ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 13–18

РЕЗУЛЬТАТЫ ДЕЯТЕЛЬНОСТИ ИССЛЕДОВАТЕЛЬСКИХ ЦЕНТРОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

УДК 004.8

ДОВЕРЕННЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: ВЫЗОВЫ И ПЕРСПЕКТИВНЫЕ РЕШЕНИЯ

© 2022 г. Д. Ю. Турдаков[1,4,6,*], академик РАН А. И. Аветисян[1,4,5,6], К. В. Архипенко[1], А. В. Анциферова[1], Д. С. Ватолин[4], С. С. Волков[1], А. В. Гасников[1,5], Д. А. Девяткин[4,7], М. Д. Дробышевский[1,5], А. П. Коваленко[1], М. И. Кривоносов[1], Н. В. Лукашевич[4], В. А. Малых[5], С. И. Николенко[6], И. В. Оселедец[2,3], А. И. Перминов[1,4], И. В. Соченков[2,7], М. М. Тихомиров[4], А. Н. Федотов[4], М. Ю. Хачай[8]

Поступило 28.10.2022 г.
После доработки 29.10.2022 г.
Принято к публикации 01.11.2022 г.

Широкое внедрение технологий искусственного интеллекта привело к возникновению новых угроз, эффективное противодействие которым не может быть реализовано текущими средствами разработки безопасного ПО. Для ответа на этот вызов в 2021 г. в рамках федерального проекта “Искусственный интеллект” на базе ИСП РАН был создан Исследовательский центр доверенного искусственного интеллекта, задачами которого является создание научно-технологической базы для обеспечения доверия к технологиям ИИ. В статье рассмотрены риски применения технологий искусственного интеллекта, а также представлены направления и промежуточные результаты работ Центра доверенного искусственного интеллекта ИСП РАН.

Ключевые слова: доверенный искусственный интеллект, атаки на машинное обучение, объяснимый искусственный интеллект, доверенные интеллектуальные системы

DOI: 10.31857/S2686954322070207

1. ВВЕДЕНИЕ

Современные интеллекутальные системы используют целый стек технологий, который состоит из методов и алгоритмов искусственного интеллекта, фреймворков машинного обучения (например, TensorFlow, PyTorch), а также инфраструктурных решений для их поддержки (облачные системы, специализированные аппаратные системы и др.). Обеспечение доверия к таким системам является долгосрочным вызовом, активным поиском ответа на который мировое сообщество занимается уже несколько лет. Однако на текущий момент не существует научно-технологической базы для разработки высоконадежных доверенных и одновременно эффективных систем, использующих технологии искусственного интеллекта (интеллектуальных систем), в том числе отсутствуют инструменты для поиска новых видов уязвимостей и противодействия новым типам угроз, специфичным для этих технологий.

Работа над ответом на этот вызов ведется в двух встречных направлениях. Первое направление – выработка требований к интеллектуальным системам и разработка государственных стандартов [1, 2].

Второе направление – создание научно-технологической базы, поддерживающей разрабатываемые стандарты. Без создания инструментальных средств, обеспечивающих безопасность функционирования интеллектуальных систем, невозможно говорить о полноценной реализации раз-

[1] *Институт системного программирования им. В.П. Иванникова Российской академии наук, Москва, Россия*
[2] *Сколковский институт науки и технологий, Москва, Россия*
[3] *Институт искусственного интеллекта AIRI, Москва, Россия*
[4] *Московский государственный университет имени М.В. Ломоносова, Москва, Россия*
[5] *Московский физико-технический институт, Долгопрудный, Россия*
[6] *Национальный исследовательский университет “Высшая школа экономики”, Москва, Россия*
[7] *Институт системного анализа Федерального исследовательского центра “Информатика и управление” Российской академии наук, Москва, Россия*
[8] *Институт математики и механики Уральского отделения Российской академии наук, Екатеринбург, Россия*
*****E-mail: turdakov@ispras.ru*

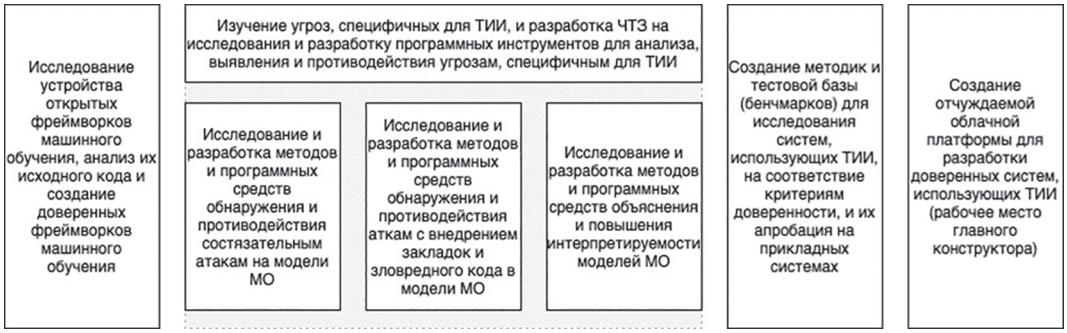

Рис. 1. Направления работ Центра.

работанных и перспективных стандартов. Работы в этом направлении ведутся исследовательским центром доверенного искусственного интеллекта ИСП РАН.

Программа Центра определяет следующие ключевые направления работ (рис. 1)

- Исследование устройства открытых фреймворков машинного обучения, анализ их исходного кода и создание доверенных фреймворков машинного обучения;
- Исследование и разработка программных инструментов для анализа, обнаружения и противодействия угрозам, специфичным для технологий искусственного интеллекта, включающее три раздела
- Исследование и разработка методов и программных средств обнаружения и противодействия состязательным атакам на модели машинного обучения;
- Исследование и разработка методов и программных средств обнаружения и противодействия атакам с внедрением закладок и зловредного кода в модели машинного обучения;
- Исследование и разработка методов и программных средств объяснения и повышения интерпретируемости моделей машинного обучения;
- Создание методик и тестовой базы (“бенчмарков”) для исследования интеллектуальных систем на соответствие критериям доверенности, и их апробация на прикладных системах;
- Создание отчуждаемой облачной платформы для разработки доверенных интеллектуальных систем.

В следующих разделах будут рассмотрены результаты текущих исследований и разработки по каждому направлению.

2. ДОВЕРЕННЫЕ ФРЕЙМВОРКИ МАШИННОГО ОБУЧЕНИЯ

Для продуктивной разработки интеллектуальных систем используется большое количество библиотек и фреймворков машинного обучения. Эти программные инструменты существенно ускоряют создание прикладных продуктов. Но как любое программное обеспечение они могут содержать ошибки и недокументированные возможности на уровне своего исходного кода или своих программных зависимостей. Такие уязвимости могут быть использованы злоумышленником для проведения атак на целевую систему. Таким образом, обеспечение доверия к интеллектуальным системам в целом невозможно без обеспечения доверия к этим ключевым системным компонентам.

Создание доверенного промышленного программного обеспечения регламентируется циклом безопасной разработки (SDL). Важным аспектом применения практик SDL является анализ не только разработанных компонент, но и заимствованного открытого программного обеспечения. Основными методами анализа, которые применяются в SDL, являются статический и динамический анализ (фаззинг). С помощью этих технологий возможно выявить критические дефекты в ПО и затем устранить их до выхода новой версии продукта.

В рамках работы Центра поставлена задача создания доверенных версий популярных фреймворков Tensorflow и PyTorch. Для этих фреймворков был проведен анализ поверхности атаки. Перспективным вектором атаки был выбран компонент загрузки моделей, так как обученные модели часто могут браться из недоверенного источника. Кроме того для фреймворка TensorFlow были восстановлены фаззинг-цели в проекте OSS-Fuzz, обеспечивающем непрерывный фаззинг для программного обеспечения с открытым исходным кодом [3].

Фазинг фреймворков производился с помощью инструмента Sydr [4]. В результате фаззинга было обнаружено 7 ошибок для фреймворка PyTorch [5, 6] и одна ошибка для фреймворка TensorFlow [7]. Также исходный код фреймворков был проанализирован инструментом статическо-

го анализа Svace [8]. В результате в фреймворке PyTorch было обнаружено 13 ошибок [9]. В фреймворке TensorFlow было обнаружено 8 ошибок [10].

Информация о найденных ошибках и предложения по исправлению некоторых из них были доведены до сообщества разработчиков. В целях дальнейшего использования все исправления были собраны в доверенных версиях фреймворков, созданных на базе PyTorch v1.11.0 и TensorFlow v2.8.2.

При этом сами фреймворки постоянно развиваются и появляются новые версии. Поэтому необходимы постоянная проверка новых изменений кода и синхронизация доверенных версий с оригинальными репозиториями. Для обеспечения этого непрерывного процесса создана аппаратно-программная инфраструктура обеспечения доверия к базовым фреймворкам машинного обучения, объединяющая инструменты SDL и инструменты для автоматизации их применения.

3. УГРОЗЫ, СПЕЦИФИЧНЫЕ ДЛЯ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ, И МЕТОДЫ ПРОТИВОДЕЙСТВИЯ

Одним из центральных вопросов разработки систем доверенного интеллекта является проблема неустойчивости отображений, выучиваемых моделями машинного обучения, к изменениям входных данных. Даже незначительное изменение входного объекта, например, добавление невидимого глазу шума к картинке, может существенно изменить предсказание модели на новом объекте, который как внешне, так и по метрикам схожести практически не отличается от исходного объекта. Этот феномен привел к возникновению так называемых состязательных атак (adversarial attack). Актуальность методов противодействия состязательным атакам на модели машинного обучения подтверждена экспериментами, проведенными сотрудниками Центра на моделях и наборах данных его научных партнеров (Университет Иннополис, ННГУ им. Н.И. Лобачевского). Проведена успешная атака белого ящика на систему сегментации рентгеновских снимков и на модели для предсказания возраста человека по экспрессии генов, основанные на бустинге решающих деревьев.

Подходы к построению моделей, устойчивых к атакам, делятся на 2 класса. В первом подходе модифицируется (сглаживается) сама модель, что приводит к более высокой эмпирической устойчивости. Возможно получить теоретические гарантии на величину атаки, при которой модель не будет менять свои предсказания. В рамках работы центра доверенного ИИ исследователями из Сколтеха впервые предложен [11] универсальный вероятностный подход к созданию моделей с сертифицированной устойчивостью, основанный на границах Чернова–Крамера. Подход позволяет формально оценить вероятность отказа модели, если атака выбрана из определенного распределения. Теоретические выводы подтверждены экспериментальными результатами на различных наборах данных.

Во втором подходе изменяется способ обучения. Сотрудниками Центра проведены исследования по этому направлению, в части задач оптимизации в условиях (злонамеренных) помех. Проведено исследование black-box моделей оптимизации, в которых атаки моделируются небольшим шумом к выдаваемому значению целевой функции [12]. Предложена общая схема, позволяющая сводить выпуклую задачу безградиентной оптимизации к выпуклой гладкой задаче оптимизации со стохастическим градиентным оракулом. Впервые удалось показать, что оптимальный метод с точки зрения числа оракульных вызовов (числа вычислений целевой функции), оказывается оптимальным и с точки зрения числа последовательных итераций, но, самое главное, что чувствительность такого метода к неточности в значении функции у него доказуемо наилучшая. Таким образом, был предложен подход, который одновременно оптимален по всем трем критериям (число оракульных вызовов, максимальный параллелизм и максимальный допустимый уровень шума). Кроме того удалось решить проблемы, связанные с формализацией атак на параметры задачи оптимизации таким образом, что исходная оптимизационная задача, в конечном итоге, заменяется седловой задачей. Предложены подходы к решению таких оптимизационных задач с доказанной оптимальностью [13, 14].

Кроме того, исследователями Центра протестирован прикладной подход противодействия состязательным атакам – состязательное обучение (adversarial training) моделей, т.е. обучение на атакованных некоторым методом атаки данных, а также методы аугментации данных, направленные на защиту от атак [15]. А также разработаны новые нейросетевые архитектуры, более устойчивые к состязательным атакам с общепринятыми видами возмущений [16]

4. ПОВЫШЕНИЕ ИНТЕРПРЕТИРУЕМОСТИ МОДЕЛЕЙ

Нейронные сети широко используются в качестве мощных инструментов моделирования, и большинство крупных поставщиков включили их в свое программное обеспечение для интеллектуального анализа данных. Моделирование, однако, является лишь частью процесса интеллектуального анализа данных. Дополнительно необходимо проанализировать влияние входных

переменных на результат модели. Результаты некачественной интерпретации применяемой модели могут быть использованы злоумышленником для проведения атак на целевую систему, например, через “отравление” данных. Таким образом, обеспечение доверия к интеллектуальным системам требует обеспечения интерпретируемости применяемых моделей.

В рамках работы Центра поставлены задачи повышения интерпретируемости и выявления некорректных результатов работы многослойных нейронных сетей. Разработан тестовый стенд для визуализации полносвязной нейронной сети с сублинейными функциями активации для задачи бинарной классификации в двумерном пространстве признаков. С его помощью проанализирован способ геометрической интерпретации обученной нейросетевой модели, где каждому нейрону соответствует разделяющая гиперплоскость в исходном признаковом пространстве. На основе нее предложен ряд новых методов анализа модели. Для входного примера в качестве интерпретации решения сети предъявляется k ближайших соседей с точки зрения модели из обучающей выборки. Предложен метод построения решающего дерева, имитирующего работу исходной нейросети, листовые вершины которого соответствуют упомянутым секторам или их объединениям. Статистический анализ в них может показать, например, стоит ли доверять решению сети в этой области, или что невозможно однозначно принять решение. Предложенные подходы могут быть применимы для полносвязных нейросетей с произвольным числом нейронов и слоев. Для сверточных нейросетей предложен метод восстановления интерполированных изображений из точек признакового пространства. Также разрабатываются методы анализа фильтров, значимых для определения заданного класса изображений.

Созданы [17] вычислительно-эффективные методы выявления некорректных результатов работы многослойных нейронных сетей с архитектурой “Трансформер”. Первый метод заключается в оценке принадлежности анализируемых объектов к определенным классам с помощью ансамбля нейронных сетей, в каждой из которых маскируются отдельные веса выходного слоя. Второй метод – оценка расстояния между скрытыми векторными представлениями (эмбеддингами) анализируемых объектов и ближайших к ним объектов из обучающей выборки. Результаты экспериментов на задачах классификации текстов и извлечения именованных сущностей показали, что применение предложенных методов позволяет существенно повысить надежность обнаружения ошибок классификации по сравнению с более вычислительно затратными методами.

Кроме того, исследован подход получения семантически интерпретируемых категорий для векторных представлений на основе семантических сетей типа WordNet. В качестве интерпретируемых размерностей используются семантические классы (суперпонятия), объединяющие множества слов.

5. СОЗДАНИЕ МЕТОДИК И ТЕСТОВОЙ БАЗЫ (БЕНЧМАРКОВ) ДЛЯ ОЦЕНКИ ДОВЕРИЯ К ПРИКЛАДНЫМ СИСТЕМАМ

Наличие или отсутствие злоумышленника, сценарии воздействия злоумышленника на процессы жизненного цикла прикладных интеллектуальных систем, разнообразие самих прикладных задач и интеллектуальных систем требуют тщательного анализа и систематизации этих сценариев и угроз. В результате проведения такого анализа были разработаны критерии доверия к интеллектуальным системам, которые в дальнейшем будут апробированы на реальных интеллектуальных системах.

При этом доверие к прикладным интеллектуальным системам не может рассматриваться в отрыве от эффективности таких систем. Так, тривиальной является разработка модели машинного обучения, неуязвимой, например, к состязательным атакам, если не задавать минимальные требования к точности и скорости распознавания для этой модели. Таким образом, важной задачей является создание тестовой базы из наборов данных и моделей на основе реальных задач из различных прикладных областей, и разработка методик для оценки соответствия интеллектуальных систем и их компонентов требованиям в области доверия и эффективности.

Были разработаны бенчмарки, решающие различные прикладные задачи, используя данные разного типа: текст, изображения, видео, графы, таблицы. При этом в качестве бенчмарков мы используем как простые (baseline) модели, так и создаем state-of-the-art решения, на которых можно демонстрировать эффективность методов обеспечения доверия в реальных условиях.

В частности, разработана новая процедура создания искусственного движения для обучения более устойчивых нейросетевых алгоритмов обработки видео. Предложенная процедура была использована для обучения нейросетевого алгоритма семантического матирования видео с людьми [18]. Акцент в данном алгоритме был сделан на стабильности результата работы во времени. При помощи предложенной процедуры удалось многократно повысить размер обучающей выборки, а также повысить устойчивость метода матирования к разным видам движения.

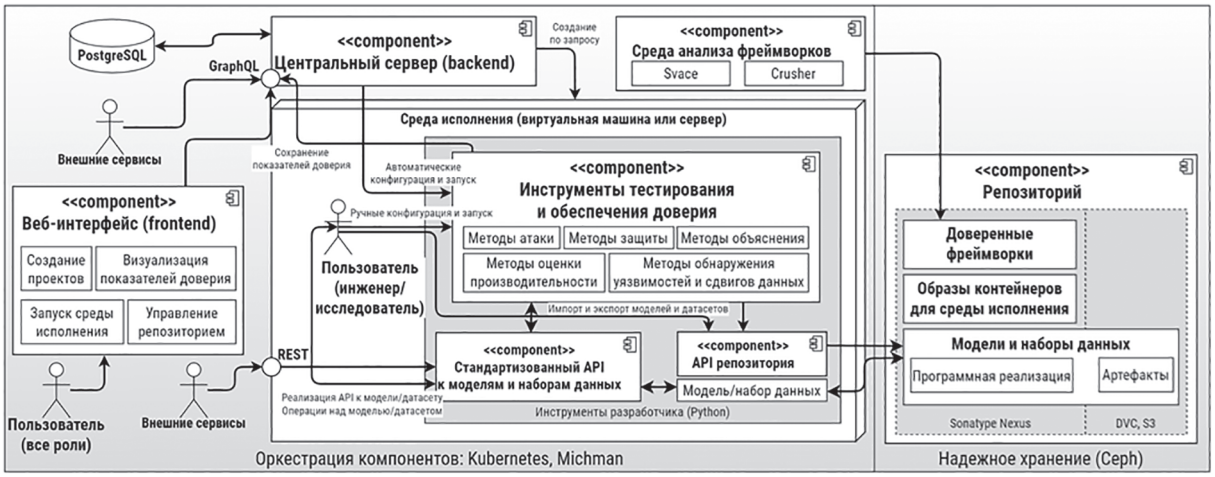

Рис. 2. Высокоуровневая архитектура Платформы.

Также разработан новый однопроходный метод для открытого извлечения информации из текстов, вдохновленный алгоритмами детектирования объектов из области компьютерного зрения [19]. Подход использует независимую от порядка следования отдельных слов функцию потерь, основанную на двудольном сопоставлении слов и отдельных элементов триплетов, которое обеспечивает однозначные предсказания модели, и архитектуру, использующую только кодировщик на основе “Трансформер” для маркировки последовательностей. Предлагаемый подход демонстрирует превосходящую или аналогичную производительность с точки зрения как показателей качества, так и времени вывода по сравнению с современными моделями на стандартных бенчмарках.

6. ОБЛАЧНАЯ ПЛАТФОРМА ДЛЯ РАЗРАБОТКИ ДОВЕРЕННЫХ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ

Основываясь на результатах проведенных исследований была разработана концепция облачной платформы для разработки интеллектуальных систем. Платформа будет

• содержать инструменты анализа наборов данных (датасетов) и анализа моделей машинного обучения, в том числе объяснения моделей, исследования их устойчивости к атакам и производительности;

• накапливать доверенные модели, алгоритмы и наборы данных для обучения моделей и проведения экспериментов (бенчмарки);

• предоставлять компоненты жизненного цикла разработки безопасного программного обеспечения.

Современные методы машинного обучения требуют большого количества вычислительных ресурсов в момент обучения модели. В остальное время эти вычислительные ресурсы могут использоваться неэффективно. Модель облачных вычислений позволяет решить эту проблему за счет обеспечения доступа к вычислительным ресурсам по требованию. Отсюда вытекает требование возможности развертывания Платформы в публичных и приватных облаках. Работа Платформы в связке с облаком будет продемонстрирована на примере облачной платформы Asperitas и оркестратора Michman [20], обеспечивающего развертывание необходимых программных компонентов по запросу.

Еще одно ключевое требование к платформе – расширяемость. Так как новые программные инструменты анализа и реализации атак и защиты появляются постоянно, будет обеспечена возможность их добавления в Платформу. В частности, разрабатываются программные интерфейсы (рис. 2) и методика добавления таких инструментальных средств.

7. ЗАКЛЮЧЕНИЕ

Исследовательский центр доверенного искусственного интеллекта поставил перед собой задачу разработки научно-технической базы, которая позволит создавать доверенные интеллектуальные системы. С этой целью сотрудниками и партнерами Центра проводились работы по следующим направлениям: создание доверенных фреймворков машинного обучения; исследование угроз, специфичных для искусственного интеллекта и методов противодействия им; повышение интерпретируемости моделей машинного обучения; создание методик и бенчмарков для оценки доверия; а также объединение прикладных инстру-

ментов в облачную платформы для разработки доверенных интеллектуальных систем. Научные исследования опубликованы в 9 статьях, представленных на конференциях A*, и 7 журнальных статьях из первого квартиля (Q1). Основным прикладным результатом в 2022 г. являются созданные сотрудниками Центра доверенные версии фреймворков TensorFlow и PyTorch, уже переданные в опытную эксплуатацию индустриальным партнерам Центра.

СПИСОК ЛИТЕРАТУРЫ

1. ГОСТ Р 59276–2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения.
2. ГОСТ Р 59921–2021. Системы искусственного интеллекта в клинической медицине.
3. Pull request на восстановление фаззинг целей для фреймоврка TensorFlow. https://github.com/google/oss-fuzz/pull/7704. Дата обращения: 2022-10-26.
4. V*ishnyakov A., Fedotov A., Kuts D., Novikov A., Parygina D., Kobrin E., Logunova V., Belecky P., Kurmangaleev S.* Sydr: Cutting edge dynamic symbolic execution. In 2020 Ivannikov ISPRAS Open Conference (ISPRAS) (pp. 46–54). IEEE. 2020 December.
5. Pull request в PyTorch [https://github.com/pytorch/pytorch/pull/79192]. Дата обращения: 2022-10-26.
6. Pull request в PyTorch [https://github.com/pytorch/pytorch/pull/84343]. Дата обращения: 2022-10-26.
7. Pull request, Fix endless loop in TF. [https://github.com/tensorflow/tensorflow/pull/]. Дата обращения: 2022-10-26.
8. *Ivannikov V.P., Belevantsev A.A., Borodin A.E., Ignatiev V.N., Zhurikhin D.M., Avetisyan A.I.* Static analyzer Svace for finding defects in a source program code. Programming and Computer Software. 2014. V. 40 (5). P. 265–275.
9. [https://github.com/pytorch/pytorch/pull/85705]. Дата обращения: 2022-10-26.
10. [https://github.com/tensorflow/tensorflow/pull/57892]. Дата обращения: 2022-10-26.
11. *Pautov M., Tursynbek N., Munkhoeva M., Muravev N., Petiushko A., Oseledets I.* (2022, June). CC-Cert: A probabilistic approach to certify general robustness of neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 7, pp. 7975–7983).
12. *Gasnikov A., Novitskii A., Novitskii V., Abdukhakimov F., Kamzolov D., Beznosikov A., Takáč M., Dvurechensky P., Gu B.* The power of first-order smooth optimization for black-box non-smooth problems. ICML 2022.
13. *Kovalev D., Gasnikov A., Richtárik P.* Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. NeurIPS 2022.
14. *Kovalev D., Gasnikov A.* The First Optimal Algorithm for Smooth and Strongly-Convex-Strongly-Concave Minimax Optimization. NeurIPS 2022.
15. *Chistyakova A., Cherepnina M., Arkhipenko K., Kuznetsov S.D., Oh C.S., Park S.* September. Evaluation of interpretability methods for adversarial robustness on real-world datasets. In 2021 Ivannikov Memorial Workshop (IVMEM) (pp. 6-10). IEEE. 2021.
16. *Курденкова Е.О., Черепнина М.С., Чистякова А.С., Архипенко К.В.* Влияние трансформаций на успешность состязательных атак для классификаторов изображений Clipped BagNet и ResNet. Иванниковские чтения, 2022.
17. *Vazhentsev A., Kuzmin G., Shelmanov A., Tsvigun A., Tsymbalov E., Fedyanin K., Zhukov L.* Uncertainty Estimation of Transformer Predictions for Misclassification Detection //Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022. С. 8237–8252.
18. *Molodetskikh I., Erofeev M., Moskalenko A., Vatolin D.* Temporally coherent person matting trained on fake-motion dataset. Digital Signal Processing. 2022. V. 126. P. 103521.
19. *Vasilkovsky M., Alekseev A., Malykh V., Shenbin I., Tutubalina E., Salikhov D., Stepnov M., Chertok A., Nikolenko S.* DetIE: Multilingual Open Information Extraction Inspired by Object Detection. In Proceedings of the 36th AAAI Conference on Artificial Intelligence. 2022.
20. *Aksenova E., Lazarev N., Badalyan D., Borisenko O. and Pastukhov R.* December. Michman: an Orchestrator to deploy distributed services in cloud environments. In 2020 Ivannikov Ispras Open Conference (ISPRAS) (pp. 57–63). IEEE. 2020.