Робастное распределенное обучение, состязательные атаки и многоагентные системы

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

Руководитель проекта: старший научный сотрудник лаборатории дискретной и комбинаторной оптимизации МФТИ к.ф-м.н. Рогозин Александр Викторович (aleksandr.rogozin@phystech.edu, +7 985 166-65-44).

Исполнитель проекта: доцент Лаборатории инновационных технологий обработки видеоконтента университета Иннополис к.ф.-м.н. Бадер Рашид (b.rasheed@innopolis.university, +7 968 998-67-33).

Исполнитель проекта: научный сотрудник лаборатории математических методов оптимизации МФТИ, к.ф.-м.н. Курузов Илья Алексеевич (kuruzov.ia@phystech.edu, +7 926 452 63-77).

Исполнитель проекта: научный сотрудник лаборатории математических методов оптимизации МФТИ, к.ф.-м.н. Ярмошик Демьян Валерьевич (yarmoshik.dv@phystech.edu, +7 977 812-13-37).

Исполнитель проекта: старший научный сотрудник МФТИ, к.ф-м.н. Матюхин Владислав Вячеславович (vladmatyukh@gmail.com +7 925 991-59-49).

2. Описание проекта

Краткое описание для сайта (2-3 предложения по структуре: Что? Для кого? Каким образом? Что благополучатели получат?)

Большие модели машинного обучения могут быть неустойчивы к враждебному изменению входных данных, а также к неправильному поведению узлов в сети, если речь идет о мультиагентной системе. Например, изменение всего нескольких десятков пикселей в изображении может привести к тому, что оно будет неверно классифицировано нейросетью, а присутствие враждебно настроенных агентов может привести к сбоям в работе всей сети. Чтобы избежать таких неожиданных эффектов в работе модели, необходимо использовать специальные методы обучения.

Подробное описание проекта (опишите, что конкретно вы собираетесь делать).

Планируется исследование устойчивости нейросетей к разным видам нарушений. С одной стороны, на вход модели могут подаваться враждебно измененные данные. Даже незначительные изменения (например, искажение нескольких пикселей изображения) могут привести к ошибкам в ответе модели. Такие искажения входных данных называются состязательными атаками (adversarial attacks [1,2]). С другой стороны, в случае мультиагентного обучения неустойчивость может происходить со стороны неисправных агентов, необязательно враждебно настроенных (Byzantine attacks). И в случае подделки

входных данных, и в случае нарушения работы отдельных агентов модель может работать не так, как ожидается.

Нейросеть все-таки можно защитить от состязательных атак, но для этого нужно обучать ее определенным образом. Обучение должно производиться с учетом возможных искажений входа. Имеются и другие методы, например, определенные способы регуляризации модели (например, sharpness aware minimization) или ограничение ее весов во время обучения. Конечная цель всех подходов - сделать модель менее чувствительной к изменению входа, чтобы при незначительных его изменениях не было бы резкого изменения ответов модели.

В условиях огромного количества параметров моделей и больших объемов данных обучение происходит на распределенных системах. Модель и/или данные обучающей выборки распределяются между узлами вычислительной сети. В процессе самого распределенного обучения некоторые узлы могут выпадать из общего процесса в случае их неисправности или внешнего взлома. Чтобы сделать процесс обучения устойчивым, необходимо устанавливать таких враждебных агентов и исключать их из общего процесса, т.е. игнорировать сообщения, которые они посылают. Кроме того, в ходе распределенного обучения возможен учет пересечения моделей, лежащих на разных узлах, по данным или по параметрам.

[1] Madry A. et al. Towards deep learning models resistant to adversarial attacks //arXiv preprint arXiv:1706.06083. – 2017.

[2] Carlini N. et al. On evaluating adversarial robustness //arXiv preprint arXiv:1902.06705. – 2019.

Участники проекта имеют публикации по темам состязательных атак и мультиагентного обучения.

[ICLR 2025, A* конференция] Yarmoshik D. et al. Decentralized Optimization with Coupled Constraints //The Thirteenth International Conference on Learning Representations.

[NeurIPS 2024, A* конференция] Kuruzov I., Scutari G., Gasnikov A. Achieving linear convergence with parameter-free algorithms in decentralized optimization //Advances in Neural Information Processing Systems. – 2024. – Т. 37. – С. 96011-96044.

[IEEE Access 2024] Rasheed, Bader, et al. "Exploring the impact of conceptual bottlenecks on adversarial robustness of deep neural networks." IEEE Access (2024).

[Applied Sciences 2023] Rasheed, Bader, Adil Khan, and Asad Masood Khattak. "Structure estimation of adversarial distributions for enhancing model robustness: A clustering-based approach." Applied Sciences 13.19 (2023)

[International Journal of Computational Intelligence Systems 2023] Rasheed, Bader, et al. "Boosting adversarial training using robust selective data augmentation." International Journal of Computational Intelligence Systems 16.1 (2023): 89.

3. Актуальность

Зачем нужен этот проект, какую проблему он решает?

Проект направлен на увеличение устойчивости больших моделей машинного обучения по отношению к изменениям входных данных, а также по отношению к сбоям в многоагентной сети. Обучение современных больших языковых моделей проходит именно на распределенных архитектурах, поэтому использование алгоритмов, робастных к помехам в сети (например, отключению ее участников или ненадежному их поведению), является актуальной темой для машинного обучения в целом.

4. Кто или что является благополучателем результата проекта?

Благополучателями проекта является Институт искусственного интеллекта МФТИ и университет Иннополис. В команду входят сотрудники этих двух институтов, имеющие опыт работы в разных областях - робастности нейросетей, выпуклой оптимизации и мультиагентных системах. В ходе работы над проектом ожидается обмен опытом и возможное установление научных связей между двумя университетами.

5. Планируемый результат проекта

Непосредственный результат

Планируется подготовить публикацию на тематическую конференцию или в журнал по машинному обучению. Также планируется выступление на российских конференциях по оптимизации и машинному обучению.

Как результат проекта влияет на развитие МФТИ и/или комьюнити вокруг него.

В первую очередь, полезным будет налаживание новых научных связей и обмен опытом со специалистами из других областей. В команду проекта входят сотрудники университета Иннополис. Участники проекта со стороны МФТИ имеют опыт в выпуклой оптимизации и децентрализованных системах, со стороны Иннополиса - в робастном машинном обучении и состязательных атаках. Задачи робастного обучения порождают много интересных постановок с точки зрения теоретической выпуклой оптимизации. Для лаборатории математических методов оптимизации МФТИ (и в целом для ФПМИ и Института искусственного интеллекта МФТИ) применение методов к реальным задачам машинного обучения позволит расширить область научных интересов. Сами

темы, заявленные в проекте связаны с робастным машинным обучением и устойчивым распределенным обучением - и то, и другое имеет прикладное значение в современном машинном обучении.

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

В мае-июне 2026г. планируется подать статью по итогам работы на зарубежные и российские тематические конференции. Еще одна волна конференций будет в сентябре 2026г. К декабрю 2026г. - марту 2027г. планируется публикация поданных работ в тезисах соответствующих конференций.

7. Бюджет проекта *a. Общий бюджет проекта b. Сумма, запрашиваемая от ФЦК МФТИ; c. Есть ли софинансирование? В каком объеме и его источник? Укажите потенциальных партнеров, в том числе тех, которые поддерживают не только деньгами. d. Почему проект не может быть полностью профинансирован из бюджета МФТИ или иных источников? e. Прикрепите отдельным от презентации xls-файлом таблицу с постатейной расшифровкой всего бюджета, в которой укажите, на какие статьи расхода вы запрашиваете денег у ФЦК, а на какие планируете получить спонсорское финансирование;*

Общий бюджет проекта – 3 000 000 рублей.

Сумма, запрашиваемая от ФЦК – 3 000 000 рублей.

Статья расходов	**Сумма**
Заработные платы (с учетом налогов и резервов)	*3 000 000 (три миллиона) рублей*

8. Долгосрочное развитие проекта

Является ли проект разовой акцией или планируется его повторение на регулярной основе? Планируете ли вы в будущем искать другие источники финансирования?

Тема устойчивого мультиагентного обучения больших моделей может быть полезна, в том числе, индустриальным заказчикам, обучающим собственные модели для компаний. В случае, если в рамках проекта будут получены яркие результаты, это может быть интересно заказчикам, выход на которых есть, например, через Проектный офис ФПМИ.

9. Подразделение МФТИ, через которое будет проходить финансирование проекта. *Укажите, какие договорённости есть с руководителем подразделения. В отдельных случаях мы можем запросить одобрение курирующего проректора. Укажите, кто будет заниматься документооборотом для оплаты счетов через МФТИ. a. ФИО и контакты руководителя проекта/подразделения, который будет подписывать ФЛС. b. ФИО и контакты ответственного исполнителя по проекту.*

Проект будет реализовываться в Лаборатории математических методов оптимизации (заведующий Лабораторией - Гасников А.В., gasnikov@yandex.ru).

Руководитель проекта Рогозин А.В. (aleksandr.rogozin@phystech.edu, +7 985 166-65-44).
Ответственным исполнителем будет являться Матюхин Владислав, старший научный сотрудник лаборатории, vladmatyukh@gmail.com +7 925 991-59-49.

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

Планируется указывать ФЦК, как источник финансирования в публикуемых в рамках проекта статьях. Кроме того, в новостях по проекту будет отмечаться, что в рамках его выполнения при поддержке ФЦК между МФТИ и университетом Иннополис налаживались научные связи, происходил обмен опытом.