Приложение 4
к протоколу заседания
ученого совета НИУ ВШЭ
от 24.01.2020 № 1

Концепция
научно-учебной лаборатории компании Яндекс департамента больших данных и информационного поиска факультета компьютерных наук Национального исследовательского университета «Высшая школа экономики»

Национальный исследовательский университет «Высшая школа экономики» (далее соответственно – ФКН, НИУ ВШЭ) в рамках сотрудничества с ООО «Яндекс» создает научно-учебную лабораторию компании Яндекс (далее – Лаборатория) под руководством кандидата физико-математических наук Артёма Валерьевича Бабенко. Тематика Лаборатории будет включать в себя компьютерное зрение, анализ текстов на естественном языке, краудсорсинг, а также информационный поиск и рекомендации. Направление исследований Лаборатории является крайне актуальным и востребованным. Поиск и извлечение информации, в том числе из больших коллекций плохо структурированных данных, является важной составляющей любой интеллектуальной системы. Поиск, в том числе по изображениям, и рекомендательные системы являются краеугольным камнем бизнеса крупнейших IT-компаний, таких как Amazon, Google и Яндекс.

Актуальность создания Лаборатории

В 2010-2016 гг. методы глубинного обучения позволили добиться значительных прорывов в этих областях. Во многих задачах из области анализа изображений удалось превзойти возможности человека, а в ряде задач анализа текстов приблизиться к этому уровню. Тем не менее, для создания эффективных и надёжных технологий ИИ на основе этих методов необходимы значительные усилия по их осмыслению и совершенствованию. Так, в том, что касается более сложно устроенных, чем изображения и тексты, неоднородных данных, нейросетевые методы пока проигрывают классическим, а задача подбора глубинных архитектур сейчас решается в основном эвристически. Более того, обучение глубинных нейронных сетей невозможно в отсутствии больших датасетов, и задача их сбора является одним из частых камней преткновения на пути к решению любой достаточно сложной задачи. Фактически единственным на сегодняшний день способом получения обучающих датасетов является разметка людьми (асессорами). Развитие технологий краудсорсинга позволит значительно упростить эту часть процесса. Однако в этой области есть ещё много вызовов, связанных с агрегацией и валидацией ответов асессоров. Научные интересы Лаборатории покрывают практически все аспекты промышленной работы с данными, включая их сбор, подбор и построение моделей и их последующий анализ. Следует отметить, что в российских компаниях практически отсутствуют исследовательские группы, совмещающие высокий исследовательский потенциал с глубокой погружённостью в проблемы, приходящие из продакшн-подразделений.

Необходимость создания Лаборатории при наличии базовой кафедры компании Яндекс (далее – кафедра) объясняется тем, что кафедра поддерживает образовательный процесс, а Лаборатория создаётся для проведения научных исследований и привлечения студентов и аспирантов к научной деятельности. Так, в структуре

кафедры не может быть стажеров-исследователей. Поэтому для организации работы со студентами и встраивания их в научный коллектив необходимо создание Лаборатории.

Организация-партнер, совместно с которой создается Лаборатория

Исследовательский отдел ООО «Яндекс» является сейчас одним из лидеров в России в области прикладного машинного обучения и информационного поиска. Сотрудники этого отдела проводят исследования на мировом уровне, регулярно публикуясь на ведущих международных конференциях по машинному обучению и смежным дисциплинам, имеющих ранг A* по рейтингу CORE (NeurIPS – Neural Information Processing Systems, ICML - International Conference on Machine Learning, ACL – Annual Meeting of the Association for Computational Linguistics, ICCV – International Conference on Computer Vision и др.). В частности, руководителем группы А.В. Бабенко, работающим на ФКН, за 2017-2019 гг было опубликовано 5 работ на конференциях уровня A*. В настоящее время исследовательский отдел ООО «Яндекс» представляет собой коллектив молодых сотрудников, успешно ведущих исследования в области машинного обучения, опирающихся на последние разработки в этой области.

Исследовательская группа ООО «Яндекс» поддерживает контакты с рядом ведущих зарубежных научных групп по машинному обучению (Кембриджский университет, университеты Эдинбурга и Амстердама и др.). Сотрудники исследовательского отдела ООО «Яндекс» активно преподают в Школе анализа данных ООО «Яндекс», а также на ФКН и на Физтех-школе прикладной математики и информатики, участвуют в различных митапах, конференциях и мини-курсах. В частности, Еленой Войта был разработан уникальный для России курс по Natural Language Processing, который читается в Школе анализа данных ООО «Яндекс».

А.В. Бабенко участвовал в создании ФКН, и с тех пор он и активно преподаёт, и руководит курсовыми и дипломными работами.

Задачи создаваемой Лаборатории:

1) Проведение научных исследований по следующим тематикам:

- компьютерное зрение;
- анализ текстов на естественном языке и машинный перевод;
- краудсорсинг;
- машинное обучение;
- речевые технологии;
- информационный поиск и рекомендации

2) Публикация статей на международных научных конференциях, имеющих ранг A*, по тематикам из п.1;

3) Привлечение студентов и аспирантов ВШЭ и других образовательных организаций высшего образования для работы в качестве стажёров-исследователей в лаборатории;

4) Работа со студентами ВШЭ в рамках руководства КР, ВКР и проектной работы, преподавания и проведения мероприятий.

Основные направления деятельности, описание будущих исследований

В рамках Лаборатории планируется проведение научно-исследовательской деятельности по следующим направлениям:

- компьютерное зрение;
- анализ текстов на естественном языке, в том числе машинный перевод и речевые технологии;
- краудсорсинг;

- информационный поиск и рекомендации.

Среди задач, которые планируется решать в первую очередь, можно упомянуть следующие:

1) Улучшение алгоритмов градиентного бустинга.

Градиентный бустинг – это один из самых широко применяемых классов ансамблевых алгоритмов машинного обучения, незаменимый в работе со сложно устроенными данными, особенно если они содержат пропуски, категориальные переменные и пр. Алгоритмы, основанные на градиентном бустинге, реализованы в таких популярных библиотеках, как XGBoost и Catboost (библиотека с открытым исходным кодом, разрабатываемая ООО «Яндекс») и используются во многих продакшен-системах. Улучшение этих алгоритмов позволит решать эти задачи эффективнее и принесёт несомненную пользу как для теории машинного обучения, так и для бизнеса. В рамках исследования планируется, в частности, изучить возможности улучшения градиентного бустинга с помощью дропаута деревьев и обучения градиентного бустинга с монотонными ограничениями для части признаков. Полученные улучшения будут имплементированы в библиотеке Catboost.

2) Получение теоретических гарантий для поиска ближайших соседей с помощью графовых подходов.

Имеется серьёзная эмпирическая база, подтверждающая эффективность графовых подходов к поиску ближайших соседей. Тем не менее, предпринято очень мало попыток получения теоретических гарантий эффективности. Планируется провести теоретический анализ графовых подходов в данной задаче.

3) Повышение разнообразия предсказаний ансамбля моделей.

Ансамбли моделей довольно часто позволяют добиться повышения производительности ML-систем; кроме того, эмпирически показано, что они позволяют получать робастные меры неопределённости. Важным требованием является то, что ансамбль должен демонстрировать консистентное поведение на привычных данных, но выдавать достаточно разнообразные предсказания на данных из других распределений. К сожалению, ансамбли не всегда демонстрируют достаточно разнообразное поведение. Таким образом, повышения разнообразия является важной и актуальной проблемой, решение которой позволит получить более надёжные и робастные меры неопределённости предсказаний.

4) Ускорение и сжатие генеративных нейронных сетей.

Современные нейросетевые архитектуры позволяют решать большое количество прикладных задач компьютерного зрения с высокой точностью. Как правило, наилучшее качество достигается при использовании больших моделей, занимающих много места и использующих много видеопамяти в процессе предсказания. В связи с этим возникла потребность сжатия и ускорения таких ресурсоёмких архитектур, для использования их на мобильных устройствах. У задачи сжатия дискриминативных моделей, занимающихся предсказанием некоторых величин (метка класса, глубина, сегментация) на основе исходного изображения есть множество изученных подходов, однако сжатие генеративных моделей, генерирующих новые данные на основе некоторых величин, исследовано гораздо хуже. В частности, планируется изучить возможности применения следующих техник: прунинг (обнуление некоторых весов/нейронов на основе обученной модели), дистилляция (использование выходов и/или промежуточных слоёв большой модели для предоставления дополнительных меток для маленькой модели), квантизация (замена float вычислений внутри сети на вычисления с uint/bool параметрами) и

факторизация/декомпозиция (уменьшение эффективного размера матрицы весов/преобразований с помощью матричных/тензорных разложений).

5) Исследование устойчивости нейронных сетей.

С чувствительностью к изменениям в данных связаны две глобальные проблемы, стоящие на текущий момент перед глубинным обучением. С одной стороны, существующие архитектуры уязвимы к преднамеренным атакам на них (adversarial attacks); кроме того, имеет место проблема дестабилизации обучения при обучении моделей класса GAN (Generative Adversarial Networks). Предлагается проводить анализ нейронных сетей с помощью изучения их локальных линеаризаций. Будет исследована связь поведения сети в окрестностях точек из обучающей выборки и глобального поведения модели при преднамеренных атаках, либо в контексте обучения генеративных состязательных сетей.

Распространение результатов научной деятельности

Результаты научной деятельности Лаборатории будут распространяться путём публикации на международных конференциях ранга A*, а также в рамках однодневных открытых мероприятий на ФКН, которые планируется проводить каждый семестр.

Формат работы со студентами

Предполагается работа со студентами и аспирантами НИУ ВШЭ в рамках исследовательских стажировок, а также в рамках КР, ВКР или проектной деятельности под руководством работников Лаборатории.

Предполагаемые результаты

В случае создания такой Лаборатории ФКН получит (1) команду молодых активно практикующих исследователей, которые будут не только заниматься исследованиями и руководить курсовыми и дипломными работами студентов и диссертациями аспирантов, но и проводить семинары и мастер-классы, рассчитанные на всех студентов и работников НИУ ВШЭ; (2) перспективы для углубления научных связей с ООО «Яндекс».

В свою очередь, ООО «Яндекс» получит площадку для работы со студентами в рамках исследовательских задач и возможность выращивать своих будущих исследователей из студентов ФКН.

Открытие такой Лаборатории позволит увеличить число совместных (НИУ ВШЭ и ООО «Яндекс») статей в высокоуровневых изданиях, в первую очередь на международных научных конференциях, имеющих ранг A*; студенты и аспиранты ФКН и молодые работники Лаборатории получат необходимый опыт проведения исследований на высочайшем научном уровне, работая в тесном контакте с исследователями, уже имеющими успешный опыт таких публикаций. Это приведет к росту как числа высокорейтинговых публикаций, так и их цитируемости, т.к. работы по исследованию методов машинного обучения, опубликованные на ведущих конференциях, обычно хорошо цитируются. Наконец, под эгидой Лаборатории предполагается провести серию совместных (НИУ ВШЭ и ООО «Яндекс») научных и просветительских мероприятий.

Кадровый состав Лаборатории

Работниками Лаборатории будут:

к.ф.-м.н. Бабенко Артем Валерьевич (заведующий лабораторией);

Ph.D. Малинин Андрей Алексеевич;

к.ф.-м.н. Прохоренкова Людмила Александровна.

Следующие студенты войдут в начальный состав стажёров-исследователей лаборатории:

Башаев Никита Константинович, студент 3 курса бакалаврской программы «Прикладная математика и информатика» ФКН;

Емельяненко Дмитрий Викторович, студент 3 курса бакалаврской программы «Программная инженерия» ФКН;

Рябинин Максим Константинович, студент 1 курса магистерской программы «Науки о данных» ФКН;

Червонцев Сергей Сергеевич, студент 1 курса магистерской программы «Науки о данных» ФКН.

После официального открытия Лаборатории будет запущен дальнейший поиск стажёров-исследователей.

Имеющийся задел по тематикам Лаборатории

Планируемые работники Лаборатории активно выступают на наиболее престижных конференциях в области Data Science. Так, в 2018 и 2019 годах были представлены следующие результаты:

1. Beyond Vector Spaces: Compact Data Representation as Differentiable Weighted Graphs. D. Mazur, V. Egiazarian, S. Morozov, A. Babenko. 2019, NeurIPS
2. Reverse KL-Divergence Training of Prior Networks: Improved Uncertainty and Adversarial Robustness. A. Malinin, M. Gales. 2019, NeurIPS
3. Unsupervised Neural Quantization for Compressed-Domain Similarity Search. S. Morozov, A. Babenko. 2019, ICCV
4. Learning to Route in Similarity Graphs. D. Baranchuk, D. Persiyanov, A. Sinitsin, A. Babenko. 2019, ICML
5. Community detection through likelihood optimization: in search of a sound model. L. Prokhorenkova, A. Tikhonov. 2019, WWW
6. CatBoost: unbiased boosting with categorical features. L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, A. Gulin. 2018, NeurIPS
7. Non-metric Similarity Graphs for Maximum Inner Product Search. S. Morozov, A. Babenko. 2018, NeurIPS
8. Predictive Uncertainty Estimation via Prior Networks. A. Malinin, M. Gales. 2018, NeurIPS
9. Revisiting the Inverted Indices for Billion-Scale Approximate Nearest Neighbors. D. Baranchuk, A. Babenko, Y. Malkov. 2018, ECCV

Лаборатория берёт на себя следующие обязательства:

Показатель	**2020 год**	**2021 год**	**2022 год**
Публикации на конференциях уровня A* по рейтингу CORE	3	3	4
Число стажеров-исследователей Лаборатории-студентов бакалавриата (не менее)	2	2	2

Число стажеров-исследователей Лаборатории-студентов магистратуры (не менее)	4	5	6
Число стажеров-исследователей Лаборатории-аспирантов (не менее)	0	1	2
Процент стажеров-исследователей Лаборатории-студентов и аспирантов НИУ ВШЭ среди всех стажеров-исследователей (не менее)	50%	70%	70%
Руководство КР, ВКР и проектами	4	8	10
Преподавание: число курсов, НИС и групп, в преподавание которых будут вовлечены работники Лаборатории	2	3	4
Защиты аспирантов	0	0	0
Проведение однодневных мероприятий (семинаров, мастер-классов) на ФКН: число проведённых мероприятий	2	2	2