

1. Команда проекта (ФИО, должность, телефон, почта)

Руководитель проекта: главный научный сотрудник кафедры дискретной математики ФПМИ МФТИ, руководитель научного коллектива “Квантовые вычисления” ФПМИ МФТИ, д.ф.-м.н. **Холодов Ярослав Александрович** (kholodov@crec.mipt.ru, kholodov.iaa@mipt.ru, +7 903 758-91-34).

Исполнитель проекта: старший научный сотрудник кафедры дискретной математики ФПМИ МФТИ, к.ф.-м.н. **Лобачев Виктор Анатольевич** (vlob1969@yahoo.com, +7 915 377 1992).

Исполнитель проекта: младший научный сотрудник кафедры дискретной математики ФПМИ МФТИ, аспирант 3 года обучения университета Иннополис **Али Джнади**.

Исполнитель проекта: младший научный сотрудник кафедры дискретной математики ФПМИ МФТИ, аспирант 1 года обучения ФПМИ МФТИ **Яцулевич Владимир Владимирович**.

Исполнитель проекта: инженер кафедры дискретной математики ФПМИ МФТИ, магистр 1 года обучения университета Иннополис **Ахмад Сархан**.

Исполнитель проекта: техник лаборатории кафедры дискретной математики ФПМИ МФТИ, бакалавр 4 года обучения университета Иннополис **Хади Саллум**.

Исполнитель проекта: техник кафедры дискретной математики ФПМИ МФТИ, бакалавр 3 года обучения университета Иннополис **Торшин Егор Константинович**.

2. Описание проекта

Название проекта: разработка методов применения алгоритмов нелинейной дискретной оптимизации для сжатия нейронных сетей.

Что? Для кого? Результат выполнения проекта? Проект реализуется группой исследователей кафедры дискретной математики ФПМИ МФТИ, которая имеет большой опыт проведения как теоретических исследований, так и развития их прикладных реализаций вместе с индустриальными партнерами. Итогом реализации проекта станут публикации в научных журналах первого квартала и выступления на конференциях A^* , а также популяризация этого направления исследований искусственного интеллекта среди студентов МФТИ: в том числе через научные семинары, научный трек инновационного практикума и научную летнюю школу «ЛИПС-26».

Подробное описание:

Глубокие нейронные сети (Deep Neural Networks, DNN) демонстрируют отличные результаты во множестве задач — от распознавания изображений до обработки естественного языка. Однако подобные успехи зачастую сопровождаются высокой вычислительной сложностью и значительными требованиями к ресурсам. Это значительно усложняет развертывание крупных моделей на ресурсно-ограниченных устройствах, таких как мобильные телефоны или встраиваемые системы.

Сжатие нейронных сетей, и в особенности прореживание ее фильтров, направлено на устранение избыточных параметров обученной сети при минимально возможной потере точности. Задача прореживания по своей сути представляет собой сложную комбинаторную оптимизацию: имея N фильтров, требуется найти оптимальное подмножество из K фильтров, минимизирующее ошибку модели после прореживания. Область исследования методов и алгоритмов прореживания активно развивается. Существующие обзоры систематизируют таксономию методов, подходы к оцениванию и открытые проблемы структурного прореживания.

Современные подходы к данной задаче в основном делятся на две категории. Наиболее распространенный подход использует итеративные эвристики, основанные на метриках важности, таких как L_1 -норма или градиентные метрики. Метод на основе разложения в ряд Тейлора стал популярным базовым методом, позволяющим оценивать важность фильтра по влиянию его на функцию потерь. Несмотря на эффективность, такие методы остаются «жадными»: они оценивают важность каждого фильтра по отдельности и не учитывают корреляцию между ними.

Второй, менее распространенный подход заключается в формулировке задачи прореживания как формальной задачи комбинаторной оптимизации. Например, прореживание можно выразить в виде задачи квадратичной неограниченной бинарной оптимизации (Quadratic Unconstrained Binary Optimization – QUBO), подходящей для специализированных алгоритмов, таких как квантовый отжиг или методы имитации отжига. Однако существующие алгоритмы отжига, как правило, опираются на простые метрики, такие как L_1 -норма, для оценки важности фильтров, что приводит к ухудшению качества по сравнению с более сильными градиентными методами, такими как прореживание на основе Тейлора.

Команда проекта является одной из немногих, кто активно исследуют данную тематику в России. В составе кафедры дискретной математики ФПМИ начал свою работу научный коллектив «Квантовых вычислений» (заведующий – руководитель проекта Холодов Я.А.). В рамках работы команда выполнила ряд важных шагов в развитии продвижении алгоритмов дискретной оптимизации.

Публикации по теме:

Кроме этого, участники проекта имеют заметное число публикаций в ведущих Q1 журналах и в препринтах A* конференций по данной тематике:

Q1 журнал: Salloum H., Zhanalin S., Badr A.A. & Kholodov Y.A. Mini-scale traffic flow optimization: an iterative QUBOs approach converting from hybrid solver to pure quantum processing unit. *Sci Rep* **15**, 22904 (2025). <https://doi.org/10.1038/s41598-025-04568-2>.

A* конференция по ИИ: Hadi Salloum, Kamil Sabbagh, Ruslan Lukin, Yaroslav Kholodov, Gleb Ryzhakov. Performance evaluation of the tensor train sampler in QUBO-based machine learning ADMET classification //FPI-ICLR 2025.

A* конференция по ИИ: Hadi Salloum, Kamil Sabbagh, Osama Orabi, Amine Trabelsi, Ruslan Lukin, Yaroslav Kholodov. Tensor-Train Unsupervised Image Segmentation //FPI-ICLR 2025.

A* конференция по ИИ: Kamil Sabbagh, Hadi Salloum, Yaroslav Kholodov. RepFair-QGAN: Alleviating Representation Bias in Quantum Generative Adversarial Networks Using Gradient Clipping //FPI-ICLR 2025.

Команда проекта является либо выпускниками, либо действующими аспирантами и студентами МФТИ и университета Иннополис.

Планы на 2026 год:

За 2026 год предполагается получить решение, которое устраняет разрыв между подходами, использующими разложения Тейлора и QUBO. Мы выдвигаем гипотезу, что низкое качество предыдущих QUBO-постановок связано не с алгоритмом оптимизации, а с выбором целевой функции. Поэтому мы предлагаем новую гибридную постановку задачи, эффективно использующую оценку важности фильтра нейросети на основе включения разложения в ряд Тейлора в линейный член алгоритма QUBO. Кроме того, мы включаем метрику схожести активаций в квадратичный член QUBO, что позволяет учитывать функциональную избыточность фильтров в процессе оптимизации.

Также, мы будем вводить двухэтапный процесс верификации. Сначала мы решаем гибридную QUBO-задачу, получая высококачественную начальную маску прореживания. Затем это решение используется в качестве начального для стадии «TT Refinement» — оптимизации на основе безградиентного тензорного обучения, работающего как «черный ящик» и выполняющего тонкий локальный поиск непосредственно по заданной метрике качества (например, PSNR. Такая двухэтапная схема устраняет расхождение между целевой функцией QUBO оптимизации и фактической целевой функцией модели в исходной задаче.

3. Актуальность

Внедрение больших языковых моделей на потребительских устройствах и устройствах Интернета вещей приобретает все большее значение в различных отраслях, что обусловлено необходимостью обработки данных в режиме реального времени, повышения конфиденциальности и снижения зависимости от больших корпораций, которые являются основными поставщиками ИИ ассистентов и прочих услуг в сфере искусственного интеллекта. Этот переход особенно важен для малых и средних предприятий, которые, несмотря на отсутствие развитой инфраструктуры крупных корпораций, нуждаются в возможностях, которые даёт использование искусственного интеллекта, чтобы оставаться конкурентоспособными.

Современные большие языковые модели обычно содержат от десятков до сотен миллиардов параметров, что значительно повышает вычислительные затраты как на хранение таких моделей в памяти GPU, так и на скорость инференса. Это делает их использование невозможным на устройствах с серьёзными ограничениями на вычислительные ресурсы. Модели таких размеров не получится даже загрузить в память устройства, не говоря уже об их запуске. Поэтому методы сжатия нейросетей, такие как прореживание, дистилляция знания или квантизация, критически важны для использования в быту.

Также со сжатием за счет упрощения структуры нейросети растёт её обобщающая способность. Это позволяет нейросетям лучше работать с данными, непохожими на данные из обучающего датасета. Хорошее сжатие делает нейросеть одновременно более ресурсоёмкой и устойчивой к работе с разнородными данными.

Прореживание нейронных сетей является самым распространённым подходом к сжатию, поэтому актуальность его исследований не оставляет сомнений.

4. Кто или что является благополучателем результата проекта?

В качестве благополучателя может выступать ФПМИ МФТИ и МФТИ в целом (подробнее в пункте 5).

Также в качестве благополучателя могут выступать индустриальные партнеры МФТИ. Технология важна для многих телекоммуникационных и IT-компаний, таких как Huawei, с которой у коллектива уже есть опыт реализации проектов.

5. Планируемый результат проекта

Непосредственный результат проекта:

Формальным результатом работы в 2026 году должна стать минимум две статьи (в Q1 журнал и препринтах A*/A конференций – см. пример в пункте 2) в области дискретной оптимизации, а также несколько выступлений на ведущих мировых конференциях по ИИ.

Как результат повлияет на развитие МФТИ:

Дискретная оптимизация на основе квантовых и квантово-вдохновленных алгоритмов – перспективная тема исследований, важный шаг в развитии искусственного интеллекта, как в России, так и в мире. МФТИ, как один из ведущих исследовательских центров, всегда был на передовых ролях в перспективных исследованиях. Поэтому кажется естественным выпускникам Физтеха не просто развивать алгоритмы дискретной оптимизации, но и делать это в рамках МФТИ с прицелом на поддержание ведущих позиций Физтеха в научном сообществе в области ИИ.

Помимо проведения научных и индустриальных исследований, коллектив активно вовлекает в свою работу мотивированных студентов МФТИ. Мы стараемся поддерживать любые начинания и попытки начать свои первые исследования студентов и молодых ученых. Все больше сильных студентов 3-4 курсов уходят в индустрию, даже не попробовав себя в науке, хотя ровно для раннего старта в науку и задумывался Физтех. Даже небольшая финансовая поддержка с нашей стороны может помочь удерживать ребят вокруг интересных научных задач и в некотором смысле популяризировать науку среди студентов МФТИ.

6. План реализации проекта, его этапы и их сроки.

На данный момент работы ведутся по всем направлениям. Примерный сроки представления результатов:

- февраль 2026 года - май 2026 года: участие коллектива в организации научно-образовательных проектов в рамках научного трека инновационного практикума ФПМИ (обязательная для всех программ ПМИ дисциплина в 6 семестре).
- май 2026 года: окончание работ над 3-4 научными статьями в области дискретной оптимизации, отправка работ на рецензирование в ведущие журналы и мировые конференции;
- август 2026 года: окончание разработки технологии применения алгоритмов нелинейной дискретной оптимизации для сжатия нейронных сетей;
- июль - декабрь 2026 года: выступление на ведущих мировых конференциях по ИИ с результатами проекта;
- октябрь – декабрь 2026 года: новостное освещение технологии применения алгоритмов нелинейной дискретной оптимизации для сжатия нейронных сетей;
- сентябрь 2026 года – февраль 2027 года: выход научных публикаций с результатами проекта.

Также на протяжении всего 2026 года будет проводиться еженедельный открытый научный семинар, куда смогут присоединяться студенты МФТИ.

7. Бюджет проекта

Общий бюджет проекта – 18 000 000 рублей.

Сумма, запрашиваемая от ФЦК – 5 000 000 рублей.

Команда проекта реализует свои исследования на внебюджетные источники финансирования ФПМИ. Продвижение, административно-управленческое сопровождение проекта осуществляется офисом сопровождения научно-образовательных проектов ФПМИ.

Необходимость привлечения дополнительных источников финансирования обусловлена тем, что команда начала свою деятельность только в октябре 2025 года. В этой связи требуется поддержка на начальном этапе исследований. В дальнейшем планируется также финансировать деятельность команды внебюджетными источниками ФПМИ, привлекать новые средства от партнеров, заинтересованных в поддержке развития фундаментальной науки с индустриальными приложениями, а также подавать заявки на конкурсы Российского научного фонда, искать дополнительные возможности получения государственного финансирования.

Статья расходов	Сумма
<i>Заработные платы (с учетом налогов и резервов)</i>	<i>5 000 000 (пять миллионов) рублей</i>

8. Долгосрочное развитие проекта

Проект планируется как долгосрочный. Как уже было описано выше, применения алгоритмов нелинейной дискретной оптимизации для сжатия нейронных сетей является важной фундаментальной ступенью в развитии искусственного интеллекта нового поколения. Важно приумножать успехи в этой области не только в мировом сообществе, но и в России. Поэтому работа в этой области будет продолжена как по научной составляющей, так и по индустриальной. Предполагается искать спонсирование, как со стороны государства (РНФ и другие фонды), так и со стороны индустриальных заказчиков (Huawei, Сбер, Т-Банк, Яндекс, Газпромнефть). В идеале есть надежда, что бизнес поддержит не только исследования в данной области, но и не побоится массово использовать технологию для реальных задач и приложений.

9. Подразделение МФТИ, через которое будет проходить финансирование проекта.

Проект будет реализовываться на базе кафедры дискретной математики ФПМИ в рамках деятельности научного коллектива “Квантовые вычисления”, руководитель проекта Холодов Ярослав Александрович (yar.kholodov@gmail.com). Реализация проекта курируется непосредственно Райгородским Андреем Михайловичем, заведующим кафедры дискретной математики, директором ФПМИ.

Ответственным исполнителем будет являться Байнова Виктория Константиновна, руководитель офиса сопровождения научно-образовательных проектов ФПМИ, (bainova.vk@mipt.ru, +7 996 188 40 93), которая также будет заниматься всем необходимым документооборотом.

10. Как ваш проект будет способствовать популяризации деятельности ФЦК МФТИ среди студентов, сотрудников и выпускников, какие конкретные действия вы планируете для этого предпринять.

- 1) Планируется указывать ФЦК, как источник финансирования и поддержки в новостях, связанных с темой проекта.
- 2) Планируется указать ФЦК, как спонсора в презентациях и постерных выступлениях на научных конференциях и семинарах.
- 3) Команда также планирует продвижение бренда ФЦК среди студентов. Как было сказано в пунктах 2 и 5, коллектив видит одним из своих приоритетов вовлечение студентов МФТИ в свои научные исследования. Вовлечение происходит через организацию открытого оффлайн научного семинара в МФТИ, участие в научных школах и проектной деятельности в рамках «инновационного практикума» на ФПМИ. Во всех анонсах мероприятий коллектива планируется указание ФЦК как основного источника поддержки и финансирования. В том числе коллектив будет поощрять инициативных студентов также развивать свои научно-образовательные проекты под брендом ФЦК.