данных компании Яндекс

Дискуссия «Исследовательские работы: что мы хотим от студентов, что студенты хотят от нас?»



Модератор дискуссии:

Ратников Ф. Д. [✉] доцент ДБДИП, в.н.с. НУЛ МАБД, к.ф.-м.н.

Навык проведения научной работы и грамотной презентации результатов — это один из ключевых пререквизитов научной работы в целом. Соответственно, это одна из важнейших компонент универсального университетского образования, наряду с широким набором базовых знаний и углубленного изучения специальных направлений.

Курсовые и дипломные исследовательские работы под руководством научных сотрудников — уникальная возможность получить практические навыки исследовательской работы. Однако, чтобы научная работа была успешной, желательно, чтобы научные руководители обучили руководимых студентов определенным методологическим принципам, что и в какой последовательности делать.

Мне бы хотелось обсудить с коллегами принципы научных исследований, сформировать некоторый чеклист необходимых компонентов успешного исследования и последующей успешной защиты результатов этого исследования нашими студентами.


Участники: Все желающие обсудить тему участники конференции

Доклады основной программы

В данно разделе приводятся аннотации всех докладов научной конференции ФКН 2024 года. В случае заинтересованности темой можно связаться с авторами доклада, для чего приводится ссылка на их личную страницу, где имеется дополнительная информация, а также их контакты для связи.

Сложность данных и простота классификаторов



Кузнецов С. О.  руководитель ДАДИИ, профессор, д.ф.-м.н.

В докладе будут представлены текущие работы по сложности данных с точки зрения замкнутых множеств признаков и импликационных зависимостей в данных, которые они порождают.

Возможности динамического полиморфизма при процедурно-параметрическом программировании



Легалов А. И. [✉] профессор ДПИ, д.т.н.

Косов П. В. [✉] аспирант ФКН

Разработка программ зачастую связана с инкрементальным расширением функциональности. Повышение надежности и качества в этом случае могут быть достигнуты минимизацией изменений в уже написанном коде. Для инструментальной поддержки эволюционной разработки предложена процедурно-параметрическая парадигма программирования, расширяющая возможности процедурного подхода. Она обеспечивает безболезненное расширение как данных, так функций, используя при этом статическую типизацию вместо разыменования типов. Проводится сравнение процедурно-параметрической парадигмы с объектно-ориентированным подходом по поддержке эволюционной разработке. Рассматриваются альтернативные реализации паттернов проектирования и других ситуаций, требующих гибкой разработки программного обеспечения. Демонстрируется расширение языка программирования С механизмом, который обеспечивает поддержку процедурно-параметрического полиморфизма.

Использование дискурсивного анализа для выявления манипулятивных приемов и умышленных искажений в текстовых данных



Ильвовский Д. А. [✉] доцент ДАДИИ, к.т.н.

Распространенность манипуляции информацией в Интернете создала потребность в системах обнаружения манипулятивных текстовых фрагментов. Такие системы, как правило, фокусируются на отдельных словах, игнорируя лингвистическую структуру. Мы стремимся преодолеть это упущение. В нашей работе мы представляем первую попытку использования анализа дискурса для решения такого рода задачи. Мы рассматриваем как классификацию на уровне абзацев, так и на уровне токенов и используем архитектуру Transformer, учитывающую дискурс. Наши эксперименты для текстов на английском и русском языках демонстрируют значительный прирост производительности по сравнению с рядом базовых моделей. Более того, наше исследование подчеркивает важность определенных типов дискурсивных отношений, а наш углубленный анализ выявляет сильную корреляцию между техниками манипуляции и дискурсивными фрагментами.

Выявление коннектома у глухих посредством решения обратной задачи



Харламов А. А. [✉] профессор ДПИ, д.т.н.

Верхлютов В. М. [✉] старший научный сотрудник Института высшей нервной деятельности и нейрофизиологии РАН, к.м.н.

Бородин Н. С. студент НИУ ВШЭ

Выявление коннектома является наиболее актуальной задачей в электроэнцефалографии и магнитоэнцефалографии в настоящее время. Одним из путей решения этой задачи является использование свертки несущего (в данном случае – речевого) сигнала с записями ЭЭГ (МЭГ). В этом случае матрица J диполей тока как источников на промежутке времени записи ЭЭГ, связана с матрицей S признаков – огибающей речевой волны на том же промежутке времени – через так называемую матрицу функций нейротокowego отклика Φ . Матрица измеренных ЭЭГ-сигналов Y связана с матрицей диполей тока J через так называемую матрицу поля L . Если бы нам была известна фоновая активность мозга, то вычисление диполей тока могло бы быть осуществлено решением задачи максимального правдоподобия минимизацией функции $\|Y - L\Phi S\|^2$. Чтобы смягчить некорректный характер этой постановки (отсутствие знаний о фоновой активности), гладкий временной ряд аппроксимируется ядрами Габора из словаря G : $\Phi = \Theta G^T$ и $\tilde{S} = G^T S$, и теперь надо решать задачу максимального правдоподобия для функции $\|Y - L\Theta\|^2$ с использованием алгоритма Champ-Lasso.

Представлены коннектомы нормально слышащих людей и глухих. Коннектом является важной информацией при диагностике глухоты, что позволяет, по крайней мере, разделить глухих на две категории: патология сигнальной обработки, или патология представления информации в коре.

Diffusion models for controllable image generation



Аланов А. [icon] заведующий ЦГОБМ,
исследователь AIRI, к.комп.н.

In this presentation, we explore the advancements in controllable generation using diffusion models, a rapidly evolving area in AI research. We will delve into the underlying mechanisms that enable diffusion models to produce high-quality, diverse, and controllable outputs across various applications.

A significant highlight of this presentation is our recent work, «Guide-and-Rescale», which has been accepted to ECCV 2024. This paper introduces a novel approach for editing images using diffusion models, showcasing how precise control can be achieved in image manipulation tasks.

We will also discuss our ongoing research in personalized generation, where diffusion models are tailored to produce content that aligns with individual preferences and styles. Additionally, we'll cover our innovative work on generating architectural floorplans, demonstrating how diffusion models can be leveraged for creative and practical design applications.

Подводные камни в бенчмарках табличного ML и новый бенчмарк TabReD



Рубачёв И. В. [📧] м.н.с. НУЛ компании Яндекс,
преподаватель БК Яндекс

We introduce TabReD, a collection of industry-grade tabular datasets, filling the gaps in academic benchmarks. Our evaluation reveals performance differences for various models and techniques in a new setting.

Текст работы: <https://arxiv.org/abs/2406.19380v2>

Symbolic regression for properties prediction and defect structure generation in MoS₂ and WSe₂



Лазарев М. В. [✉] доцент ДБДИП, научный сотрудник НУЛ МАБД (Lambda), Ph.D.

The recent advancements in neural network approaches and large databases of materials have catalyzed the application of deep learning methods to atomistic predictions. Machine Learning (ML) techniques, grounded in databases generated through Density Functional Theory (DFT), have found wide applications. These machine learning methods facilitate the design of new materials by predicting properties with accuracy nearly on par with ab-initio calculations, yet at significantly lower computational expenses. Over the past few years, a myriad of swift and precise deep learning architectures have emerged. Among these, graph neural networks, such as MEGNet, CGCNN, SchNet, and GemNet, have proven to be particularly effective. Nonetheless, this approach is not without its limitations, which include the necessity for substantial data, costly model training, lack of result interpretability, and generalization challenges. Meanwhile, symbolic expressions elucidate a clear relationship between observations and the target variable.

Инструмент на основе методов NLP для анализа журналов событий среды .NET CLR



Ставцев М. А. [📧] стажёр-исследователь НУЛ ПОИС

Шершаков С. А. [📧] заместитель руководителя, доцент ДПИ, к.комп.н.

В докладе мы представляем инструмент для анализа журналов событий среды .NET CLR, разработанный на основе метода, вдохновленного подходами обработки естественного языка (NLP). Мы рассмотрим актуальную проблему мониторинга и оптимизации программных систем через детальный анализ журналов событий. В основе инструмента лежит архитектура BERT с усовершенствованной токенизацией, специально адаптированной для работы с журналами событий. Инструмент реализован на Python с использованием библиотек и базы данных SQLite, что делает его подходящим как для академических исследований, так и для решения практических задач в индустрии. В ходе доклада мы продемонстрируем результаты работы предложенного подхода в сжатии последовательностей событий, выявлении повторяющихся паттернов и обнаружении аномалий.

Проект программно аппаратного комплекса для ухода за растениями с элементами ИИ



Александров Д. В. [📧] профессор ДПИ, заведующий
НУЛ ОиМТ, д.т.н.

Резуник Л. [📧] стажёр-исследователь НУЛ ОиМТ

Прозорский М. А. [📧] стажёр-исследователь НУЛ
ОиМТ

Доклад посвящен разработке программно-аппаратного комплекса для автоматизированного ухода за растениями, интегрирующего элементы искусственного интеллекта. В нем представлена архитектура системы, включающая в себя датчики для сбора данных о состоянии растений (температура, влажность) и исполнительные механизмы для управления поливом. Искусственный интеллект используется для идентификации растений и их заболеваний. Особое внимание уделяется разработке iOS-приложения, которое позволяет пользователю отслеживать состояние растений, получать уведомления о необходимости полива, менять настройки системы, просматривать историю полива и получать информацию о возможном заболевании растения. В докладе будут рассмотрены следующие аспекты: выбор аппаратных компонентов, разработка программного обеспечения и обучение алгоритмов машинного обучения.

Анализ автоматического распознавания, генерации и интерпретации юмора



Баранов А. М. [📧] стажёр-исследователь НУЛ ММВП, приглашённый преподаватель ДБДИП

Доклад посвящен анализу автоматического распознавания, генерации и интерпретации юмора. В нем будет сделан акцент на двух работах: «Dataset of wordplay in news headlines» и «You Told Me That

Joke Twice: A Systematic Investigation of Transferability and Robustness of Humor Detection Models». Кроме того, в докладе будут затронуты и другие актуальные работы по вычислительному юмору, что позволит комплексно оценить текущее состояние области.

В первой работе на тему детекции юмора «A Systematic Investigation of Transferability and Robustness of Humor Detection Models» анализируется переносимость и устойчивость моделей распознавания юмора.

Эксперименты показали, что, хотя модели показывают хорошие результаты на отдельных датасетах, они склонны к переобучению и демонстрируют нестабильное поведение на данных из других доменов. Также были рассмотрены большие языковые модели, такие как GPT-3.5 и Flan-UL2, которые показали конкурентоспособные результаты, но сталкиваются с проблемами в переносе навыков юмора.

Во второй работе рассматривается игра слов в новостных заголовках. Для ее анализа был создан набор данных, который включает 2700 заголовков из газеты «Коммерсантъ». Из них около половины аннотированы как содержащие игру слов. Также датасет выделяется из других похожих наборов данных тем, что примеры игры слов снабжены контекстом в виде описания новостного события.

Обе работы объединяет задача улучшения моделей, способных эффективно распознавать и интерпретировать юмор. Датасет игры слов предоставляет возможности для изучения интерпретации юмора через ссылки на исходные выражения, в то время как «A Systematic Investigation of Transferability and Robustness of Humor Detection Models» фокусируется на переносимости и устойчивости моделей.

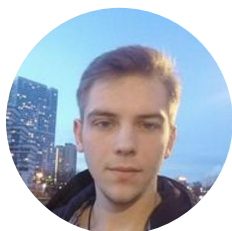
Синтез иерархических моделей процессов: подход на основе разбиения событий на множества



Бегичева А. К. [📧] стажёр-исследователь НУЛ ПОИС, преподаватель ДПИ

Process mining — это область компьютерных наук, которая занимается синтезом и анализом моделей процессов на основе автоматически генерируемых журналов событий. Синтезированная стандартными алгоритмами модель процесса может быть слишком подробной, сложной и трудной для понимания экспертами. В докладе рассматривается задача синтеза иерархической модели бизнес-процесса из низкоуровневого журнала событий, то есть, задача автоматического синтеза более удобочитаемых и понятных моделей процессов на основе данных, хранящихся в журналах событий информационных систем. Также предлагается алгоритм синтеза иерархических моделей процессов в виде двухуровневых сетей потоков работ. Алгоритм основан на предопределенном разбиении событий на множества, которые определяют подпроцессы, соответствующие переходам на верхнем уровне двухуровневой сети потоков работ. В отличие от существующих решений, представленный алгоритм не накладывает ограничений на поток управления процессом, а также допускает параллелизм и итерации.

Решение задачи непарного переноса стиля на основе дистилляции диффузионных моделей



Ракитин Д. Р. [📧] м.н.с. ЦГОБМ,
преподаватель ДБДИП

Цель методов дистилляции диффузионных моделей — их сжатие в одношаговые генераторы с сохранением настолько близкого к оригиналу качества, насколько возможно. Среди этих методов

Distribution Matching Distillation (DMD) представляет собой фреймворк для обучения генераторов произвольного вида, применимых за пределами безусловной генерации. В докладе будет рассказано о нашей с коллегами недавней работе, в которой предлагается модификация метода DMD, применяемая к задачам непарного переноса стиля. Мы показываем его применимость к различным задачам перевода между доменами, включая изображения, где наш метод демонстрирует сравнимое и/или лучшее качество, чем многошаговые бейзлайны.

Декодирование BOLD сигнала по данным ЭЭГ



Семенов И. В. [✉] аспирант ДАДИИ, аналитик ИКН

Функциональная магнитно-резонансная томография (фМРТ) является важным неинвазивным методом нейровизуализации, который позволяет исследовать активность глубоких структур мозга. Однако применение фМРТ связано с

необходимостью лабораторных условий, дорогостоящего оборудования и квалифицированного персонала. В отличие от этого, электроэнцефалография (ЭЭГ) представляет собой более доступный и практичный метод, способный проводить измерения даже в домашних условиях или в рамках умной одежды, хотя напрямую не фиксирует активность глубоких структур мозга, что ограничивает её применение в некоторых областях. Тем не менее, в ряде исследований была обнаружена связь между сигналом, измеряемым фМРТ (BOLD), и многоканальным ЭЭГ сигналом, что открывает возможности для моделирования этой зависимости с помощью методов машинного обучения. В данной работе мы представляем новую методику оценки BOLD сигнала глубоких структур мозга на основе ЭЭГ, использующую нейронные сети, и продемонстрировали её эффективность по сравнению с современными подходами. Наш метод позволяет предсказывать средний BOLD сигнал и одновременно оценивать дифференциальную активность в нескольких интересующих областях (ROI) мозга. Эта методика представляет собой важный шаг к созданию технологии цифрового двойника фМРТ на базе ЭЭГ. Разрабатываемая технология имеет потенциал для открытия новых перспектив в исследовании функций подкорковых структур головного мозга человека и создании доступных диагностических инструментов для выявления заболеваний нервной системы, связанных с нарушениями в работе глубоких отделов мозга.

Genetic engineering algorithm (GEA) and genetic speciation in solving combinatorial optimization problems



Сохраби М. [✉] стажёр-исследователь НУЛ ММВП, преподаватель ДАДИИ

In combinatorial optimization, solving complex problems efficiently is often a challenge due to the vast and discrete solution spaces. Traditional algorithms like Genetic Algorithms (GAs) have been widely used due to their heuristic search capabilities, simulating the process of natural selection through crossover and mutation. However, GAs often struggle with issues like premature convergence and local optima, limiting their performance in finding high-quality solutions. Two advanced modifications of GAs aimed at overcoming these limitations. First, the Genetic Engineering Algorithm (GEA), which draws inspiration from genetic engineering principles, offers a novel search mechanism. By mimicking processes like gene insertion, purification, and material exchange from elite solutions, GEA enhances population diversity and accelerates convergence towards optimal solutions, effectively addressing challenges in large-scale combinatorial problems. The second focus is on the concept of Genetic Speciation, which mirrors the process of species divergence in nature. This approach introduces mechanisms for creating sub-populations or "species" within the overall population, allowing the algorithm to maintain diversity and explore multiple promising regions of the solution space simultaneously. This helps avoid premature convergence and allows the algorithm to better adapt to the problem's landscape. Through this interplay of genetic engineering and speciation concepts, these algorithms provide a more robust framework for solving a wide range of combinatorial optimization problems. By exploring these innovative approaches, this research will highlight their potential for improving solution quality and efficiency in complex optimization tasks.

Майнинг ациклических DFG моделей для ациклических процессов



Шаимов Н. Д. [✉] стажёр-исследователь НУЛ ПОИС

Ломазова И. А. [✉] профессор ДПИ,
заведующая НУЛ ПОИС, д.ф.-м.н.

Мицюк А. А. [✉] доцент ДПИ, с.н.с. НУЛ ПОИС,
к.комп.н.

В докладе рассматривается проблема майнинга моделей ациклических процессов. Особенностью ациклических процессов является отсутствие повторов действий экземплярами процесса. Для представления моделей процессов используются различные нотации, но наиболее популярным способом представления является Directly-Follows Graph (DFG). Модели в виде DFG представляют собой ориентированный граф отражающий действия процесса и переходы между ними.

Существующие методы майнинга DFG моделей могут порождать циклы в DFG моделях для ациклических процессов. Циклы в контексте ациклического процесса значительно снижают точность модели, а также усложняют визуальный анализ. Для устранения циклов предлагается новый алгоритм майнинга DFG. Алгоритм позволяет получать ациклические DFG модели ациклических процессов и основан на слиянии нескольких ациклических DFG с допущением повторов вершин в модели. В рамках доклада будет рассмотрен пример работы алгоритма на примерах и дана оценка эффективности алгоритма на реальных данных.

Анализ применения Web-фреймворков: React, Angular, VUE.JS и Svelte в проектах с различным уровнем вложенности компонентов



Каравеева Е. А. [✉] стажёр-исследователь НУЛ ОиМТ, пригл. преп. ДПИ

Думкин Н. А. [✉] стажёр-исследователь НУЛ ОиМТ, пригл. преп. ДПИ

Исследование направлено на выявление преимуществ и недостатков фреймворков: Vue.js, Angular, Svelte и React в зависимости от уровня вложенности компонент, что позволит разработчикам принимать обоснованные решения при выборе технологий для своих проектов.

MiSTi: метод оценки времени разделения и скорости миграции при переменном размере популяции



Щур В. Л. [📧] профессор ДБДИП, заведующий МЛ СВГ, д.ф.-м.н.

Оценка изменений эффективного размера популяции (N_e) во времени – фундаментальная задача популяционной генетики, но интерпретация N_e как эффективного числа особей, способных к размножению, осложняется популяционной структурой. Фактически, изменения N_e , обсуждаемые во многих работах, могут быть следствием миграции между популяциями, а не изменением их размеров. Мы подходим к этой давней проблеме, рассматривая модель, учитывающую изменения размера популяции, миграцию и время расхождения этих популяций. Таким образом мы можем корректировать оценки изменений N_e во времени и оценивать реальный размер популяции, связанной с другой популяцией через миграцию. Мы также разработали метод для оценки времени расхождения и скорости миграции, учитывающий сложные сценарии изменения размера популяции. Мы применили наш метод к ранее опубликованным человеческим полногеномным данным, и показали, что с учетом миграции и изменений N_e , оценка времени расхождения между африканскими популяциями бушменов и динка равна 108 тысячам лет вместо ранее опубликованной даты в 255 тысяч лет. При помощи компьютерного моделирования мы показали, что ранее опубликованное и удивительно глубокое время расхождения между бушменами и динка на самом деле вызвана смещением, обусловленным изменениями N_e во времени.

ИИ и генетический компьютер



Попцова М. С. [✉] доцент ДБДИП,
заведующая МЛ Биоинформатики, к.ф.-м.н.

Клетка как единица всего живого действует согласно генетическим программам, закодированным в геноме. Задача современной науки понять алгоритмы и принципы кодирования генетических программ с целью создания цифрового двойника клетки. Информация в геноме закодирована на разных уровнях, и современные эксперименты позволяют получать сигналы с разных уровней по всей длине генома. Объем омиксных данных измеряется в сотнях терабайтов, и задачи анализа и поиска новых взаимосвязей между функциональными элементами возложена на искусственный интеллект. В докладе будут представлены разработки систем ИИ на основе архитектур глубинного обучения – от сверточных и рекуррентных до больших языковых моделей, с помощью которых осуществляется обратный инжиниринг геномных программ. Будут представлены работы, в которых с помощью систем ИИ были открыты новые механизмы регуляции геномных процессов посредством флипонов – вторичных структур ДНК.

Нейросетевое определение популяционной принадлежности человека по графу общих сегментов генома



Глызин Д. С. [✉] н.с. МЛ СВГ, к.ф.-м.н.

Рассматриваются полученные из массива анализов ДНК графы общих по происхождению сегментов генома между участниками, указавшими свое происхождение из некоторого географического региона или свою принадлежность к

определенному этносу.

Решается задача классификации для популяционной принадлежности нового участника.

Показано, что в случае графа с неполной разметкой вершин графовые сверточные нейросети получают преимущество над классическими методами классификации.

Инновационные медицинские приложения: Проект EYAS для анализа фундус-изображений



Рябцев Д. И. [\[✉\]](#) доцент ДПИ, к.м.н.

Проектная группа «Информационные системы для медицинских приложений» (Information Systems for Medical Applications, ИСМед) занимается соединением передовых технологий программного обеспечения и медицинской практики,

разрабатывая инновационные медицинские информационные системы, которые улучшают клинические рабочие процессы. В основе нашей текущей работы - проект EYAS, инновационное приложение для анализа фундус-изображений. В отличие от традиционных скрининговых инструментов, EYAS использует комплексный подход, анализируя как нормальные, так и патологические особенности структур глазного дна, предоставляя врачам подробные, объяснимые результаты, а не предиктивные диагнозы.

Благодаря модульной системе моделей глубокого обучения, интегрированных с традиционными методами компьютерного зрения, EYAS улучшает процесс принятия клинических решений, генерируя описания изображений глазного дна на естественном языке. Эта новая система не только повышает точность диагностики, но и легко интегрируется в существующую медицинскую инфраструктуру благодаря безопасным и удобным интерфейсам. Способствуя сотрудничеству между информатикой и здравоохранением, EYAS служит примером нашего стремления развивать медицинскую диагностику с помощью технологий, создавая инструменты, которым клиницисты могут доверять в своей работе.

Исследование механизмов лекарственной устойчивости возбудителя туберкулеза к лекарственным препаратам с помощью моделей машинного обучения



Неверов А. Д. [✉] в.н.с. МЛ СВГ, приглашённый преподаватель ДБДИП, к.б.н.

Возбудитель туберкулеза, *Mycobacterium tuberculosis* (МТВ), являлся главной причиной смертности от инфекционных заболеваний в период до пандемии COVID-19 и вернул первенство сразу после ее окончания. Благодаря эффективной терапии заболеваемость туберкулезом в РФ и в мире в целом снижается, при этом угрожающе растет доля устойчивых форм туберкулеза, в связи с чем, раннее выявление ЛУ и индивидуальный подбор комбинаций эффективных препаратов приобретает высокую значимость. В настоящее время в мире, особенно в развитых странах, в рутинную практику внедряются генетические методы выявления ЛУ, основанные на определении наличия специфических мутаций в определенных генах. Генетические методы для отдельных препаратов с хорошо изученными механизмами действия имеют точность, сопоставимую с золотым стандартом - фенотипическими методами, однако, для других препаратов их чувствительность существенно ниже.

В попытках найти новые гены, участвующие в развитии лекарственной устойчивости, используются статистические методы машинного обучения (МО). Эти методы используют большие коллекции изолятов с известными полногеномными последовательностями и фенотипами к различным препаратам. Однако высокая корреляция между наличием или отсутствием устойчивости к препаратам, которые используются вместе в одной схеме лечения, затрудняет обнаружение мутаций, являющихся причиной устойчивости, с помощью методов МО.

Мы применили разнообразные модели МО для поиска ассоциаций мутаций в геноме МТВ с лекарственными фенотипами (устойчивостью и чувствительностью) для 13 противотуберкулезных препаратов:

линейные модели (логистическую регрессию с различными штрафными функциями (SCAD, MCP, L1)), полиномиальный алгоритм поиска наилучшего подмножества признаков минимального размера (ABESS) и LSTM со вниманием. Для обучения и тестирования моделей мы использовали множество 11 822 генотипов, охарактеризованных на наличие/отсутствие устойчивости как минимум к одному из препаратов. Мы сравнивали методы по способности отбирать мутации, вызывающие устойчивость к каждому конкретному препарату, и не отбирать мутации в генах, которые, как известно из литературы, связаны с устойчивостью к другим препаратам. ABESS значительно превзошел остальные методы, отобрав наборы мутаций, с минимальным числом неправильных ассоциаций. Однако по точности предсказания ABESS уступает методам логистической регрессии с регуляризацией. С помощью ABESS мы обнаружили четыре новых гена, мутации в которых ассоциированы с наличием ЛУ (perC, Rv2017, varB1, fadA).

Гибридизация методов машинного обучения и классических моделей



Деркач Д. А. [✉] директор по прикладным исследованиям и разработкам ИИИиЦН, заведующий НУЛ МАБД (Lambda), доцент ДБДИП, Ph.D.

Искусственный интеллект (ИИ) больше не просто модное слово; это неотъемлемая часть современного бизнеса и общества. От автоматизации рутинных задач до создания персонализированного клиентского опыта ИИ формирует наш мир с головокружительной скоростью. Одной из важных задач является интеграция знаний, полученных человечеством, до прихода ИИ и гибкости новых методов анализа данных. В своём докладе я намечу направления исследований и расскажу про результаты применения таких подходов в задачах реального мира.

Применение диффузионных моделей для генерации текстур трёхмерных моделей



Струминский К. А. [✉] стажёр-исследователь НУЛ компании Яндекс, к.комп.н.

Создание текстур для окраски трёхмерных моделей трудоемкий и нетривиальный процесс. В то же время диффузионные модели оказываются способны достоверно передавать знания о внешнем виде окружающих нас предметов при генерации изображений. В докладе пойдет речь о ряде подходов, которые позволяют применить вышеупомянутые диффузионные модели для генерации физически достоверных текстур в трёхмерном моделировании для упрощения труда художника.

Нормальная аппроксимация и мультипликативный бутстреп для алгоритмов стохастической аппроксимации



Самсонов С. В. [✉] н.с. МЛ СААМД (HDI Lab),
старший преподаватель БК ИППИ РАН

В данной работе нами получены оценки типа Берри-Эссена точности нормальной аппроксимации для итераций линейной стохастической аппроксимации (LSA) с усреднением Поляка-Рупперта и убывающим шагом. Наши результаты показывают, что в случае независимого шума наилучшая скорость нормальной аппроксимации достигается при выборе наиболее агрессивного шага $\alpha_k \sim k^{-1/2}$. Мы также покажем, как этот результат может быть использован для построения неасимптотических доверительных интервалов для оценки параметров в LSA с использованием мультипликативного бутстрепа. В докладе также будут рассмотрены обобщения на случай марковского шума.

Доклад основан на результатах работы <https://arxiv.org/abs/2405.16644>

Автоматизированное обучение сверточных нейронных сетей



Болдырев А. С. [✉] доцент ДБДИП,
н.с. НУЛ МАБД (Lambda), к.ф.-м.н.

Стремительное развитие приложений машинного обучения влечет необходимость обучения большого количества моделей. Одним из способов организации обучения множества моделей является

подход автоматического машинного обучения (AutoML), в котором отсутствует человеческий контроль обучения отдельной модели. Вследствие этого важным требованием AutoML является стабильность обучения моделей внутри подхода. В докладе представлен алгоритм совместного поиска архитектуры искусственной нейронной сети и её гиперпараметров, позволяющий получить стабильно обучающиеся модели. Представлены результаты применения такого алгоритма для сверточных нейронных сетей. Обсуждается влияние размера обучающей выборки и способа инициализации весов модели на стабильность её обучения.

Эффективная структурированная параметризация ортогональных матриц



Рахуба М. В. [📧] доцент ДБДИП,
заведующий НУЛ МТММО, к.ф.-м.н.

Мы предлагаем новый малопараметрический класс ортогональных матриц (GS матрицы). Этот класс эффективно параметризуется, что позволяет накладывать ограничения ортогональности на обучаемые в задачах глубинного обучения матрицы. В частности, мы применяем его для тонкой настройки LLM, в диффузионных моделях, а также в сверточных архитектурах для повышения их устойчивости.

О преимуществах использования сходства данных в задачах обучения для сокращения числа коммуникаций в распределённых алгоритмах



Двинских Д. М. [✉] доцент БК ИППИ РАН, с.н.с. МЛ СААМД (HDI Lab), Ph.D.

Использование распределенной системы вместо одного вычислительного устройства позволяет обучать современные модели машинного обучения на огромных наборах данных. Однако эта мощь распределенных систем сопряжена с различными трудностями: коммуникации сервера (основного устройства) с вычислительными машинами может существенно замедлить процесс обучения, особенно для систем, обладающих большими вычислительными ресурсами. Существуют различные методы для уменьшения количества таких коммуникаций, среди них методы, учитывающие структуру задачи, например сходство данных, довольно часто встречающееся в приложениях машинного обучения. В докладе будет показано, как использовать сходство данных в задачах обучения с учителем, для того чтобы более эффективно коммуницировать с сервером.

Обнаружение разладки во временных рядах с помощью отслеживания лучших экспертных стратегий



Пучкин Н. А. [✉] доцент БК ИППИ РАН,
с.н.с. МЛ СААМД (HDI Lab), к.комп.н.

Предложен новый алгоритм обнаружения разладки в режиме реального времени. В основе алгоритма лежит метод параметрического моделирования и последовательная оценка градиента логарифма плотности. Обобщенная версия предсказателя Fixed Share автоматически отслеживает наилучшее значение параметра на участках до и после разладки, не требуя при этом значительных вычислительных ресурсов. В результате предложенный алгоритм показывает многообещающие результаты на синтетических и реальных данных.

Формирование ленты новостей на основании технических событий с помощью технологий больших языковых моделей



Пашигорев К. И. [✉] приглашённый преподаватель
БК ПАО Сбербанк, руководитель направления Сбер

Резников А. О., главный разработчик Сбер

Одним из основных вызовов ближайшего времени является внедрение технологий искусственного интеллекта во все производственные процессы, начиная от самых рутинных и до высокотехнологичных операций. Искусственный интеллект становится главным помощником для всех сфер деятельности, в том числе и для экспертов, работающих в сфере ИТ. В докладе рассматривается приложение для формирования новостей на основании техноданных: новости формируются с помощью большой языковой модели, и далее выполняется построение рекомендаций с помощью технологий рекомендательных систем. Целевой аудиторией являются эксперты категории data analysts, data scientists, data engineers, chief data officers. В докладе приводится пример архитектуры такой системы, а также варианты промптов для формирования новости.

Архитектурный подход с одной точкой входа и выхода: использование Istio для управления микросервисами

Кензин И. И. [✉] приглашённый преподаватель БК ПАО Сбербанк, ведущий инженер Сбер

С ростом числа микросервисов в современных распределённых системах возрастают требования к управлению трафиком, безопасности и мониторингу. Одним из решений для этих задач является сервисная сетка (Service Mesh), обеспечивающая централизованное управление взаимодействиями между сервисами. В докладе я расскажу о концепции архитектуры с одной точкой входа и выхода (Ingress и Egress), её роли в обеспечении безопасности, управляемости и гибкости в микросервисных системах. Особое внимание будет уделено практическому применению этой концепции с использованием Istio — популярной платформы сервисной сетки. Я продемонстрирую, как Istio помогает централизовать управление трафиком, улучшить видимость и повысить безопасность взаимодействий между сервисами, обеспечивая упрощённое управление сложными архитектурными системами.

Типизированные сети Петри с переменными весами дуг



Ломазова И. А. [✉] профессор ДПИ,
заведующая НУЛ ПОИС, д.ф.-м.н.

Мицюк А. А. [✉] доцент ДПИ, с.н.с. НУЛ ПОИС,
К.комп.н.

Ривкин А. [✉] доцент Датского технического
университета (DTU) в Люнгбю-Торбек, Ph.D.

В последние годы все больший интерес исследователей привлекают процессы, основанные на взаимодействии объектов. Одним из наиболее популярных формализмов для описания жизненных циклов объектов и фиксации отношений между ними являются объектно-центрические сети Петри (Object-Centric Petri nets – OC-nets). Важная особенность таких сетей – наличие переменных дуг (variable arcs), позволяющих при срабатывании перехода задействовать произвольное количество объектов одного типа. В этой работе мы обобщаем эту конструкцию, вводя довольно простой и универсальный механизм параметризации веса дуг с помощью линейных комбинаций весовых переменных и определяем новый формализм типизированных сетей Петри с переменными весами дуг. Мы доказываем, что переменные дуги, удобные для моделирования, могут быть эффективно элиминированы, т.е. что новая модель эквивалентна по выразительности классическим P/T-сетям. Это позволяет применять для OC-сетей методы анализа, разработанные для классических сетей Петри.

Доклад основан на статье: Lomazova I. A., Mitsyuk A. A., Rivkin A. Typed Petri Nets with Variable Arc Weights, in : Business Process Management Workshops. BPM 2023 International Workshops, Utrecht, The Netherlands, September 11–15, 2023, Revised Selected Papers. / Ed. by J. De Weerd, L. Pufahl. Vol. 492 Switzerland: Springer, 2024. p. 483–495.

Двусторонний вывод с режимом применения для зависимых типов с неявными аргументами



Соколов П. П. [✉] преподаватель ДБДИП

Системы типов с зависимыми типами, изначально применявшиеся в инструментах интерактивного доказательства теорем в качестве универсального логического основания, постепенно начинают вводиться в существующие языки

программирования общего назначения или даже становятся основой для новых языков программирования с целью более тонкого контроля за поведением программы, уменьшения вероятности ошибки программиста и повышения его удобства.

Прежде всего, использование той или иной системы типов в языке выражается в компиляторе в виде алгоритма, разрешающего задачу типизации: имеет ли данное выражение t данный тип T ? И, вообще говоря, в системах с зависимыми типами решение этой задачи нетривиально и в классическом варианте требует от программиста большого числа аннотаций типами, что мешает эргономике и загрязняет код. В данном докладе мы рассмотрим недавно появившуюся технику двустороннего вывода типов с режимом применения и применим её в построении системы с зависимыми типами с неявными аргументами, критическим образом сокращающую необходимое от программиста число аннотаций типов.

Дизъюнктивная сложность



Вялый М. Н. [📧] профессор ДБДИП, в.н.с. МЛТИ,
к.ф.-м.н.

При изучении одной из задач в теории формальных языков возникла новая модель вычислений: дизъюнктивное представление булевых функций (или, что то же самое, семейств подмножеств). Свойства

этой модели радикально отличаются от свойств стандартных неоднородных моделей вычисления (скажем, схемной сложности). В докладе будет рассказано о первых результатах исследований в этом направлении и возможных направлениях дальнейших исследований.

Моделирование асинхронного взаимодействия в мультиагентных системах на основе журналов событий



Нестеров Р. А. [✉] доцент ДПИ, м.н.с. НУЛ ПОИС, К.комп.н.

Структура модели процесса, синтезированной напрямую по журналу событий мультиагентной системы часто не дает представления о поведении отдельных агентов, а также о способе их взаимодействия. Для локализации действий, которые выполняются различными агентами и которые вовлечены в их асинхронное взаимодействие мы выделяем и анализируем отношения между событиями в журнале. В результате модель мультиагентной системы представляет собой композицию моделей поведения отдельных агентов, между которыми добавляются каналы асинхронного обмена сообщениями. В работе предложен и обоснован алгоритм выделения и анализа отношений между событиями в журнале событий мультиагентной системы. Результаты экспериментальной оценки разработанного алгоритма подтверждают общее улучшение качественных оценок моделей процессов, синтезированных по журналам событий мультиагентных систем с помощью отношений между событиями в сравнении с монолитными моделями, которые синтезируются напрямую.

Using LLMs for materials design



Аль-Маини А. [✉] младший научный сотрудник
НУЛ МАБД (Lambda)

This study explores the tunability of transition metal dichalcogenides (TMDCs), specifically MoS_2 , through point defect engineering for next-gen optoelectronic and high-tech applications. Using density functional theory (DFT) calculations, we analyze the effects of various defect types and concentrations on material properties. By employing pre-trained large language models, we efficiently predict defect-induced property changes, distinguishing our method from traditional, random high-throughput approaches. Our transformer-based models shows the ability to learn a joint distributions of defects positions, types and physical properties, deepening the understanding of defect-property relationships and providing a robust framework for designing materials with tailored properties, advancing materials science and technology.

HiFi-Stream: решение для потокового усиления речи с помощью генеративных состязательных сетей



Каледин М. Л. [✉] доцент ДБДИП,
н.с. НУЛ МАБД (Lambda), к.комп.н.

Техники усиления речи стали широко используемыми технологиями в мобильных устройствах и программных решений для обработки голоса, упрощающими решение конечных задач. К сожалению, современные решения на основе глубоких нейронных сетей идут вместе с достаточно высокими требованиями к вычислительным ресурсам, что усложняет их применения на низкоресурсных устройствах. Мы представляем модель HiFi-Stream, улучшенную версию модели HiFi++, которая оптимизирована под использование в аудиопотоке. Наши эксперименты показывают, что модель практически не теряет в качестве по сравнению с другими моделями несмотря на её размер и вычислительную сложность: самая лёгкая версия имеет порядка 490k параметров, что в 3.5 раза меньше, чем у HiFi++ и что делает эту модель самой лёгкой и быстрой из решений похожего класса, доступных на текущий момент.

Heterogeneous GNN-based neutron reconstruction in the HGND at the BM@N experiment



Бочарников В. О. [✉] м.н.с. НУЛ МАБД (Lambda), преподаватель ФФ НИУ ВШЭ

The Highly Granular Neutron Detector (HGND) is designed for the BM@N experiment to study neutron emission in heavy ion collisions at beam energies up to 4A GeV. This detector allows the identification of neutrons and the reconstruction of their energies using time-of-flight method, facilitating the assessment of neutron yields and azimuthal flow. The challenging neutron energy range of \$0.5-4\$ GeV and large background contributions require sophisticated reconstruction algorithms. In this contribution, we present a heterogeneous Graph Neural Network-based approach to the neutron reconstruction problem and discuss preliminary results of the proposed algorithm.

Density of sphere packings



Ферик Т. [✉] доцент ДАДИИ, Ph.D.

Packing spheres densely is not just about putting the most oranges on a market stall, but also has applications in the theory of corrector codes and in the search for new materials. I'll tell you about some remarkable results in this field in an accessible (and

illustrated) way.

Introduction to visual analytics for explainable deep learning



Джин С. [icon] доцент ДБИ ВШБ, к.комп.н.

Visual Analytics is a scientific field that involves the visual analysis of large datasets to extract meaningful insights, making it particularly useful for interpreting complex deep learning models. In recent years, there has been a significant increase in the use and development of machine learning-based solutions across various domains, with deep learning playing a major role in this growth. However, deep learning models often lack explainability and function as black-box models. Visual Analytics offers several proposals to address these drawbacks, supporting the emerging field of Explainable Deep Learning (XDL).

This presentation aims to introduce the core concepts of Visual Analytics and its applications in enhancing the explainability of deep learning models. We will explore how Visual Analytics can help identify visualization techniques and analytical tools that are well-suited to support researchers, developers, and users of deep learning solutions. By systematically reporting the contributions of Visual Analytics to Explainable Deep Learning, we aim to identify current gaps and challenges, and serve as a resource for the deep learning community to better understand and utilize these tools. This research highlights the role of Visual Analytics in effectively supporting Explainable Deep Learning and encourages its adoption within the community.

PAC-bayesian learning meets synthetic Ricci curvature bounds



Пари К. [\[📧\]](#) доцент ДАДИИ, Ph.D.

We present a connection between the performance of a range of randomised statistical learning procedures, known as PAC-bayesian methods, and the concept of synthetic Ricci curvature lower bounds for metric-measure spaces. In particular, we show that under limited assumptions on the loss function, and provided the parameter space equipped with the chosen prior has lower bounded Ricci curvature, the expected excess risk of PAC-bayesian procedures converges to 0 with the sample size at rates otherwise obtained in the literature under different (and often more restrictive) assumptions.

Lost in Translation: chemical language models and the misunderstanding of molecule structures



Тутубалина Е. В. [✉] н.с. НУЛ ММВП,
исследователь AIRI, д.комп.н.

This talk will explore the recent integration of chemistry with natural language processing (NLP) and its impact on advancing drug discovery. The critical role of molecule representation in language models (LMs) for enhancing chemical understanding will be discussed. Augmented Molecular Retrieval (AMORE), a versatile zero-shot framework, will be introduced to assess the trustworthiness of various Chemical LMs—those trained solely on molecular data for chemical tasks and those trained on a mixed corpus of natural language texts and string-based structures. The framework utilizes molecule augmentations that maintain essential chemical properties, such as kekulization and cycle replacements. Both encoder-only and generative LMs are evaluated by calculating a metric based on the similarity scores between distributed representations of molecules and their augmentations. Findings from experiments on the ChEBI-20 and QM9 benchmarks reveal that these models consistently yield lower scores compared to graph-based molecular models trained without language modeling objectives. Additionally, it will be discussed how the augmentation of SMILES representations adversely affects performance on chemical property prediction tasks within the MoleculeNet benchmark.

Поймай бота, или широкомасштабная структура естественного языка



Громов В. А. [✉] зам. руководитель,
профессор ДАДИИ, с.н.с. МЛ ИССА, д.ф.-м.н.

В настоящей работе методами топологического анализа данных выделяются и оконтуриваются наиболее устойчивые персистентные гомологии первого порядка в векторных пространствах слов, биграмм и триграмм русского и английского языков (“дырки” в семантических пространствах данных естественных языков). Из гомологий первого порядка были отобраны те, которые действительно отражают крупномасштабную структуру языка, а не неоднородности выборки – в каждом из рассмотренных языков (английский и русский) таких оказалось достаточно мало. Полученные контуры окружают «слепые зоны» языка – области в которых отсутствуют слова/биграммы/триграммы, т.е., области существования концепций, не различимых “в оптике” соответствующего языка.

В рамках решения задачи идентификации ботов, для текстов различной природы: текстов художественной литературы и текстов, сгенерированных большими языковыми моделями, были посчитаны характеристики, основанные на расстоянии входящих в тексты слов/биграмм/триграмм до выделенных персистентных гомологий. На указанных признаках были обучены модели классификации с сильной постановкой – модели обучались на текстах, сгенерированных одним набором ботов, а тестировались на текстах, сгенерированных оставшимися ботами. Несмотря на то, что тексты в тестовых выборках были сгенерированы ботами, не встречавшихся в обучающей выборке, во всех случаях точность классификации получалась не ниже 0.7, а усреднённые по всем комбинациям значения превышали 0.8.

Построение доверенного ИИ: проблемы, стратегии и перспективы



Силаев Ю. В. [✉] старший преподаватель ДПИ, заведующий лаборатории доверенного искусственного интеллекта МИРЭА

В современном мире искусственный интеллект играет все более значимую роль в различных сферах экономики. Однако с ростом его применения возрастает и необходимость обеспечения доверия к этим технологиям. Особую важность наличия доверия у систем ИИ подтверждается в том числе включением в феврале 2024 года понятия «доверенные технологии искусственного интеллекта» в Национальную стратегию развития искусственного интеллекта на период до 2030 года.

В докладе планируется показать результаты исследования актуальных вызовов, стоящих как перед разработчиками систем ИИ, так и перед регуляторами и органами по сертификации. Мы рассмотрим ключевые стратегии, направленные на преодоление этих проблем, включая внедрение этических принципов и стандартов ТК 164 «Искусственный интеллект». В том числе имеются результаты проведения работ с различными регуляторами (на примере ФСТЭК) по направлению осуществления оценки соответствия систем ИИ требованиям безопасности.

В связи с высокой важностью и необходимостью оперативного решения вопроса построения доверенных систем ИИ, регуляторы информационной безопасности рассматривают в том числе возможность внесения весьма жестких требований и ограничений на системы ИИ, в первую очередь для наиболее важных задач по обеспечению безопасности государства. В докладе рассмотрим возможные перспективы применения подобных мер и как они коснутся большинства разработчиков систем ИИ.

О перспективных направлениях исследований методов изучения человеко-машинного взаимодействия в области разработки компьютерных игр



Поздняков Д. А. [✉] преподаватель ДПИ, совладелец и продуктовый директор студий «Винторог» и «Contrast Games», ведущий эксперт проектной группы «ПИКИ»

В докладе обсуждаются существующие и перспективные методы исследования человеко-машинного взаимодействия применительно к разработке компьютерных игр. Авторы выделяют специфику подходов, используемых для изучения взаимодействия игроков с системами цифровых развлечений различных типов с учётом разнообразия игровых механик. Выделенные свойства устройств ввода/вывода, программных компонентов, нарративных техник и поведенческих шаблонов позволяют системно взглянуть на тренды и вызовы в области.

Доклад основан на результатах как на теоретическом анализе области человеко-машинного взаимодействия в компьютерных играх, так и на опыте реальной разработки и поддержки проектов в области цифровых развлечений с оценкой удовлетворенности игроков.

Пучки на графах и рекомендательные системы



Тараканов А. А. [📧] доцент ДБДИП, Ph.D.

Ермилов С. И. [📧] разработчик ВК

Сверточные графовые нейронные сети широко используются для решения различных задач на графе, например, классификация вершин графа и рекомендательные системы. В последнем случае задача сводится к оценке вероятности того, что две вершины соединены ребром. Одним из ключевых шагов в графовых нейронных сетях является передача сообщений от соседних вершин (message passing) с последующей агрегацией векторов признаков каждой из вершин. С одной стороны, данная процедура позволяет учитывать информацию соседей, что позволяет делать более точные рекомендации. С другой стороны, нередко случаи, когда message-passing приводит к потере информативности векторов признаков за счет слишком сильного усреднения по соседям.

Теория пучков на графе предлагает инструменты модификации алгоритма передачи сообщений для решения проблемы излишнего сглаживания (oversmoothing). В данной работе мы рассматриваем один из методов аппроксимации морфизмов пучка и построения рекомендательной системы на основе пучка на графе. Разбираются как и синтетические примеры, так и классические датасеты для рекомендательных систем.

Сегментация и идентификация кузовных деталей легкового автомобиля POLO VW 5



Пантюхин Д. В. [✉] ст. преподаватель ДПИ, МИЭМ, РТУ МИРЭА

В докладе представлена программа сегментации и идентификации кузовных деталей легкового автомобиля POLO VW 5, работающая в составе приложения БЕРТУМ, представленная в виде браузерной и мобильной (Андроид) версии. Программа прошла тестовые испытания и принята в эксплуатацию.

Сегментация и идентификация деталей реализована на основе нейросетевых моделей YOLO v5, дообученных на собранных и размеченных наборах изображений.

Было размечено 130 пар «изображение – маска» неповрежденных автомобилей (4716 сегментов, 74 вида деталей) и 181 изображение поврежденных автомобилей (501 аннотированная деталь, 39 классов деталей) с выделением степени повреждения дефект\потертость.

Показаны результаты идентификации деталей автомобиля со средней точностью mAP 85-97% на 74 классах деталей. Для большинства классов деталей точность составляет от 85% до 98%, но имеются классы деталей (в основном маленького размера), для которых точность составляет ~50%. В среднем эти результаты сопоставимы с мировым уровнем, а для некоторых классов деталей даже превосходят его. Кроме того, в литературе рассматривается существенно меньшее число классов деталей.

Высокая точность идентификации достигнута только для одной марки автомобиля - POLO VW 5. В дальнейших планах расширить этот перечень, для чего будет использован инструмент для автоматизации разметки данных для сегментации, новые изображения поврежденных автомобилей из смет Ингосстраха, и изображения деталей, представленные в автомагазине Автоланс.

Работа выполнена при поддержке Фонда содействия развитию малых форм предприятий в научно-технической сфере в рамках проекта «Разработка и испытания опытного образца программного обеспечения ассистента автовладельца на основе искусственного интеллекта для целей идентификации, поиска запчастей и расчета ремонта легкового автомобиля» (Заявка № С1-244762, Проект № 85560, договор №4996ГС1/85560 от 03.07.2023)

Towards reliable false discovery rate control in classification problems under distribution shift



Боревский А. О. [✉] стажёр-исследователь
НУЛ ИИВБ

The performance of machine learning-based classifiers can significantly degrade in software production due to domain shifts or batch effects that were not considered during training and validation. This can particularly be problematic in automated medical diagnostic systems. Here, we present a simple yet robust heuristic method to mitigate the bias in false discovery rate control in case of data shift. This method is based on the adjustment of the null distribution of test prediction scores, thereby we termed it Test Null Adjustment (TNA) method. The adjusted null distribution then can be used to calculate empirical p-values (EPVs) to be used in the Benjamini-Hochberg (BH) protocol to maintain false discover rate (FDR) control. The main advantage of our TNA method is that it's application does not require any domain specific knowledge, it is fully data-driven, and it operates in the 1-dimensional space of the prediction scores of a classifier ensuring stable operation. We have performed four experimental tests on various biomedical image analysis domain, demonstrating that the introduced TNA method almost entirely can mitigate the bias in FDR induced by data shift.

Improving generative flow networks through reinforcement learning



Морозов Н. В. [✉] стажёр-исследователь ЦГОБМ, преподаватель ДБДИП

Generative Flow Networks (GFlowNets) are models that treat sampling from distributions over compositional discrete spaces as a sequential decision-making problem, training a stochastic policy to construct objects step by step. This talk will cover recent results on the theoretical equivalence between GFlowNet training and specific reinforcement learning (RL) problems, as well as how these results can be utilized in practice to apply existing RL algorithms for improving GFlowNets.

<https://arxiv.org/abs/2310.12934>

<https://arxiv.org/abs/2406.13655>

Особенности применения архитектурных паттернов в мобильной разработке



Сосновский Г. М. [✉] ведущий разработчик компании Звук

Абросов С. А. [✉] стажёр-исследователь НУЛ ОиМТ

Мы рассмотрим несколько популярных архитектурных шаблонов для разработки iOS-приложений, включая Composable Architecture (TCA), MVVM и SVIP. Попытаемся понять зачем они нужны, обсудим ключевые принципы каждого из них, их преимущества и недостатки. В рамках лекции разберём, как эти архитектуры помогают улучшить модульность, тестируемость и поддерживаемость кода, ускоряя процесс разработки.

Плакаты постерной сессии

1	Бхимани К.	стажёр-исследователь НУЛ ИИВБ	Sequence variation identification in high resolution MS/MS data with accurate FDR control
2	Вылегжанин Ф. Е.	стажёр-исследователь МЛ АТиП	Гомотопические группы некоторых торических многообразий
3	Гришина Е. Р.	стажёр-исследователь НУЛ МТМО	Tight and efficient upper bound on spectral norm of convolutional layers
4	Данилов А. А.	м.н.с. НУЛ ОиМТ	Применение Apache Kafka в высоконагруженных приложениях
5	Джоши К.	стажёр-исследователь НУЛ ИИВБ	Large-scale evaluation of the generalization performance of deep neural networks in tandem mass spectrometry data analysis
6	Иванова А. А.	стажёр-исследователь Л ТОМИИ	Guide-and-Rescale: self-guidance mechanism for effective tuning-free real image editing
7	Кривулец Н. В.	стажёр-исследователь НУЛ ИИВБ	Spectrum annotation with transformer models
8	Резуник Л.	стажёр-исследователь НУЛ ОиМТ	Подход к созданию сервиса генерации программного кода мобильных приложений
9	Ромашкина А. Д.	стажёр-исследователь НУЛ ИИВБ	False discovery rate control in classification problems with knock-off classes.
10	Талайкова Н. А.	менеджер НУЛ МАБД	Принципы формирования трехмерных изображений или количественная фазовая визуализация
11	Юдин Н. Е.	стажёр-исследователь НУЛ МТМО	Group and Shuffle: efficient structured orthogonal parametrization (NeurIPS'24)
12	Юсупов В. А.	стажёр-исследователь НУЛ МТМО	Knowledge graph completion with mixed geometry tensor factorization

Sequence variation identification in high resolution MS/MS data with accurate FDR control

Бхимани К. [ORCID] стажёр-исследователь научно-учебной лаборатории искусственного интеллекта для вычислительной биологии

Sequence variation identification in high resolution MS/MS data with accurate FDR control



Kishankumar Bhimani¹, Attila Kertész-Farkas^{*}

¹ * Laboratory on AI for Computational Biology, Department of Data Analysis and Artificial Intelligence, HSE University, Moscow

[†] kbhimani@hse.ru, * akerteszfarkas@hse.ru



Introduction

This research focuses on extracting the highest scoring peptide sequence from a dynamic programming table to maximize the PSM score for a given spectrum, similar to how the Needleman-Wunsch algorithm traces sequence alignments. We track the origin of each cell to identify amino acids and use a greedy approach when multiple paths exist. By tracing from the highest score back to the origin, we can determine the sequence, as shown by the red arrows in Figure 1.

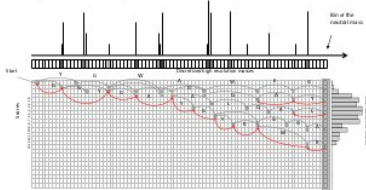


Figure 1: Figure illustrates the score calculation using dynamic programming for a given spectrum. The table starts at $D(c, m) = 1$, where c is the score row and m is the N -terminal residue bin. The exact null distribution of PSM scores is found in the column corresponding to the precursor ion's m/z (shaded gray), with red arrows indicating backtracking.

Dataset & Pipeline

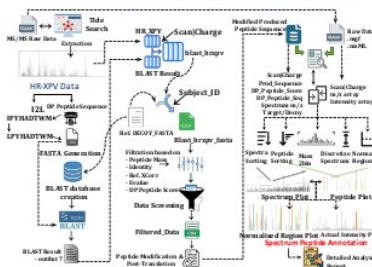


Figure 2: Data Pipeline to find Accurate sequence variation identification in MS/MS.

Table 1: Summary of mass spectrometry data sets

ID	Topic	Instrument	No. of spectra
PXD010324	Lung cancer	Orbitrap Fusion Lumos	7,865,965
PXD017407	Breast cancer, Brain cancer	Q Exactive HF	705,218

Sequence Variations

This novel analysis in order to identify protein sequence variation in high resolution MS2 data, that is based on SeVa Algorithm and the High-Resolution Exact P-Value (HR-XPV) [1] method, which is illustrated in Figure 3.

Sequence Variation



Figure 3: Figure shows BLAST sequence extension combined with post-translational modifications.

Main results

Main result analysis done on filtration parameters to obtain optimized FDR threshold.

Table 2: Our approach resulted in 2.98% FDR on PXD00224 and 2.02% FDR on PXD017407.

No.	Filtration			CPTAC LUAD			MHC				
	Peptide Mass	MinPct	Ref. X _{com}	eval _{com}	score	Target	Decoy	FDR	Target	Decoy	FDR
1	≤ 1500	≥ 90	≤ 2	≤ 1	≥ 3	13	0	0	3	0	0
2	≤ 1600	≥ 85	≤ 2	≤ 1	≥ 3	613	4	0.064	15	0	0
3	≤ 2000	≥ 85	≤ 1	≤ 1	≥ 3	2192	36	0.072	23	0	0
4	≤ 2000	≥ 90	≤ 1	≤ 1	≥ 3	1186	5	0.042	0	0	0
5	≤ 2000	≥ 90	≤ 2	≤ 1	≥ 3	2077	13	0.062	22	0	0
6	≤ 2000	≥ 90	≤ 2	≤ 10	≥ 3	5677	109	0.192	56	0	0
7	≤ 2000	≥ 90	≤ 3	≤ 1	≥ 1	4672	21	0.044	629	0	0
8	≤ 2000	≥ 90	≤ 3	≤ 1	≥ 2	4521	20	0.044	443	0	0
9	≤ 2000	≥ 90	≤ 3	≤ 1	≥ 2.5	3650	19	0.052	197	0	0
10	≤ 2000	≥ 90	≤ 3	≤ 1	≥ 3	2108	13	0.061	32	0	0
11	≤ 2000	≥ 90	≤ 3	≤ 2	≥ 3	3042	26	0.082	48	0	0
12	≤ 2000	≥ 90	≤ 3	≤ 5	≥ 3	4417	52	0.117	79	1	0.028
13	≤ 2000	≥ 92	≤ 2	≤ 1	≥ 3	1790	30	0.055	18	0	0
14	≤ 2000	≥ 95	≤ 2	≤ 1	≥ 3	1521	3	0.019	18	0	0
15	≤ 2500	≥ 85	≤ 2	≤ 1	≥ 3	2192	36	0.072	29	0	0
16	≤ 2500	≥ 85	≤ 2	≤ 10	≥ 1.5	16,950	280	0.169	2,165	44	0.020

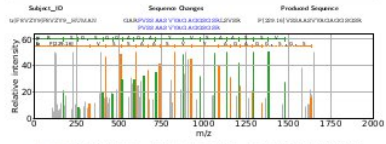


Figure 4: PXD0024 Peptide Spectrum Match plot for Scan(209), Charge 2, Seq: FLIPII4YGSANLTVAGAGDGSK

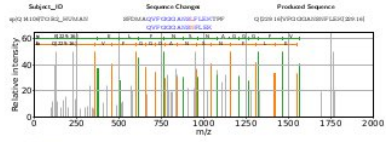


Figure 5: PXD017407 Peptide Spectrum Match plot for Scan(209), Charge 2, Seq: QGAPVAV AAH YVAG AEGSDGSLGVK

Bibliography

- Bhimani, K., Pansolina, A., Vornik, D., & Kertész-Farkas, A. (2024). Exact p-value calculation for XCorr scoring of high-resolution MS/MS data. *Proteomics*, 24(5), 2300145.
- Kertész-Farkas, A., Néi-Mátay Agayay, F. L., Bhimani, K., Eng, J. K., Földesi, W. E., Grant, C., ... & Noble, W. S. (2023). The GrantLab for analysis of bottom-up tandem mass spectrometry proteomics data. *Journal of proteome research*, 22(2), 961-969.

Гомотопические группы некоторых торических многообразий

Вылегжанин Ф. Е. [ORCID] стажёр-исследователь международной лаборатории алгебраической топологии и ее приложений

Гомотопические группы некоторых торических многообразий

Вылегжанин Фёдор Евгеньевич
ФКН НИУ ВШЭ, Москва



Симплициальные комплексы и верны

Определение. Симплициальный комплекс без граничных вершин \mathcal{K} на множестве вершин \mathcal{V} — это набор подмножества $I \subset \mathcal{V}$ (граней \mathcal{K}), содержащий все одноэлементные подмножества и замкнутый относительно включения ($I \subset J, J \in \mathcal{K} \Rightarrow I \in \mathcal{K}$).

Определение. Особый simplicial комплекс Σ с m вершинами в решетке \mathbb{Z}^d — это пара (\mathcal{K}, Λ) , где \mathcal{K} — simplicial комплекс на множестве $[m] = \{1, \dots, m\}$, а $\Lambda_1, \dots, \Lambda_m \in \mathbb{Z}^d$ — набор векторов со следующими свойствами:

- Для каждой грани $I \in \mathcal{K}$ набор $\{\Lambda_i; i \in I\}$ можно дополнить до базиса решетки \mathbb{Z}^d .
- Для любых двух граней $I_1, I_2 \in \mathcal{K}$ соответствующие открытые конусы $\sum_{i \in I_1} \mathbb{R}_+ \Lambda_i \subset \mathbb{R}^d, i = 1, 2$, не пересекаются.

На рисунке ниже изображены особый simplicial комплекс на \mathbb{Z}^2 соответствующий simplicial комплекс \mathcal{K} и матрица отображения $\Lambda: \mathbb{Z}^2 \rightarrow \mathbb{Z}^4, e_i \mapsto \Lambda_i$.

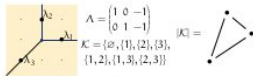


Рис. — Пример особого simplicial комплекса

Торические многообразия

Пусть $K^n = \mathbb{K} \setminus \{0\}$ — мультипликативная группа поля \mathbb{K} .

Определение. Торическое многообразие над \mathbb{K} — это нормальное алгебраическое многообразие X с регулярным алгебраическим тором $T = (K^n)^d$ в качестве открытого подмногообразия, такое что действие $T \curvearrowright X$ не имеет деления и про- долевается до алгебраического действия $T \curvearrowright X$.

Торические многообразия [KS] естественно возникают в алгебраической геометрии (как замкнутый орбит действительного тора), а также в комбинаторике и алгебраической статистике [DS].

Мы изучим топологию гладких торических многообразий над полем \mathbb{C} . Все такие многообразия получаются из следующей конструкции.

Конструкция Вейлера-Кона. Пусть $\Sigma = (K, \Lambda)$ — особый simplicial комплекс на \mathbb{Z}^d . Рассмотрим множество

$$\mathbf{U}(\mathcal{K}) := \bigcup_{I \in \mathcal{K}} (\mathbb{C}^*)^I \subset \mathbb{C}^m,$$

где

$$(\mathbb{C}^*)^I = \mathbb{Y}_1 \times \dots \times \mathbb{Y}_m, \quad \mathbb{Y}_i = \begin{cases} \mathbb{C}^*, & i \in I; \\ \{1\}, & i \notin I. \end{cases}$$

На нём координатно действует тор $(\mathbb{C}^*)^m$. Рассмотрим подгруппу

$$H_\Sigma := \text{Ker}(\Lambda_*: (\mathbb{C}^*)^m \rightarrow (\mathbb{C}^*)^d) \subset T^m.$$

Тогда действие H_Σ на $\mathbf{U}(\mathcal{K})$ свободно и собственнo, а $X_\Sigma := \mathbf{U}(\mathcal{K})/H_\Sigma$ — ториче- ские многообразие, на котором действует тор $T = (\mathbb{C}^*)^d/H_\Sigma \cong (\mathbb{C}^*)^d$.

Пример веры на рис. 1 соответствует $\mathbf{U}(\mathcal{K}) = \mathbb{C}^3 \setminus \{(0, 0, 0)\}$ и $H_\Sigma = \{(t, t, t) : t \in \mathbb{C}^*\}$; таким образом, $X_\Sigma = (\mathbb{C}^2 \setminus \{0\})/\text{diag}(\mathbb{C}^*) \cong \mathbb{C}P^2$. Другим тор $(\mathbb{C}^*)^2 = \{(t_1, t_2)\}$ действует на $\mathbb{C}P^2$ по формуле $\{t_1, t_2\} \cdot [x : y : z] = [x : t_1 y : t_2 z]$.

Связанные конструкции торической топологии

Обычно конструкции для пары пространства $A \subset X$ и simplicial комплекса \mathcal{K} на множестве вершин $[m]$ определяют топологические инварианты:

$$|K(A)^J| := \bigcup_{I \in \mathcal{K}} X^I \times A^{[m] \setminus I} \subset X^m.$$

Из частности $\mathbf{U}(\mathcal{K}) = (\mathbb{C}^*)^J$. Вместо того $\mathbf{U}(\mathcal{K})$ топологически эквивалентно по момент умнож. $Z_\Sigma := (\mathbb{D}^2, S^1)^J$, кроме того, для каждого гладкого торического многообразия X и некоторого топологического расслоения HP^2

$$(S^1)^{m-d} \rightarrow Z_\Sigma \rightarrow X, \quad (S^1)^d \rightarrow X \rightarrow (\mathbb{C}P^2, s)^J.$$

Это позволяет считать вычисление многих топологических инвариантов гладких торических многообразий и расслоения эквивалентной задаче. Например, если ве- ер Σ полный, то $H^*(X_\Sigma; \mathbb{Z}) \cong K[\mathcal{K}]/I_\Sigma$ (формула Дельфинга-Кларинга); здесь $K[\mathcal{K}] := K\langle \mathbb{Y}_1, \dots, \mathbb{Y}_m \rangle / (\prod_{i \in \mathcal{V}} \mathbb{Y}_i = 0, A \notin \mathcal{K})$ — кольцо Стейна-Рисера кон- дика \mathcal{K} и алгебра $I_\Sigma \subset K[\mathcal{K}]$ определяется в терминах отображения Λ .

Из выше вычислены гомоморфизмы групп гладких торических многообразий $\pi_n(X_\Sigma)$ удается проинтерпретировать сией следствием: для $n \geq 3$ и всех $\pi_n(X_\Sigma) \cong \pi_n(\mathbb{Z}^d) \cong \pi_n(\Omega \mathbb{Z}^d)$, а пространство веры $\Omega \mathbb{Z}^d$ иногда топологически эквива- лентно произведению более простых пространств.

Основной результат

Для каждого simplicial комплекса \mathcal{K} на множестве вершин \mathcal{V} и подмнож- ства $J \subset \mathcal{V}$ определяем полный simplicial комплекс $\mathcal{K}_J := \{I \in \mathcal{K} : I \subset J\}$. Элерова характеристика комплекса \mathcal{K} — это $\chi(\mathcal{K}) := 1 - \sum_{I \in \mathcal{K}} (-1)^{|I|}$.

Simplicial комплекс \mathcal{K} фантомный, если для любого подмножества $I \subset \mathcal{V}$ такое что $\{I, J\} \in \mathcal{K}$ для всех $J \in I$ веры $I \in \mathcal{K}$. Фантомные комплексы — это в частности алгебраические комплексы (например, комплексы Вейлера-Ривса из топологического анализа данных).

В следующей теореме топологическая эквивалентность упоминаема Л.Стеингома, а числа D_n вычислены доказательно.

Теорема. Пусть \mathcal{K} — фантомный simplicial комплекс на множестве вершин $[m]$. Тогда $\Omega \mathbb{Z}^d \cong \prod_{n \geq 3} (\Omega S^{2n})^{D_n}$, где числа $D_n \geq 0$ восстанавливаются из равенства формальных степенных рядов

$$\sum_{I \in \mathcal{K}} (1 - \chi(\mathcal{K}_I)) \cdot e^{II} = \prod_{n \geq 3} (1 - e^{2n})^{D_n}.$$

Следствие. Пусть $\Sigma = (K, \Lambda)$ — особый simplicial комплекс на \mathbb{Z}^d с m вершинами, где \mathcal{K} — фантомный комплекс. Тогда топологические группы соответствующего торического многообразия X_Σ выражаются через топологические группы сфер:


$$\pi_n(X_\Sigma) \cong \mathbb{Z}^{m-d} \oplus \bigoplus_{n \geq 3} \mathbb{Z}^{D_n} (S^{2n})^{D_n}, \quad \forall n \geq 3. \square$$

Для обобщения этих результатов на неформальный случай необходимо изу- чить комбинаторно-алгебраические объекты — алгебры Понка-масса Стейна-Ривера $\text{Ext}_{K[\mathcal{K}]}(K[\mathcal{K}], K[\mathcal{K}])$ (она позволяет как вычислять пространство веры $H_*(\mathbf{U}(\mathcal{K}; \mathbb{C}P^2, s)^J)$. За счет этого удается обобщить результаты выше на случай "почти фантомных" комплексов \mathcal{K} .

Список литературы

[1] V. M. Buchstaber, T. F. Frenk, "Topology of Math. Surv. Monogr. 204, AMS, 2015.
[2] D. A. Cox, J. B. Little, H. Sturmfels, "Toric varieties", Grad. Studies in Math. 130, AMS, 2011.
[3] M. Dehn, B. Sturmfels, S. Sullivant, "Combinatorics in Algebraic Statistics", Springer, 2010.
[4] L. Steenrod, "Long exact decompositions of moment-angle complexes associated to flag complexes", Quart. J. Math. 73(3) (2014) 417–427.
[5] F. Velephakis, "Long homology of moment-angle complexes in the flag case", arXiv:2401.18426

Tight and efficient upper bound on spectral norm of convolutional layers

Гришина Е. Р.  стажёр-исследователь научно-учебной лаборатории матричных и тензорных методов в машинном обучении



Tight and Efficient Upper Bound on Spectral Norm of Convolutional Layers

Ekatерина Grishina
Mikhail Gorbunov
Maxim Rakhuba
HSE University

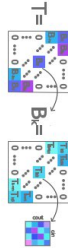


1. Motivation

- Controlling Lipschitz constant of convolutional layers
- Improve robustness of CNNs
- Enhances generalization performance
- Prevents explosion of the gradients

Prüschitt constant ($\rho(\text{Conv}) = \|\mathbb{T}\|_2$)

Consider the convolution of a $n \times n$ image with the kernel $K \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$. Its Jacobian $\mathbb{T} \in \mathbb{R}^{(n-n_1+1) \times (n-n_2+1) \times n_3 \times n_4}$ is a doubly block banded matrix.



Controlling the spectral norm is challenging. Computing $\|\mathbb{T}\|_2$ typically requires $\mathcal{O}(n^7)$ operations, where the image size n can be large.

2. Spectral norm of tensor

For a tensor $A \in \mathbb{R}^{n_1 \times \dots \times n_d}$ and vectors $u_i \in \mathbb{C}^{n_i}$ we define the following multilinear functional:

$$[A, u_1, u_2, \dots, u_d] = \sum_{i_1, \dots, i_d} A_{i_1, \dots, i_d} u_{1, i_1} \dots u_{d, i_d}$$

The spectral norm of a matrix $A \in \mathbb{R}^{n_1 \times n_2}$ is defined as:

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sqrt{\lambda_{\max}([A, u_1, u_2])} = \sqrt{\lambda_{\max}([A^T, u_2, u_1])}$$

Spectral norm naturally extends from matrices to tensors. For a d -dimensional tensor $A \in \mathbb{R}^{n_1 \times \dots \times n_d}$, it is defined as:

$$\|A\|_2 = \sqrt{\lambda_{\max}([A, u_1, \dots, u_d])}$$

Real and complex rank-1 approximations of a real tensor may not coincide.

3. Computation

Spectral norm of tensor is computed with Nesterov-Ostrov Power Method [1].

```

Algorithm 1 Power Method
Input: Tensor  $A \in \mathbb{R}^{n_1 \times \dots \times n_d}$ 
Output: Spectral norm  $\|A\|_2$ 
For  $i = 1, \dots, d$  do
     $u_i = \text{rand}(n_i)$ 
     $u_i = A u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_d$ 
     $u_i = \text{eig}(u_i) / \|u_i\|_2$ 
     $u_i = A u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_d$ 
     $u_i = \text{eig}(u_i) / \|u_i\|_2$ 
return  $\|A\|_2 = \sqrt{\lambda_{\max}([A, u_1, \dots, u_d])}$ 
    
```

4. Theoretical results

Theorem 1 For the 2D convolution with kernel $K \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$, the spectral norm of Jacobian is bounded as follows:

$$\|K\|_2 \leq \sqrt{1 + 3} \sqrt{\|K\|_1}$$

Theorem 2 For the convolution with stride s and kernel $K \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$,

$$\|K\|_2 \leq \sqrt{1 + 3} \sqrt{\frac{1}{s} \sum_{i=1}^n |K_i|}$$

where $Q \in \mathbb{R}^{(n-n_1+1) \times (n-n_2+1)}$ is zero padded and centered tensor K .

Theorem 3 For the 3D convolutional convolution with kernel $K \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$,

$$\|K\|_2 \leq \sqrt{1 + 3} \sqrt{\sum_{i=1}^n |K_i|}$$



Figure 2. Unfolding for convolution with kernel $1 \times 2 \times 3 \times 3$.

5. Connection to unfoldings

A matrix is called an **unfolding** of a tensor $A \in \mathbb{R}^{n_1 \times \dots \times n_d}$, if it is obtained by first reordering indices and then by reshaping the tensor into a matrix.

$$A_{i_1, \dots, i_d} \rightarrow (A_{i_1, \dots, i_d})_{(i_1, \dots, i_d)}$$

"Frobenius four" ($\| \cdot \|_F$) bound the spectral norm of tensor convolution with the minimum of spectral norms of four kernel unfoldings:

$\| \mathbb{T} \|_2 \leq \sqrt{\max}(\|A_{i_1, \dots, i_d}\|_F, \|A_{i_2, \dots, i_d}\|_F, \|A_{i_3, \dots, i_d}\|_F, \|A_{i_4, \dots, i_d}\|_F)$

Lemma 1 $\|K\|_2 \leq \sqrt{1 + 3} \|K\|_F$ for any reshaping matrix R of the tensor K .

Thanks to this lemma, our tensor norm (TV) bound is guaranteed to be more accurate than "Frobenius four" bound.

Sizes	$\ \mathbb{T}\ _2$	$\ R\ _F$	$\sqrt{1 + 3} \ R\ _F$	$\frac{\ \mathbb{T}\ _2}{\sqrt{1 + 3} \ R\ _F}$
64, 64, 3, 3	48.25	81.87	51.51	1.04
512, 512, 5, 5	227.1	501.8	229.6	1.01
512, 512, 7, 7	315.25	834.35	326.2	1.035

Table 1. Accuracy of the bounds, $n_2 \times n_3$ input size.

References:
 [1] Nesterov, A., Ostrov, S. "Power method for computing spectral norm of tensor." In: Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), pp. 2089–2092. IEEE, 2012.
 [2] Grishina, E. R., Gorbunov, M., Rakhuba, M. "Tight and efficient upper bound on spectral norm of convolutional layers." In: Proceedings of the 2024 International Conference on Artificial Intelligence and Deep Learning (AIDL 2024), pp. 1–10. IEEE, 2024.

6. Comparison to existing bounds

- Tensor norm bound is:
- Guaranteed to be more accurate than "Frobenius four" bound [3].
 - Independent of input size n .
 - Differentiable.
 - Fast to compute.
 - Applicable for convolution with any padding.

Method	Acc.	Mem.	Diff.	Ind. of n	Padding	Complexity (\mathcal{O})
Power method	+	+	+	+	zero	$n_1^2 n_2^2 n_3^2 n_4^2$
Singular [2]	+	+	+	+	scalar	$n_1^2 n_2^2 n_3^2 n_4^2$
Frobenius four [3]	+	+	+	+	any	$\mathcal{O}(n_1 n_2 n_3 n_4)$
TV bound	+	+	+	+	any	$\mathcal{O}(n_1 n_2 n_3 n_4)$
Gram basis	+	+	+	+	definite	$n_1^2 n_2^2 n_3^2 n_4^2$
PowerQR	+	+	+	+	definite	$n_1^2 n_2^2 n_3^2 n_4^2$
Tensor Norm (ours)	+	+	+	+	any	$\mathcal{O}(n_1 n_2 n_3 n_4)$

Table 2. Comparison of existing spectral norm computation methods. See our paper for their abbreviations.

7. Regularization

Regularization loss $\mathcal{L}_R = \mathcal{L}_{\text{mse}} + \lambda \|\mathbb{T}\|_2$.

$\|\mathbb{T}\|_2$ denotes the bound on spectral norm of the d th convolutional layer.

Method	Acc.	Info. wrt. Acc.	w/ wrt. Baseline	Method	Acc@1	Acc@5
F_1	73.10	23.94		Baseline	73.368	91.438
F_1	73.96	74.91		F_1	73.388	91.400
$T\mathbb{T}$ (Ours)	73.96	74.99		$T\mathbb{T}$ (Ours)	73.510	91.400

Table 3. Results on CIFAR100.

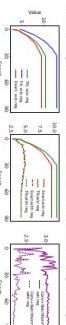


Figure 3. One layer of ResNet-18.


Orthogonality of convolutional maps has been shown to improve training stability for CNNs and reduce model overfitting.

Orthogonal Loss $\mathcal{L}_{\text{Orth}} = \|\text{Conv}(K, \text{padding}) - \text{"convd"}\|_F$.

Method	Acc@1	Acc@5	Time (s)
Baseline	73.38	93.16	141.0
F_1	73.39	93.14	141.0
$T\mathbb{T}$ (Ours)	73.51	93.14	141.0
200m Conv	73.99	94.08	121.8

Table 5. Results on CIFAR100 with orthogonal loss.

Применение Apache Kafka в высоконагруженных приложениях

Данилов А. А.  м.н.с. научно-учебной лаборатории облачных и мобильных технологий



Основные задачи

Фронтенд (WebUI) – сервер, состоящий из фронтенда. Основная задача фронтенда – взаимодействие с клиентом. Тестовый сервер – сервер, который проверяет данные. Тестовый сервер – сервер, который проверяет данные. Тестовый сервер – сервер, который проверяет данные. Тестовый сервер – сервер, который проверяет данные.

Продуктивная способность

Одним из ключевых преимуществ Apache Kafka является способность обрабатывать огромные объемы данных в реальном времени. Благодаря этому, Kafka идеально подходит для приложений, требующих высокой скорости и масштабируемости. Кроме того, Kafka поддерживает различные форматы данных, что позволяет легко интегрировать данные из различных источников. Это делает Kafka идеальным решением для высоконагруженных приложений, требующих высокой скорости и масштабируемости.

Технология Zero-сюр

Одним из ключевых преимуществ Apache Kafka является способность обрабатывать огромные объемы данных в реальном времени. Благодаря этому, Kafka идеально подходит для приложений, требующих высокой скорости и масштабируемости. Кроме того, Kafka поддерживает различные форматы данных, что позволяет легко интегрировать данные из различных источников. Это делает Kafka идеальным решением для высоконагруженных приложений, требующих высокой скорости и масштабируемости.

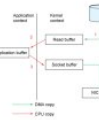


Рис. 2. Представление Zero-сюр архитектуры.

Zero-сюр архитектура позволяет избежать копирования данных между различными компонентами системы. Это достигается за счет использования общих буферов и адресов памяти, что значительно снижает нагрузку на процессор и повышает производительность.



Рис. 3. Представление Zero-сюр архитектуры.

Zero-сюр архитектура позволяет избежать копирования данных между различными компонентами системы. Это достигается за счет использования общих буферов и адресов памяти, что значительно снижает нагрузку на процессор и повышает производительность.

Применение Apache Kafka в высоконагруженных приложениях

приложениях

Данилов Алексей Андреевич
ФНИ НИУ ВШЭ, Москва

Эта работа посвящена применению Apache Kafka в высоконагруженных приложениях. В статье рассматриваются основные задачи, стоящие перед разработчиками, и предлагаются решения. Также описаны различные сценарии использования Kafka и способы интеграции с другими технологиями.

Сжатие данных при передаче и хранении

Сжатие данных является важным аспектом работы с данными, особенно при передаче и хранении. В статье рассматриваются различные алгоритмы сжатия и их применение в Apache Kafka. Также описаны способы настройки Kafka для оптимальной работы с сжатыми данными.

№	Имя	Ссылка	Тип	Статус	Комментарий
1	1	https://github.com/...	репозиторий	активен	Инициализация репозитория
2	2	https://github.com/...	репозиторий	активен	Добавление документации
3	3	https://github.com/...	репозиторий	активен	Добавление тестов
4	4	https://github.com/...	репозиторий	активен	Добавление примеров
5	5	https://github.com/...	репозиторий	активен	Добавление тестов
6	6	https://github.com/...	репозиторий	активен	Добавление документации
7	7	https://github.com/...	репозиторий	активен	Добавление тестов

1. Проверка статуса репозитория.
2. Создание репозитория Apache Kafka.
3. Добавление документации.
4. Добавление тестов.
5. Добавление примеров.
6. Добавление документации.
7. Добавление тестов.

Масштабирование и отказоустойчивость

Масштабирование и отказоустойчивость являются ключевыми аспектами работы с данными в высоконагруженных приложениях. В статье рассматриваются различные способы масштабирования и обеспечения отказоустойчивости.



В статье рассматриваются различные способы масштабирования и обеспечения отказоустойчивости. Также описаны способы настройки Kafka для оптимальной работы с данными. Кроме того, рассматриваются различные форматы данных и способы интеграции с другими технологиями.

Результаты исследования продуктивной способности

Для исследования продуктивной способности Apache Kafka были проведены различные тесты. В статье описаны результаты этих тестов и предложены способы оптимизации производительности.

1. Apache Kafka Framework CPU & RAM
2. Apache Kafka Framework CPU & RAM
3. Apache Kafka Framework CPU & RAM

Результаты тестов показывают, что Apache Kafka обладает высокой продуктивной способностью и может использоваться в высоконагруженных приложениях.



Рис. 4. Результаты тестирования продуктивной способности.

В статье описаны различные способы масштабирования и обеспечения отказоустойчивости. Также описаны способы настройки Kafka для оптимальной работы с данными.

Заключение

Apache Kafka является идеальным решением для высоконагруженных приложений, требующих высокой скорости и масштабируемости. Благодаря Zero-сюр архитектуре, Kafka обеспечивает высокую производительность и снижает нагрузку на процессор.

Список литературы

1. Apache Kafka. <https://kafka.apache.org/>
2. Kafka. <https://kafka.apache.org/>
3. Kafka. <https://kafka.apache.org/>
4. Kafka. <https://kafka.apache.org/>
5. Kafka. <https://kafka.apache.org/>
6. Kafka. <https://kafka.apache.org/>
7. Kafka. <https://kafka.apache.org/>

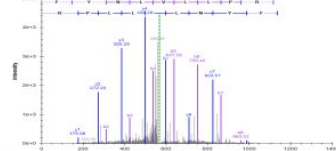
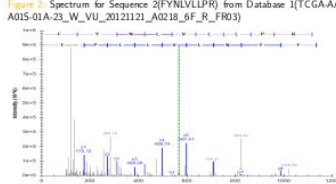
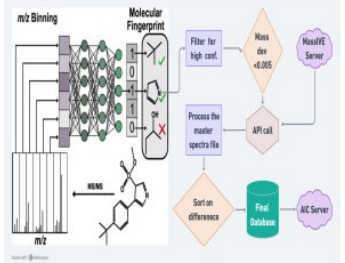
Large-scale evaluation of the generalization performance of deep neural networks in tandem mass spectrometry data analysis

Джоши К. [✉] стажёр-исследователь научно-учебной лаборатории искусственного интеллекта для вычислительной биологии

Large-scale evaluation of the generalization performance of deep neural networks in tandem mass spectrometry data analysis



Kartik Joshi*, Attila Kertesz-Farkas**
 AIC Lab, FCS NRU HSE, Moscow
 k-dzhoshi@hse.ru* Kfattila@yandex.com**

Basic concepts	Examples
<p>Tandem Mass Spectrometry (MS): Identifies compounds by fragmenting molecules and analyzing ions.</p> <p>Deep Neural Networks (DNNs): Used for peptide identification by learning from extensive datasets to match sequences to spectrum.</p> <p>Generalization in DNNs: The ability to perform well on unseen data, crucial for biological and chemical datasets like MS.</p> <p>Overfitting in MS Analysis: Occurs when models memorize data from specific databases, hindering generalization and leading to biased results in MS research [1].</p>	<p>To demonstrate the differences in spectrum for the same sequence across various databases [Fig 2 and Fig 3, we utilized the SeeMS [4] software to extract spectrum images.</p>
<p>Historical background</p> <p>Evolution to Machine Learning: MS analysis has shifted from traditional algorithms to ML methods, struggling with limited and homogeneous datasets.</p> <p>Deep Learning Adoption: DNNs excel at identifying complex patterns in MS data, yet face generalization challenges with similar training sets.</p> <p>Data Sparsity Issues: A lack of comprehensive MS datasets across various instruments and conditions hinders model performance and generalization.</p>	
<p>Key Results</p> <p>Peprnet Testing: The Peprnet DNN model processed ~40K MS sequences, identifying ~1.5K with high confidence. [2]</p> <p>Sequence Filtering: Approximately 100 correctly structured peptides were filtered, demonstrating the model's generalization ability. [3]</p> <p>Spectra Analysis: Around 80 spectra files from various databases were analyzed, with up to three perturbations per file for robustness. [3]</p> <p>Key Finding: Heterogeneous datasets are crucial; diverse testing enhances model robustness and reliability across MS environments. [1]</p> <p>Dataset Size: We processed ~30 TB of data through the pipeline shown in Figure 1.</p>	
<div style="display: flex; justify-content: space-between;"> <div data-bbox="192 1053 537 1316">  </div> <div data-bbox="543 1053 929 1382"> <p>Conclusions and further research directions</p> <p>Need for Diverse Datasets: The research emphasizes the necessity of varied datasets from different labs and instruments to assess DNN generalization in MS.</p> <p>Broader Data Inclusion: Collecting datasets with diverse environmental conditions and instrumentation is vital for accurate model performance.</p> <p>Future Testing: Improving model accessibility for MS analysis requires ongoing testing with additional perturbations and an expanded dataset.</p> <p>References</p> <ol style="list-style-type: none"> [1] Mizadeh, I., Alizadeh, K., Shalimi, H., Tazai, O., Bengio, S., & Farahfar, M. (2024). GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. <i>arXiv preprint</i>. Available at: https://arxiv.org/abs/2410.05229 [2] PeprNet: Deep CNN for De Novo Peptide Sequencing. <i>GitHub Repository</i>. Available at: https://github.com/ProtosomaTeam/PeprNet [3] MassIVE: Mass Spectrometry by Interactive Virtual Environment. <i>UCSD Center for Computational Mass Spectrometry</i>. Available at: https://massive.ucsd.edu/ProtocoolAP/atic/massive.jsp [4] SeeMS: Software for Viewing Mass Spectra. <i>ProteinWard Toolkit</i>. Available at: http://protonwizard.sourceforge.net/100.html </div> </div>	

Guide-and-Rescale: self-guidance mechanism for effective tuning-free real image editing

Иванова А. А. [ORCID] стажёр-исследователь лаборатории теоретических основ моделей искусственного интеллекта



Self-Guidance Mechanism for Effective Tuning-Free Real Image Editing

Vadim Titov^{1*}
¹HSE University ²AIRI

Madina Khaimolova^{4*} Alexandra Voronova^{2,3*}
³Skolkovo Institute of Science and Technology ⁴UNSW Sydney

Dmitry Vetrov⁵ Adhek Adipriyo²
⁵Complutense University, Belgium

* Equal contribution

Main Contributions

- Propose a novel pipeline of image editing **Guide-and-Rescale** that
- Propose **rescaling function (guides)** based on user feedback and self-attention maps to utilize them in the guidance mechanism.
- Introduce a **noise-rescaling technique** that allows balancing the prompt of classifier-free guidance and our proposed guides.



Fig. 1: Guides and Effects for real image editing.

Proposed Method

The **our method** uses a classic dim-inversion of the source image, denoising process, and denoising process. For every denoising step the noise term is modified by guides that utilize prompts s_i . From current generation process and time-aligned SDPA latents z_t .

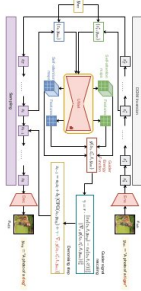


Fig. 2: Overall scheme of the proposed method. Guide-and-Rescale.

Qualitative Comparison

Guide-and-rescale allows to manipulate images for a wide range of editing prompts. Our method shows **more consistent results** than existing methods and achieves a **better trade-off** between **editing quality** and the **adversarial rejection** of the original image.



Fig. 3: Visual comparison of our method and adversarial rejection. Trade-off of editing.

Quantitative Comparison

Our method achieves the best balance between editing quality (CQIP) and preservation of the original image (UPPS) and achieves the best similarity to the real images (FID).

Table 1: Quantitative comparison of our method with baselines on image editing task.

Method	CQIP	UPPS	FID
ControlNet	0.85	0.92	0.12
ControlNet+Kandinsky	0.82	0.88	0.15
ControlNet+Kandinsky+Kandinsky	0.80	0.85	0.18
ControlNet+Kandinsky+Kandinsky+Kandinsky	0.78	0.82	0.20
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.75	0.78	0.25
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.72	0.75	0.30
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.70	0.72	0.35
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.68	0.70	0.40
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.65	0.68	0.45
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.62	0.65	0.50
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.60	0.62	0.55
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.58	0.60	0.60
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.55	0.58	0.65
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.52	0.55	0.70
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.50	0.52	0.75
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.48	0.50	0.80
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.45	0.48	0.85
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.42	0.45	0.90
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.40	0.42	0.95
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.38	0.40	1.00

Table 1: Quantitative comparison of our method with baselines on image editing task.

User Study

Our method **subjectively gives a better result** on editing quality while preserving source image prompt consistency for real editing.

Method	Mean	Std
ControlNet	0.85	0.12
ControlNet+Kandinsky	0.82	0.15
ControlNet+Kandinsky+Kandinsky	0.80	0.18
ControlNet+Kandinsky+Kandinsky+Kandinsky	0.78	0.20
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.75	0.25
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.72	0.30
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.70	0.35
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.68	0.40
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.65	0.45
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.62	0.50
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.60	0.55
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.58	0.60
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.55	0.65
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.52	0.70
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.50	0.75
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.48	0.80
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.45	0.85
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.42	0.90
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.40	0.95
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.38	1.00

Table 2: User study that we analyze the subjective preference of our method compared to all baselines.

Ablation Study

Jointly applying both guides preserves both visual characteristics of unedited regions of the image. Noise rescaling technique **improves editing** and improves its quality.



Fig. 4: (A) Standard of applying noise rescaling. (B) Standard of applying noise rescaling.

Each guide is responsible for preserving different aspects of the original image, and combining them leads to **good performance in terms of CQIP/FID**. The results with noise rescaling **surpass more details** from the original image, as reported by UPPS improvement.

Method	CQIP	UPPS	FID
ControlNet	0.85	0.92	0.12
ControlNet+Kandinsky	0.82	0.88	0.15
ControlNet+Kandinsky+Kandinsky	0.80	0.85	0.18
ControlNet+Kandinsky+Kandinsky+Kandinsky	0.78	0.82	0.20
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.75	0.78	0.25
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.72	0.75	0.30
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.70	0.72	0.35
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.68	0.70	0.40
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.65	0.68	0.45
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.62	0.65	0.50
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.60	0.62	0.55
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.58	0.60	0.60
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.55	0.58	0.65
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.52	0.55	0.70
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.50	0.52	0.75
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.48	0.50	0.80
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.45	0.48	0.85
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.42	0.45	0.90
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.40	0.42	0.95
ControlNet+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky+Kandinsky	0.38	0.40	1.00

Table 3: Quantitative comparison of our method with the results of the proposed noise rescaling components.

HF Demo
 Code
 Project Page

Spectrum annotation with transformer models

Кривулец Н. В. [✉] стажёр-исследователь научно-учебной лаборатории искусственного интеллекта для вычислительной биологии

Spectrum annotation with transformer models



Nina Krivulets, Attila Kertesz-Farkas
AIC Laboratory, ФКН НИУ ВШЭ, Москва



Core concepts

Mass spectrometry is an analytical technique used to measure the mass-to-charge ratio (m/z) of ions. It helps identify the chemical composition of a sample and determine the structure of molecules. It is widely used in chemistry, biology, and related fields for analyzing small molecules, proteins, metabolites, and other substances.

Mass spectrum: A graphical representation showing the intensity of ions at various m/z ratios.

Each peak in the mass spectrum represents an ion with a specific m/z ratio, corresponding to a molecular fragment or the intact molecule, and heights of the peaks denote the relative abundance of the various components in the sample.

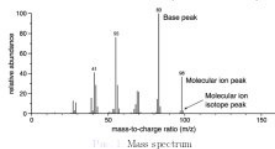


Fig. 1 Mass spectrum

Architecture overview

In this study, transformer models are used for deep learning. These models were initially introduced in the paper [1].



Fig. 2 Transformer architecture

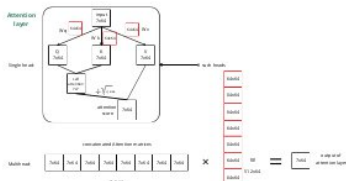


Fig. 3 Attention layer

The architecture consists of an encoder and a decoder, both made up of multiple layers. Each encoder layer has two main components: multi-headed self-attention and position-wise feedforward networks.

The decoder, similarly structured, includes additional attention layers that focus on the encoder's output.

Current progress

For now, we were able to verify the data, as one of the problems in mass spectrometry is that 60-80% of the experimental spectra cannot be annotated correctly with high confidence.

Positional encoding based on the article [1] has been implemented for the data. Formula:

$$PE_{(pos,2i)} = \sin\left(\frac{p}{10000^{2i/d_{model}}}\right), PE_{(pos,2i+1)} = \cos\left(\frac{p}{10000^{2i/d_{model}}}\right)$$

where p is the peak value, i is the position and d_{model} is the dimension of the model.

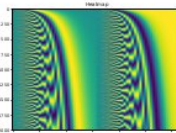


Fig. 4 Heatmap of the positional encodings

Future research

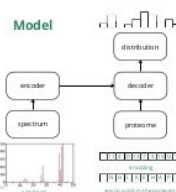


Fig. 5 Proposed model

The transformer neural network is currently being implemented, and we aim to extend this transformer model to predict the location of the peptide (which generated the observed spectrum) in the proteome. This involves incorporating the entire proteome, represented as strings of amino acids, into the decoder. Overall, the model will take peaks as an input for the encoder and strings of amino acids as an input for the decoder. Positional encoding and embedding will be applied respectively to attain the proper matrices. The outputs of the encoder will be used in the decoder, in the 'encoder-decoder attention' layer, along with the decoder's input. Afterwards, the model will produce a distribution over the proteome of the location of the peptide.

References

- [1] Ashish Vaswani and Noam Shazeer and Nikhil Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Łukasz Kaiser and Illia Polosukhin. "Attention Is All You Need" arXiv:1706.03762
- [2] Jimmy Lee Ba and Justin Ryan Kiros and Geoffrey E. Hinton. "Layer Normalization" arXiv:1607.08019.

Подход к созданию сервиса генерации программного кода мобильных приложений

Резуник Л. [2] стажёр-исследователь научно-учебной лаборатории облачных и мобильных технологий

Подход к созданию сервиса генерации

программного кода
мобильных приложений

Резуник Людмила Александровна
ФКН НИУ ВШЭ, Москва



Основные понятия

LLM (Large Language Model), большая языковая модель – модель машинного обучения, основанная на нейронной сети со множеством параметров, обученная на большом количестве текстовых данных.

“Задача «description-to-code» (описанием в код) – разновидность задачи [1], основанная на получении описания на естественном языке, также называемый промпт, и генерации программного кода на его основе.

Бенчмарк – эталон, по которому можно измерить производительность (в данном исследовании – качество ИИ).

Предпосылки

Существующие большие языковые модели редко адаптируются в сторону программной оболочки для мобильных разработчиков, хотя это привнесло бы пользу отрасли. Первоначально была поставлена практическая задача – разработать сервис для генерации программного кода мобильных приложений с использованием ИИ. Также, было необходимо исследовать производительность и качество существующих open-source LLM, предложить варианты к оценке моделей генерации программного кода.

Вопросы исследования

- Как оценить качество и производительность LLM?
- Существует ли зависимость между качеством генерации, производительностью и типом поставленной задачи?
- Какое качество ответов различных open-source LLM на вопросы? Как мы можем добиться лучших результатов?

Основные результаты

По итогам проекта был реализован сервис для генерации кода мобильных приложений. Система предоставляет пользователю основную функциональность через клиентские приложения (Рис. 1) и отдельные микросервисы. Проект также включает в себя модель искусственного интеллекта, который был создан на базе существующего LLM Gpt4Qna.



Рис. 1. Окно диалогов с ИИ-моделью в разрабатываемом сервисе

В процессе разработки были даны ответы на вопросы исследования, а также было выведено, что на данный момент не существует какого-либо универсального бенчмарка [2] или теста для оценки качества ответов LLM.

В связи с этим был создан бенчмарк и тестовый клиент. Для измерения качества модели применялись та же метрика pass@k, что и в известном бенчмарке HumanEval.

$$pass@k = 1 - ((n - c)/k)/(n/k) \quad [1]$$

где n – общее количество тестовых случаев, c – число успешно пройденных тестов, k – число попыток.

Анализ качества и производительности моделей

Были протестированы следующие LLM: Gpt4Qna, Madaider, CodeLlama, DeepMagic, Coder, Mistral. Было выведено, что качество генерации и скорость генерации может варьироваться в зависимости от типа задачи написания кода. Задачи по генерации простых алгоритмов выполняются моделями качественнее и быстрее, сложнее же генерация UI-компонентов.

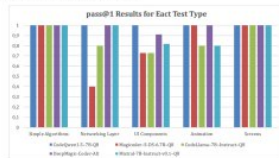


Рис. 2. Качество генерации кода в зависимости от задачи

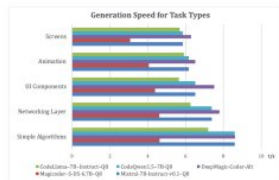


Рис. 3. Скорость генерации в зависимости от задачи

Выводы и дальнейшие направления исследований

В настоящее время разрабатываемый сервис может быть использован в образовательных целях и для тестирования данной концепции как MVP. Артефакты, полученные в результате исследования (бенчмарк, клиент для тестирования, результаты оценки производительности и качества) могут использоваться как ориентир для выбора LLM разработчиками аналогичных сервисов.


В рамках дальнейшей работы можно рассмотреть возможность автоматизации тестирования для оценки качества моделей, а также разработку алгоритмов, наиболее подходящих конкретно под задачу оценки качества сгенерированного программного кода.

Список литературы


- [1] E. DiLucca, D. Bardolino, S. Baldo, S. Gaud, W. Meert. Code Generation Using Machine Learning: A Systematic Review. 2022. IEEE Access. 10.1.1.10.1109/ACCESS.2022.3196347.
- [2] Michael Eitzinger, Egor Boronov, Y. Sotirov, and Timothy Birkus. “Out of the BLEU: How should we assess quality of the Code Generation models?”. Journal of systems and software: The Journal of systems and software, vol. 201, pp. 111741–111741, Sep. 2023.

False discovery rate control in classification problems with knock-off classes

Ромашкина А. Д. [📧] стажёр-исследователь научно-учебной лаборатории искусственного интеллекта для вычислительной биологии



Arina Romashkina, Attila Kertesz-Farkas
AIC Laboratory, ФКН НИУ ВШЭ, Москва



False discovery rate control in classification problems with knock-off classes.

Research dictionary

- Classification – a classic machine learning task implying a supervised approach to elaborate an algorithm that reveals the key features of a given number of classes to correctly categorize new observations.
- P-value – denotes the probability of events' occurrence in case of the null-hypothesis (H0) truthfulness.
- False discovery rate – a proportion of samples that belong to type I errors. In other words, this is simply a share of observations, likely rejecting H0.
- FDR control – a family of statistical procedures aimed at formulating a certain level of metric's value for particular number of discoveries.

Existing approaches

The **Benjamini-Hochberg procedure** – ubiquitously recognized FDR-controlling step-up technique, where the metric is being set to a certain level. After calculating for each observation its critical values, method formulates the boundary of acceptance as the higher rank, where critical value is greater than the corresponding p-value.

The new approach

The new approach is based on extracting features highlighted by the model at the penultimate layer and then passing them to the generative model, where the target values are from null distribution. The task of the discriminator is to determine whether this is an originally null score or one generated. After that, we compare the scores of the model and the generator for each object, and if the model's score is higher, we consider the object to be positive.




Fig. 1. The main idea.

So we take features from the model to train GAN in order to get new scores that will match the null distribution from the model.

Then calculate statistics following algorithm:

- $\text{max_score} = \max(\text{scores_GAN}, \text{scores_CNN})$
- $\text{label} = \begin{cases} 0, & \text{if scores_GAN} > \text{scores_CNN} \\ 1, & \text{else} \end{cases}$
- sort all items by max_score
- $\text{rank} = \text{index in sorted list} + 1$
- $\text{FDR} = \frac{\text{number of positive}}{\text{number of positive} + 1}$
- $q_i = \min(q_{i-1}, \text{FDR}_i)$
- then plot q_i vs q_i vs q_i

The main advantage is that this approach will get the same results if there is a data shift! For example, if we decrease intensity of pixels in the picture this approach will still give the expected result.

My experiments

I conducted my experiments on the MNIST dataset. The model is CNN, GAN includes linear generator and discriminator.

In the case of binary classification, the images are divided into two classes: 5 and not 5. Then, for each image, I determine the score produced by the CNN. The threshold is set at 0, so images with a score > 0 are considered to be images of 5. Then features of each object are assigned to target value as:

$$\text{score}_{\text{null}} = \begin{cases} \text{score}_{\text{CNN}} & \text{label} = \text{negative} \\ 0, 2 * \text{score}_{\text{CNN}} & \text{label} = \text{positive} \end{cases}$$

Then GAN is trained with two learning objectives. Consider: f - features, G - generator, D - discriminator.

1) The generator should force the discriminator to accept the generated data as real, also the distance between generated score and null should be minimised.

$$\text{loss}_{\text{SGE}} = -\frac{1}{n} \sum_{i=1}^n \log(D(G(f)_i) + \epsilon)$$

2) Also the distance between generated score and null should be minimised.

$$\text{loss}_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^n (G(f)_i - \text{null}_i)^2$$

And additional penalty is added when mistake is in region near 0.

$$\text{loss}_{\text{MSE additional}} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\text{null}(c) - 2\sigma_1} \cdot 10 \cdot (G(f)_i - \text{null}_i)^2$$

$\text{generator_loss} = \text{loss}_{\text{SGE}} + \text{loss}_{\text{MSE}} + \text{loss}_{\text{MSE additional}}$

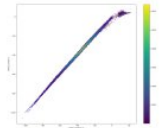


Fig. 2. The scatter plot of GAN's and CNN's scores for each object.

Then, for the test set, we compare the scores produced by the model and the generator. Wherever the GAN's score is higher, we predict the negative class, otherwise, the positive class.

Then we decrease intensity of each picture and try all steps again. And compare FDR curves with ground truth.

In the case of multiclass classification, the vector of scores and features for each picture is returned, then target distribution is formed by excluding s scores that correspond to correct label and GAN was trained. Then in each vector maximum is taken, scores of CNN and GAN are compared as in binary case.

Current results.

The FDR curves look like that now. The next step is to improve quality and accuracy.

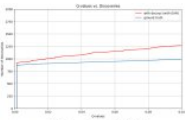


Fig. 3. FDR for binary MNIST case.

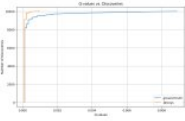


Fig. 4. FDR for multiclass MNIST case.

Принципы формирования трехмерных изображений или количественная фазовая визуализация

Талайкова Н. А. [ORCID] менеджер научно-учебной лаборатории методов анализа больших данных

Принципы формирования трехмерных изображений или количественная фазовая визуализация



Талайкова Наталья Анатольевна

LAMBDA ФКН НИУ ВШЭ, Москва



Количественная фазовая визуализация

Fig. 1. Оптическая схема.
 LS – источник освещения, ОО – исследуемый объект,
 МО – объектив оптического микроскопа, ML – окуляр микроскопа,
 1,1,1,2 – линзы, SF – пространственный фильтр,
 Cam – CMOS камера. Изображение – схема пространственного фильтра

Фото дифракционного фазового микроскопа

В данной работе мы используем метод дифракционной фазовой микроскопии, который позволяет получать не только привычные нам двумерные фотографии как в обычном фотоаппарате или оптическом микроскопе, но так же и трехмерные. Из таких изображений мы можем получать информацию не только качественную, но и количественную. Такая группа методов в микроскопии получила собственное название – методы количественной фазовой визуализации.

В этой работе мы опишем теорию формирования таких трехмерных изображений и сравним ее ранее полученными экспериментальными данными.

Теория формирования трехмерного изображения

Для описания формирования изображения можно использовать основное уравнение голографии, используя функции интенсивности объектного и оверного полей в неканоничном виде:

$$I(x, y) = I_0(x, y) + I_1(x, y) + 2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x' y' \phi(x', y') dx' dy' \quad (1)$$

где $I_0(x, y)$ и $I_1(x, y)$ – функции интенсивности объектного поля и оверного поля, соответственно, $\phi(x, y)$ – заданная среда при пропускании света через объект. Но нас интересует как именно меняется результат после обработки изображения, исходя из параметров системы или исследуемого объекта.

Предположим, что полуотражающий объект освещается светом на одной частоте, например, только красного света $E_0(x, y)$, ϕ и освещение равномерно и полностью освещает объект и проходит через него.

Для простоты представим объект $f_0(x, y)$ в виде ступенчатой функции.

$$f_0(x, y) = a_0 \cdot \text{step}(x - x_0, y - y_0) \quad (2)$$

$$f_0(x, y) = a_0 \cdot \exp(i \cdot \text{phase}(x, y)) \quad (3)$$

$$f_0(x, y) = \begin{cases} a_0 - 1, & x < x_0 \\ a_0, & x_0 \leq x \leq x_0 + b \\ a_0 + 1, & x > x_0 + b \end{cases} \quad (4)$$

$$f_0(x, y) = \text{step}(x - x_0, y - y_0) \cdot \text{step}(y - y_0) \quad (5)$$

В процессе формирования изображения микроскопом, так же устанавливая дифракционную решетку, световое поле оказывается преломленным спектовыми функциями:

$$f_1(x, y) = f_0(x, y) \otimes \delta(x - x_0, y - y_0) \quad (6)$$

где $f_1(x, y)$ – комплексная амплитуда увеличенного светового поля, x, y – комплексная амплитуда увеличенного светового поля и $\delta(x, y)$ – функция пропускания дифракционной решетки.

Далее, переходя в плоскость пространственного фильтра. Перед ней устанавливается первая линза системы, т.е. прибор, который осуществляет фурье-преобразование [1], потому в плоскости пространственного фильтра мы применим свертку для функции пропускания пространственного фильтра, светового поля, объекта и дифракционной решетки. Для простоты представим функцию пропускания дифракционной решетки в виде лабра дельта-функций [2], а пространственный фильтр в виде набора прямоугольных функций с осями h и h' :

$$F_0(x, y) = \sum_{m=1}^M \sum_{n=1}^N \text{rect}(x - x_m, y - y_n) \otimes \delta(x - x_m, y - y_n), \quad m=1, \dots, M, \quad (7)$$

$$\delta(x, y) = \delta(x - x_m) \otimes \delta(y - y_n), \quad \delta \geq \frac{1}{2} \sqrt{\frac{\lambda z}{L}}, \quad (8)$$

$$\text{when } \delta(x, y) = \begin{cases} 1, & |x - x_m| \leq \delta/2, \\ 0, & |x - x_m| > \delta/2, \end{cases} \quad \text{when } \delta(x, y) = \begin{cases} 1, & |y - y_n| \leq \delta/2, \\ 0, & |y - y_n| > \delta/2, \end{cases} \quad (9)$$

Получим поле в плоскости детектора:

$$f_2(x, y) = F_0(x, y) \otimes f_1(x, y) = \sum_{m=1}^M \sum_{n=1}^N \text{rect}(x - x_m, y - y_n) \otimes f_1(x, y) \otimes \delta(x - x_m, y - y_n) + \dots \quad (10)$$

Мы получим описание формирования изображения при подстановке функций в явном виде.

Изображение и способ его обработки

Fig. 2. Изображение с полками, из которого можно ивлечь информацию о высоте объекта по его углу наклона в объективном пространстве

Fig. 3. Изображение, полученное методом количественной фазовой визуализации результата его обработки и изображение, полученное на детекторе, на как скэне – увеличенный фрагмент изображения, б – фотографии того же объекта, в – спектр изображения, г – результат применения алгоритма обработки изображения [3]

Выводы и дальнейшие направления исследований

Так как рассмотренный метод является комбинацией нескольких методов визуализации, то описание теории формирования изображений в явном виде позволяет:

- расширять теорию формирования голографических изображений в целом,
 - расширять возможность использования микроскопа в зависимости от объекта, а так же проводить эксперименты без дорогостоящей сборки оборудования,
 - снижать ошибку измерения в методах количественной фазовой визуализации или показать в явном виде вероятность ошибки в зависимости от параметров системы.
- Например, при увеличении ося оверного порядка до "объектного", мы получаем новую информацию о высоте объекта, как на рис. 6. И далее, мы планируем расширить теорию на более практически значимые функции, например, для освещения цилиндрической гроссумо функции, цилиндрической функции освещения лазера, а так же использовать функции дифракционных решеток, которые изготавливаются в данный момент ведущими оптическими компаниями, например, Thordlabs или Edmund Optics.

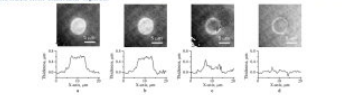


Fig. 6. Зависимость высоты объекта от размера ося оверного порядка от а до: изменение от минимального размера к максимальному (сплошная линия как бы прерывається в объектное и мы видим делительно сигнала об объекте при изменении в области изображения лагу "объектного" на угол луге на друг

Список литературы

[1] P.P. Shchegolev, "Описание голограммы", Изд. Высшеского, Днепропетров, 2014 стр. 344.
 [2] A. Kalyanas, N. Saldanova, V. Rybalchko, "Formal theory of Affine phase microscopy", Surface Ball Meeting 2014 Optical Technologies in Biophysics and Medicine XVI, Laser Phys and Photonics XVI, Computational Biophysics, 2015, vol. 9446, pp. 944617-1-8.
 [3] N. Talaykova, V. Rybalchko, "Computer Affine phase microscopy for quantitative visualization of cells in biomedical applications", Journal of Physics: Conference Series, 2016, vol. 737, pp. 1308-611.

Group and Shuffle: efficient structured orthogonal parametrization

Юдин Н. Е. [ORCID] стажёр-исследователь научно-учебной лаборатории матричных и тензорных методов в машинном обучении



Group and Shuffle: Efficient Structured Orthogonal Parametrization (NeurIPS'24)

Mikhail Gorbunov, Nikolay Yudin, Vera Soboleva, Aibek Alanov, Alexey Naumov, Maxim Rakhuda
FCS HSE, Moscow



arXiv:2405.10010

Introduction

We propose

- A new class of \mathcal{GS} matrices. It requires less matrix products than the previous approach HOFT to form a dense matrix.
- Efficient structured orthogonal parametrization of \mathcal{GS} matrices.
- To apply it in orthogonal fine-tuning (LLMs and diffusion models), and orthogonal convolutional architectures.

Orthogonal Fine-Tuning

Orthogonal Fine-tuning framework modifies forward pass of pre-trained linear layers

$$\mathbf{y} = (\mathbf{W}^0)^T \mathbf{x} \rightarrow \mathbf{y} = (\mathbf{Q}\mathbf{W}^0)^T \mathbf{x}; \quad \mathbf{Q} \text{ orthogonal matrix}$$

- OFT method uses block diagonal structure for \mathbf{Q} parameterizing it as

$$\mathbf{Q} = \text{diag}(\mathbf{Q}_1, \mathbf{Q}_2, \dots, \mathbf{Q}_k),$$

where $\mathbf{Q}_i \in \mathbb{R}^{b_i \times b_i}$ are small orthogonal matrices

Problem: restrictive structure

- HOFT method parametrizes \mathbf{Q} as a product of several orthogonal sparse matrices, aiming to form a dense orthogonal matrix

$$\mathbf{Q} = \mathbf{B}_k \mathbf{B}_{k-1} \dots \mathbf{B}_1,$$

Each matrix \mathbf{B}_i is a block diagonal matrix up to a permutation of rows and columns, consisting of r orthogonal block matrices of sizes $\mathbf{b} \times \mathbf{b}$

Problem: requires computing product of multiple matrices

\mathcal{GS} matrices

Definition 1

A matrix \mathbf{A} is in $\mathcal{GS}(L_1, P, P_k)$ class with k_1, k_2 blocks and block sizes $\mathbf{b}_1, \mathbf{b}_2$ if

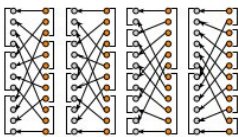
$$\mathbf{A} = \mathbf{P}_1 (\mathbf{LPR}) \mathbf{P}_k,$$

where $\mathbf{L} = \text{diag}(L_1, L_2, \dots, L_{k_1}), L_i \in \mathbb{R}^{b_1 \times b_1},$
 $\mathbf{R} = \text{diag}(R_1, R_2, \dots, R_{k_2}), R_i \in \mathbb{R}^{b_2 \times b_2}, \mathbf{P}_1, \mathbf{P}, \mathbf{P}_k$ are permutation matrices and $\mathbf{b}_1, \mathbf{b}_2 = k_1 \cdot k_2$

This definition can be extended into the case of arbitrary number of block diagonal matrices with rectangular blocks

We typically set $\mathbf{P} = \mathbf{P}_{k_1, k_2}$ constructed by $\text{tr}(\mathbf{I}) = (\mathbf{I} \text{ mod } k) \cdot \frac{k}{2} + \lfloor \frac{k}{2} \rfloor$ with $k = k_1$. This is most effective at forming dense matrices which is shown by Proposition 1

Fig. 1. Illustration of \mathbf{P}_{k_1, k_2} permutations for $k \in \{4, 3, 2, 6\}$



Properties of \mathcal{GS} matrices

We find that class $\mathcal{GS}(L, P, I)$ consists of block matrices with low rank blocks, ranks of which are defined by permutation matrix \mathbf{P}

Proposition 1

Let \mathbf{A} be a matrix from $\mathcal{GS}(L, P, I)$ with a permutation matrix \mathbf{P} defined by the function $\sigma: \{0, \dots, n-1\} \rightarrow \{0, \dots, n-1\}$. Let $\{w_i^r\}$ be the rows of the blocks $\mathbf{R}_1, \dots, \mathbf{R}_{k_2}$ and $\{u_i^c\}$ the columns of the blocks $\mathbf{L}_1, \dots, \mathbf{L}_{k_1}$ in the consecutive order. Then the matrix \mathbf{A} can be written as a block matrix with $k_1 \times k_2$ blocks using the following formula for each block \mathbf{A}_{k_1, k_2}

$$\mathbf{A}_{k_1, k_2} = \sum_{\substack{r=1 \\ i=0}}^{k_2} u_{\sigma(i)} w_i^r.$$

Note that we are also indexing for this proposition for simplicity of formulas

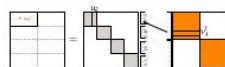


Fig. 2. Illustration of Proposition that provides block low rank interpretation of $\mathcal{GS}(L, P, I)$ matrices

In practice we achieve orthogonality through enforcing it for every block of \mathbf{L} and \mathbf{R} . This is theoretically justified, as shown by Theorem 1.

Proposition 2

Let \mathbf{A} be any orthogonal matrix from $\mathcal{GS}(P_1, P, P_k)$. Then \mathbf{A} admits $\mathbf{P}_1 (\mathbf{LPR}) \mathbf{P}_k$ representation with the matrices \mathbf{L}, \mathbf{R} consisting of orthogonal blocks.

Applications

- CSOFT: We utilize the pipeline of OFT and HOFT methods with the exception of parameterizing \mathbf{Q} with orthogonal permuted $\mathcal{GS}(P_1, P, P_k)$ matrices

$$\mathbf{Q} = \mathbf{P}^T \mathbf{LPR},$$

where $\mathbf{L} = \text{diag}(L_1, \dots, L_r), L_i \in \mathbb{R}^{b \times b},$
 $\mathbf{R} = \text{diag}(R_1, \dots, R_r), R_i \in \mathbb{R}^{b \times b}$. We use $\mathbf{P}_{\text{CSOFT}}$ as a permutation matrix \mathbf{P} .

- Doable CSOFT: We introduce an approach that multiplies pre-trained weight matrices from both sides with $\mathbf{Q}_0, \mathbf{Q}_1$ parametrized as orthogonal \mathcal{GS} matrices

$$\mathbf{y} = (\mathbf{W}^0)^T \mathbf{x} \rightarrow \mathbf{y} = (\mathbf{Q}_0 \mathbf{W}^0 \mathbf{Q}_1)^T \mathbf{x}$$

Compared to one-sided adaptation, this modification allows to adapt both left and right singular vectors of \mathbf{W}^0 , improving model's flexibility.

- \mathcal{GS} Orthogonal Convolutions: By analogy, we construct \mathcal{GS} convolution

$$\mathbf{Y} = \text{ClShuffle}(\text{GrConv}(\text{ClShuffle}(\text{GrConv}(\mathbf{X})),$$

where GrConv - grouped convolution, ClShuffle (Channel Shuffle) - channel shuffle operation, i.e. permuting order of channels. This can be used to construct orthogonal

convolutions [- orthogonal Jacobian matrices], using Taylor expansion of a matrix exponential for convolutions with slow symmetric filters

$$\mathbf{Y} = \text{ClShuffle}(\exp(\frac{\mathbf{C}}{\text{groupsize}} * (\text{ClShuffle}(\mathbf{L}_1^{\text{groupsize}} * \mathbf{X}))))$$

$$\mathbf{L}_r * \mathbf{X} = \mathbf{X} + \frac{\mathbf{L}_r * \mathbf{X}}{1!} + \dots + \frac{\mathbf{L}_r^n * \mathbf{X}}{n!} + \dots$$

Natural Language Understanding

Table 1. Results on GLUE benchmark with RoBERTa base model. Adapters were applied to all linear layers. We report Pearson correlation for SST-2. Matthews's correlation for CoLA and accuracy for other tasks

Model	MRPC	MRPC	STS-B	STS-B	QQP	QQP	MRPC	STS-B	QQP
CSOFT	0.84	0.84	0.85	0.85	0.86	0.86	0.84	0.85	0.86
CSOFT	0.84	0.84	0.85	0.85	0.86	0.86	0.84	0.85	0.86
CSOFT	0.84	0.84	0.85	0.85	0.86	0.86	0.84	0.85	0.86
CSOFT	0.84	0.84	0.85	0.85	0.86	0.86	0.84	0.85	0.86

Subject-driven generation

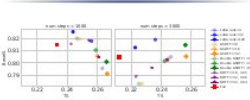


Fig. 3. Image and text similarity visualisation for different methods on subject-driven generation



Fig. 4. Visual results on subject-driven generation

\mathcal{GS} Orthogonal Convolution

Table 2. Results of training LipConvNet 15 architecture on CIFAR-100. (\mathbf{a}, \mathbf{b}) in "Groups" column denotes that we have to grouped exponential convolutions: the first one with kernel size = 3 the second with kernel size = 1. If $\mathbf{b} = \mathbf{0}$, we have only one \mathcal{GS} orthogonal convolutional layer. Before each grouped layer with k_2 groups we use a ClShuffle operator.

Conv. Layer	# Params	Groups	Similarity	Accuracy	Relative Accuracy
8x4	28.17k	(3, 1)	0.8195	29.18%	-
GS-OC	8.91k	(3, 1)	0.8186	29.20%	29.20%
GS-OC	8.91k	(3, 1)	0.8125	29.66%	29.66%
GS-OC	1.86k	(3, 2)	0.8267	28.96%	28.96%
GS-OC	1.70k	(3, 3)	0.8276	29.7%	29.7%

Knowledge graph completion with mixed geometry tensor factorization

Юсупов В. А. [ORCID] стажёр-исследователь научно-учебной лаборатории матричных и тензорных методов в машинном обучении



Knowledge Graph Completion with Mixed Geometry Tensor Factorization

Yusupov Viacheslav Maxim Rakhuba Evgeny Frolov
PCS HSE, Moscow



Introduction

Our contributions are:

- We introduce a new mixed-geometry tensor factorization (MIG-TF) model that combines Tucker decomposition defined in the Euclidean space with a new hyperbolic ternary interaction term.
- We highlight intricacies of applying geometric approach to real-world knowledge graphs and demonstrate the associated with it limitations of using single-geometry modeling.
- The proposed combined approach significantly reduces the number of model parameters compared to state-of-the-art methods. It does so without sacrificing expressive power and achieves more accurate results in most of the common benchmarks.

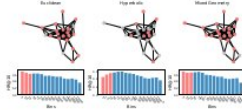


Fig. 1. Comparison of performances of the Euclidean, hyperbolic and mixed-geometry models on knowledge graph FB15k-237.

Motivation

We observe that real-world knowledge graphs are not consistently align with the assumption of strict hierarchical internal structure and may only partially follow a power-law distribution. We introduce our model extracting information from both hierarchical and non-hierarchical structured data.

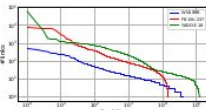


Fig. 2. Link distribution on three benchmark knowledge graphs considered in this work. Our MIG-TF approach outperforms both Euclidean and hyperbolic models, see Table 1.

Hyperbolic geometry

The Lorentz inner product for vectors $x, y \in \mathbb{R}^{n+1}$ which is used for measuring distance between hyperbolic embeddings:

$$\langle x, y \rangle_{\mathcal{L}} = -x_0 y_0 + \sum_{i=1}^n x_i y_i,$$

where

$$x_0 = \sqrt{\beta + \sum_{i=1}^n x_i^2}, \|x\|_{\mathcal{L}}^2 = \langle x, x \rangle_{\mathcal{L}}$$

The corresponding n -dimensional space $\mathcal{H}^{n,\beta} \subset \mathbb{R}^{n+1}$ is called hyperbolic and defined as follows:

$$\mathcal{H}^{n,\beta} = \{x \in \mathbb{R}^{n+1} \mid \|x\|_{\mathcal{L}}^2 = -\beta, \beta \geq 0\}.$$

The origin of the hyperbolic $\mathcal{H}^{n,\beta}$ is $(\beta, 0, \dots, 0) \in \mathbb{R}^{n+1}$.

Geodesic and Lorentz Distance

The associated geodesic distance is defined as

$$d_1(x, y) = \operatorname{arccosh}(-\langle x, y \rangle_{\mathcal{L}}).$$

We introduce the square Lorentz distance between $x, y \in \mathcal{H}^n$

$$d_2^2(x, y) = \|x - y\|_{\mathcal{L}}^2 = -2 - 2\langle x, y \rangle_{\mathcal{L}}. \quad (1)$$

This distance violates the triangle inequality:

$$d(x, z) \leq d(x, y) + d(y, z). \quad (2)$$

Tetrahedron Pooling Tensor Factorization

We propose to modify (2) to capture ternary interactions. In particular, we utilize the so-called tetrahedron inequality for the points u, v, t, ω in the Euclidean space, it holds

$$d(u, v) + d(u, t) \leq d(u, t) + d(v, t) + d(\omega, u) + d(\omega, v). \quad (3)$$

Consequently, we can naturally introduce the following score function:

$$S_0 = d_1(u, v) + d_1(t, \omega) - d_1(u, t) - d_1(v, t) - d_1(\omega, u) - d_1(\omega, v), \quad (4)$$

We replace (4) with the following 'smoothed' and differentiable everywhere score function that mimics the desired behaviour:

$$S_{\text{st}}(u, v, t) = \frac{1}{2} \left(\frac{d_2^2(u, v) + d_2^2(t, \omega) - d_2^2(u, t)}{\sqrt{d_2^2(t, \omega) + d_2^2(\omega, u) - d_2^2(\omega, v)}} + \frac{d_2^2(t, \omega) + d_2^2(\omega, v) - d_2^2(t, v)}{\sqrt{d_2^2(\omega, u) + d_2^2(\omega, v) - d_2^2(u, v)}} \right). \quad (5)$$

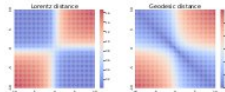


Fig. 3. Left score function landscape corresponds to our score function (5), whilst right one corresponds to (4).

In TPTF model, we use the binary cross-entropy (BCE) loss function:

$$\operatorname{lact}(y, p) = -y \log(p) - (1 - y) \log(1 - p).$$

We optimize u, v and t embeddings in the Euclidean space with the AdamV optimizer and map them to the hyperbolic as follows: $v \in \mathbb{R}^n$ maps to $\frac{1}{\|v\|_{\mathcal{L}}} \beta + \|v\|_{\mathcal{L}}^{-1} v \in \mathbb{R}^{n+1}$.

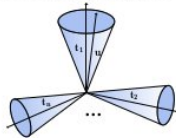


Fig. 4. Each embedding of a relation $\{t_1, \dots, t_n\}$ defines a cone that encompasses the embeddings of the entities associated with that relation.

Mixed Geometry Tensor Factorisation

To improve the link prediction in knowledge graphs we introduce a shared-factor mixed geometry model combining Euclidean TuckerET and hyperbolic TPTF models. The score function of the Euclidean model is defined as:

$$S_E(\mathbb{G}, u, v, t) = \sum_{\alpha, \beta, \gamma=1}^{d_1, d_2, d_3} G_{\alpha\beta\gamma} u_{\alpha} v_{\beta} t_{\gamma}. \quad (6)$$

The score function of our mixed-geometry model MIG-TF is the sum of score functions of the Lorentzian and Euclidean models:

$$S(\text{MIG-TF}) = (S_E) + (S_{\text{st}}), \quad (7)$$

In MIG-TF model, we utilize pretrained Tucker model (6) and optimize the hyperbolic term parameters (5) of the score function (7) to minimize the BCE loss:

$$\Delta_{\text{MIG-TF}} = \frac{1}{N} \sum_{i=1}^N \operatorname{lact}(u_i, \sigma(S(\text{MIG-TF}))),$$

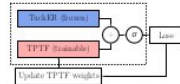


Fig. 5. The proposed MIG-TF model architecture.

Results

Model	FB15k-237	WN18RR	YAGO3-10
TuckerET _{0.5}	4 · 10 ⁶	8 · 10 ⁶	25 · 10 ⁶
Complex-N3	4 · 10 ⁶	8 · 10 ⁶	25 · 10 ⁶
RotatE	60 · 10 ⁶	60 · 10 ⁶	320 · 10 ⁶
RotS	40 · 10 ⁶	80 · 10 ⁶	320 · 10 ⁶
Our models			
TF _{0.5} - $\mathcal{H}_{0.5}$	2 · 10 ⁶	4 · 10 ⁶	12 · 10 ⁶
MIG-TF _{0.5}	5 · 10 ⁶	10 · 10 ⁶	31 · 10 ⁶

Fig. 6. Approximate number of model's parameters on knowledge graphs WN18RR, FB15k-237 and YAGO3-10

	FB15k-237	WN18RR	YAGO3-10
--	-----------	--------	----------

Model	HR@1	MRR	HR@1	MRR	HR@1	MRR
TuckerET	0.266	0.178	0.141	0.170	0.486	0.541
RotatE	0.217	0.226	0.148	0.179	-	-
Complex-N3	0.264	0.157	0.145	0.180	0.488	0.569
RotatE	0.245	0.138	0.128	0.176	0.482	0.485
RotS	0.246	0.141	0.148	0.196	0.485	0.573
RotR	0.252	0.146	0.149	0.161	0.201	0.378
Our models						
TF	0.186	0.238	0.252	0.114	0.383	0.481
MIG-TF	0.277	0.162	0.149	0.196	0.501	0.579
MIG-TF _{0.5}	0.276	0.165	0.152	0.192	0.202	0.260

Fig. 7. Metrics on knowledge graphs HR@k and MRR metrics of models on knowledge graphs WN18RR, FB15k-237 and YAGO3-10. Underlined means the best metric on the knowledge graph.

Организаторы

Организационный комитет:

Аржанцев И. В. [📧], декан ФКН, д.ф.-м.н.

Мицюк А. А. [📧], зам. декана ФКН по научной работе, к.комп.н.

Игнатов А. Д. [📧], н.с. МЛ ТИ, преподаватель ДПИ ФКН

Кузнецова К. А. [📧], менеджер НУЛ ФКН

Также принимали участие в организации конференции:

Нургаязова Д. М. [📧], **Сафонов Н. А.** [📧], **Приблудова П. А.** [📧] – информационное сопровождение, дизайн, оформление

Рябинин Н. Д. [📧] – техническое сопровождение, фото

Талайкова Н. А. [📧], **Абдуллаева С. И.** [📧], **Козлова М. С.** [📧] – полезные советы и помощь в организации

II Научная конференция ФКН

