Рефераты

УДК 004.89

Джейлбрейк больших языковых моделей с помощью мягкого перефразирования вредоносных запросов. Алексеевская И., Архипенко К. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 7–24.

Методы выравнивания (alignment) больших языковых моделей (БЯМ) позволяют встраивать в них внутренние механизмы безопасности, обеспечивающие соблюдение этических принципов при генерации ответов. Тем не менее такие модели остаются уязвимыми к атакам-джейлбрейкам, которые необходимо учитывать в процедуре выравнивания. В работе предлагается простая, но эффективная атака-джейлбрейк, основанная на перефразировании вредоносных запросов ансамблем БЯМ. Мы разрабатываем специальную инструкцию для перефразирования, которая сохраняет исходный вредоносный смысл, но убирает явно выраженную негативную окраску и изменяет стиль текста, что позволяет осуществить джейлбрейк атакуемых моделей. Предложенный алгоритм SoftPrompt атакует модели Vicuna-13B, LLaMA-2-7B и GPT-3.5 Turbo на наборе данных JBB-Behaviors за один запрос, достигая успеха на уровне существующих методов атак. Кроме того, использование перефразирования в качестве функции возмущения делает атаку более скрытной с точки зрения перплексии входных данных атакуемой модели. Мы также проверяем эффективность атаки на большом числе БЯМ, в том числе проприетарных: ChatGPT-4o, Grok-2, LLaMA-3.1-405B и многих других. Наконец, мы публикуем набор данных для red-teaming, содержащий перефразированные версии вредоносных запросов JBB-Behaviors и HarmfulQ, который может способствовать дальнейшему выравниванию БЯМ.
Библ. — 24 назв.

УДК 004.89

Клиентообусловленное спекулятивное декодирование для больших языковых моделей в мультиарендных сервисах. Ангени Г. Э. , Ершов В. А. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 25–62.

Данная работа направлена на оптимизацию применения больших языковых моделей (LLM) в мультиарендных сервисах. Приводится подробный обзор современных методов спекулятивного декодирования, а для внедрения персонализации предсказаний без оказания влияния на их качество предлагается метод клиентообусловленного спекулятивного декодирования CCSD (Client-Conditioned Speculative Decoding). Метод интегрируется поверх существующих методов спекулятивного декодирования и использует n-граммные статистики по генерациям отдельных клиентов, обновляемые в онлайн-режиме и адаптирующие распределение кандидатов драфтовой модели при осуществлении спекулятивного сэмплирования (speculative sampling). Установлено, что на клиентских срезах с предсказуемыми ответами использование биграммных статистик влечёт ускорение до +36% токенов, генерируемых за секунду, при росте acceptance rate до +82%. Также возможность адаптации CCSD в процессе работы LLM к изменениям в распределении запросов от клиента позволяет обеспечивать сохранение более значительного прироста от использования спекулятивного декодирования по сравнению с переобучением драфтовой модели под клиента. Библ. — 9 назв.

УДК 004.89

Query2Diagram: ответы на запросы по коду с помощью UML-диаграмм. Барышников О., Алексеев А. М., Николенко С. И. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 63–86.

Документация по программному обеспечению часто устаревает или вовсе отсутствует. Вместе с тем, разработчикам для понимания сложных программных систем регулярно требуются структурированные представления их составляющих. Существуют инструменты обратного проектирования, способные порождать UML-диаграммы на основе кода. Однако, как правило, они выдают чрезмерно подробные схемы и никак не учитывают текущей информационной потребности разработчика, то есть “какой аспект кода следует проиллюстрировать”. Мы предлагаем метод *генерации UML-диаграмм на основе запросов*, при котором с помощью больших языковых моделей создаются диаграммы, напрямую отвечающие на вопросы о коде, сформулированные на естественном языке. В отличие от существующих методов, наш подход

позволяет создавать семантически ориентированные диаграммы, содержащие только релевантные элементы, дополненные контекстными описаниями. Мы донастраиваем (дообучаем) модель Qwen2.5-Coder-14B используя для этого тщательно отобранный набор данных, включающий файлы с кодом, запросы разработчиков и соответствующие ответы — представления в виде диаграмм в формате JSON. Мы оцениваем качество работы модели как с помощью автоматического выявления структурных дефектов, так и с помощью оценки семантической релевантности, вручную выполненной экспертами. Результаты показывают, что дообучение на небольшом количестве данных, исправленных вручную, даёт значительный эффект: наша лучшая модель демонстрирует самые высокие показатели F1-меры, при этом количество структурных дефектов в её ответах ниже, чем в ответах современных больших языковых моделей. Сгенерированные диаграммы структурно корректны и релевантны запросам разработчиков с точки зрения семантики. Таким образом, мы демонстрируем возможность использования больших языковых моделей для масштабируемой контекстной генерации документации по запросу. Программный код и набор данных доступны по ссылке `https://github.com/i-need-a-pencil/query2diagram`.
Библ. — 47 назв.

УДК 004.852

Эффективное планирование распределённых AutoML-конвейеров с учётом дедлайнов. Бутаков Н., Вахрушев А., Ходорченко М., Захарова А., Рыжков А., Плосская О., Гайнетдинов А., Дамдинов Р., Александров Д., Насонов Д., Савченко М. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 87–99.

Ежедневное обновление ML-решений в эксплуатации сегодня является распространённой практикой. Создание новых версий требует не только наличия готовых конвейеров автоматизированного машинного обучения (AutoML) для немедленной обработки обновлённых данных, но и своевременной доставки результата, зачастую в рамках жёстких сроков, составляющих всего несколько часов. Чтобы справляться с такими ограничениями, критически важно встраивать таймеры в AutoML-решения, поскольку это позволяет распределять время между несколькими ML-моделями и исследовать различные параметры

для обеспечения хорошего качества. Это порождает сложности при выборе того, как именно конвейеры должны выполнять вычисления для достижения оптимальной производительности. Масштабируемые распределённые фреймворки, такие как Apache Spark, при использовании в качестве основы для AutoML-фреймворков могут предоставлять возможности для совмещения режимов параллелизма по данным и параллелизма по вычислениям в гибридный режим, способный обеспечить лучшую общую производительность при ограничениях по срокам. Однако настройка параметров параллелизма для нескольких компонентов с различным поведением масштабируемости может оказаться непростой задачей, требующей учёта инфраструктуры, объёма набора данных, параметров ML-модели и закона её масштабируемости. В свете этих сложностей в данной работе предлагается новый метод оптимизации связанных с параллелизмом параметров AutoML-конвейера и распределения времени между несколькими ML-моделями в рамках ограничения по дедлайну.
Библ. — 18 назв.

УДК 519.7

Новые результаты о полудуплексной коммуникационной сложности. Дементьев Ю. И., Игнатьев А. А., Сидельник В. А., Смаль А. В., Ушаков М. С. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 100–133.

В данной работе продолжается линия исследований, начатая в работе K. Hoover, R. Impagliazzo, I. Mihajlin и A. Smal (2018), где была введена модель полудуплексной коммуникационной сложности. В отличие от классической модели коммуникационной сложности, предложенной в работе A. Yao (1979), полудуплексная модель ограничивает участников так, что в каждый момент времени каждый из них может либо передавать, либо принимать, но не одновременно, как при общении по рации. Эта модель мотивирована вопросами, возникающими при изучении гипотезы KRW. Решая открытые задачи, поставленные в работе Hoover и др., мы доказываем нижние оценки для функции дизъюнктности во всех рассматриваемых вариантах полудуплексной модели, а также доказываем верхнюю оценку в полудуплексной модели с нулём. Кроме того, мы устанавливаем нижние и верхние оценки полудуплексной сложности игр Карчмера–Вигдерсона для функций

подсчёта и для рекурсивной функции большинства, адаптируя и развивая методы классической коммуникационной сложности. Помимо этого, мы вводим понятие недетерминированной полудуплексной коммуникационной сложности и доказываем оценки, связывающие её с недетерминированной коммуникационной сложностью в классической модели. Предварительная версия данной работы была опубликована в трудах конференции SOFSEM 2021, а часть результатов впоследствии была уточнена в электронном препринте ECCC (2020). Настоящая журнальная версия существенно расширяет конференционную версию и включает новый раздел об играх Карчмера–Вигдерсона для пороговой функции и функции большинства.
Библ. — 19 назв.

УДК 004.85

Направленный шум и неявное смещение в методе Lion. Елистратов С., Подивилов А., Южаков Т., Ветров Д. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 134–152.

Lion – новый метод оптимизации, превосходящий традиционные оптимизаторы, такие как Adam, в широком спектре задач. Несмотря на эмпирический успех, причины превосходства Lion остаются неясными. В данной работе мы исследуем механизмы, обеспечивающие повышенную эффективность Lion, уделяя основное внимание структурированному шуму, возникающему из-за использования знаковой функции (sign) при обновлении градиента. Мы характеризуем этот шум углом поворота между вектором и его знаком. Мы вводим этот шум в виде случайного поворота на фиксированный угол в нормализованные обновления и анализируем, как эффективность такого метода соотносится с эффективностью Lion. Мы также приводим теоретический анализ, показывающий, как этот направленный шум изменяет неявное смещение оптимизатора, перенося индуцированный штраф за остроту/кривизну с направления градиента на ортогональную компоненту. Мы демонстрируем, что в нашей постановке этот метод обладает более высокой эффективностью, чем Lion. Такой подход выявляет связь между скоростью обучения и шумом, специфическую для метода Lion, и проливает свет на причины улучшения его показателей. Кроме того, мы выявляем эффект, который мы называем “отслеживанием импульса” (momentum tracing), в нейронных сетях со слоями нормализации

и активациями ReLU, способный существенно дестабилизировать процесс обучения. Наш анализ показывает, что присущий Lion шум поворота смягчает негативное влияние “отслеживания импульса”, обеспечивая более стабильное обучение. Эти результаты дают теоретическое обоснование эффективности Lion и намечают пути разработки более устойчивых алгоритмов оптимизации.
Библ. — 23 назв.

УДК 004.89

CleanComedy: создание дружелюбного юмора с помощью генеративных методов. Галимзянова Д., Горовая С., Вихорев Д., Ямщиков И. П., Жемчужина Е. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 153–177.

Генерация юмора является сложной задачей обработки естественного языка из-за ограниченности ресурсов и качества существующих наборов данных. Доступные языковые ресурсы юмора часто страдают от токсичности и дублирования, что ограничивает их пригодность для обучения устойчивых моделей. В данной работе предлагается CleanComedy – специализированный, частично размеченный и отфильтрованный по токсичности корпус английских и русских шуток, собранных из различных источников. Эффективность предложенного подхода к фильтрации данных исследуется с помощью опроса об уровне юмора и токсичности в различных группах шуток. Кроме того, изучается прогресс в области вычислительной генерации юмора путём сравнения шуток, написанных человеком, с шутками, сгенерированными большими языковыми моделями. Несколько моделей были дообучены на корпусах CleanComedy, с их помощью, а также с помощью исходных предобученных версий были сгенерированы шутки, после чего была собрана обратная связь от людей для оценки качества и юмористичности результатов.
Библ. — 31 назв.

УДК 004.85

Архитектура глубокого обучения для обработки табличных данных с целью извлечения мета-признаков. Гарипов Р., Забашта А., Фильченков А. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 178–190.

В данной работе изучается табличное представление данных в области метаобучения. В ней была предложена архитектура глубокого обучения, которая способна обрабатывать табличные данные, извлекая из них информацию в векторном формате, аналогично традиционным подходам извлечения метапризнаков. Эта архитектура основана на модели deep-sets, а набора данных представляется как множество множеств. Этот подход позволяет обрабатывать табличные данные более естественным образом, поскольку учитывает инвариантность набора данных относительно изменения порядка строк или столбцов. Также было предложено и изучено несколько модификаций этой архитектуры, которые можно применить к задаче многоклассовой классификации. Рассматриваемые архитектуры были экспериментально сравнены в традиционной задаче выбора алгоритма. Результаты эксперимента показали, что модели, которые напрямую обрабатывают набор данных, превосходят модели, основанные только на метапризнаках, но вместе эти подходы работают еще лучше. Однако в некоторых сценариях модель на основе сверточной сети превосходила модель на основе deep-sets.
Библ. — 18 назв.

УДК 004.852

Оптимизация конвейера конструирования признаков в процессе AutoML с использованием больших языковых моделей. Иов И. Л., Никитин Н. О. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 191–222.

Одним из важных путей повышения эффективности автоматизированного машинного обучения является привлечение метаоптимизации

на всех этапах построения конвейера. В работе предлагается использовать большие языковые модели на этапах конструирования признаков одновременно в роли оптимизаторов и экспертов предметной области. Конвейер конструирования признаков кодируется на естественном языке в виде последовательности атомарных операций. Оптимизация методом “чёрного ящика” реализуется путём запроса конвейера конструирования признаков у большой языковой модели с помощью промпта, включающего заранее заданные инструкции, описание набора данных и ранее оценённые конвейеры. Для повышения временной эффективности и устойчивости оптимизации применяется популяционный алгоритм, порождающий на каждый ответ модели набор конвейеров, а не единственный. Для предоставления модели дополнительных знаний предметной области применяется многошаговая оптимизация. Для оценки применимости предложенного подхода проведён ряд экспериментов на открытых наборах данных. В качестве базового метода для задачи оптимизации выбран случайный поиск. Результаты, полученные напрямую с помощью модели gpt-3.5-turbo, близки к базовому методу при тех же временных затратах. Популяционная генерация конвейеров превосходит и базовый метод, и другие подходы. Это подтверждает, что предложенный подход способен повысить общую производительность моделей машинного обучения при тех же временных затратах на оптимизацию и меньшем числе токенов для получения результата.
Библ. — 59 назв.

УДК 004.85

Иерархическая инициализация временных масштабов в Mamba-2 для многомасштабного моделирования последовательностей. Лаэтин А. А., Пулфер Б., Смирнов С. К. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 223–234.

В работе исследуется влияние послойной инициализации параметра Δ_{bias} в архитектуре Mamba-2 на качество языкового моделирования. Рассматриваются три стратегии инициализации: убывающая, возрастающая и равномерная. Проверяемая гипотеза состоит в том, что убывающая стратегия, при которой нижние слои инициализируются с большими Δt, а верхние - с меньшими, способствует формированию многомасштабной обработки информации без изменения архитектуры

модели. Эксперименты проведены на модели Mamba-2 размером около 40M параметров в режиме предварительного обучения на корпусе OpenWebText с последующей оценкой на наборах TinyStories, SQuAD и Wiki-EN. В проведенных экспериментах убывающая инициализация улучшала сходимость по сравнению с равномерной стратегией и в ряде внешних оценок давала более низкую perplexity, тогда как возрастающая стратегия в большинстве сравнений ухудшала качество. Предложенный подход не увеличивает число параметров и может рассматриваться как простой способ повышения эффективности обучения моделей семейства Mamba.
Библ. — 11 назв.

УДК 004.932

Методология сбора и анализа субъективной заметности артефактов нейросетевых моделей суперразрешения. Молодецких И. А., Малышев К. В., Миргалеев М. Р., Богатырёв Е. Н., Ватолин Д. С. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 235–253.

Современные методы суперразрешения изображений, основанные на глубоких нейронных сетях, способны генерировать результаты высокого качества и детализации, однако часто порождают артефакты — ошибочные детали, снижающие визуальное качество изображения. Важно, что их восприятие человеком различается: одни артефакты почти незаметны, тогда как другие серьёзно ухудшают изображение. При этом существующие наборы данных и методы детектирования артефактов используют двоичную разметку (“есть/нет артефакта”), что не отражает субъективную заметность и не позволяет количественно оценивать её влияние на восприятие качества. В данной работе предложена методология субъективной разметки артефактов суперразрешения с учётом их заметности, основанная на краудсорсинге с контролем качества аннотаций. Описаны этапы подготовки данных, включая постобработку масок для упрощения восприятия асессорами, а также подход к агрегированию субъективных оценок. Полученный набор данных включает более 1300 артефактов, созданных одиннадцатью современными моделями суперразрешения. Также проведена аннотация заметности для 593 изображений из существующего набора

артефактов DeSRA и установлено, что большинство из них слабо заметны для человека, несмотря на лабораторную разметку. Предлагаемый набор данных и методология позволяют исследовать артефакты с точки зрения восприятия человека и использовать полученные данные для обучения детекторов, ориентированных на перцептивно-значимые искажения.
Библ. — 28 назв.

УДК 004.89

Методы фильтрации и оценки данных для больших языковых моделей. Николенко С. И. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 254–279.

Современные большие языковые модели критически зависят от качества и состава обучающих данных. Настоящий обзор систематически рассматривает методы фильтрации и оценки обучающих данных для LLM, охватывая широкий спектр подходов: от детерминированных правил фильтрации (ограничения по длине, языку, токсичности, дедупликация) до методов на основе данных-образцов (сходство с референтными корпусами, разница кросс-энтропии), методов на основе моделей (дообученные LLM-оценщики, разреженные автокодировщики, зондирование в контексте), методов на основе энтропии и функции потерь, а также подходов с использованием LLM-судей и классификаторов качества. Мы анализируем преимущества и ограничения каждого класса методов и показываем, что при всём их разнообразии они не обеспечивают принципиальной связи с контрфактическими оценками влияния отдельных примеров на поведение модели. Для решения задач атрибуции данных существует отдельный класс методов — функции влияния, рассматриваемые во второй части обзора.
Библ. — 46 назв.

УДК 004.85

Функции влияния данных для больших языковых моделей. Николенко С. И. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 280–326.

Функции влияния представляют собой математический инструмент для оценки того, как отдельные обучающие примеры влияют на поведение обученных моделей машинного обучения. В отличие от методов фильтрации и оценки качества данных, рассмотренных в первой части обзора, функции влияния обеспечивают принципиальную аппроксимацию контрфактических leave-one-out оценок через локальную линеаризацию и учёт геометрии пространства параметров. Настоящий обзор систематически излагает теорию функций влияния — от классического вывода через повышение весов и теорему о неявной функции до современных масштабируемых аппроксимаций. Мы детально разбираем методы вычисления обратных произведений гессиана на вектор (LiSSA, K-FAC/EK-FAC) и представляем техники, делающие атрибуцию данных практически возможной для моделей с миллиардами параметров: проекционные методы (TRAK, LoGra), методы первого порядка (RapidIn), а также подходы на ранних стадиях обучения (LinFIK, ALinFIK). В заключение мы обсуждаем сильные стороны и ограничения различных подходов и даём рекомендации по выбору методов для конкретных задач атрибуции и отбора данных.
Библ. — 114 назв.

УДК 004.89

Распознавание галлюцинаций в больших языковых моделях с RAG на основе вероятностных расстояний. Обловатный Р., Кулешова А., Полев К., Зайцев А. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 327–349.

Распознавание галлюцинаций в больших языковых моделях (large language models, LLM) критически важно для их безопасного применения во многих задачах. Без надёжного детектирования такие системы зачастую выдают вредоносные и недостоверные ответы. В последние годы LLM активно используются в системах генерации с дополненной выборкой (retrieval-augmented generation, RAG). Однако галлюцинации сохраняются и в этом сценарии, и, хотя было предложено множество методов детектирования галлюцинаций, большинство подходов не разработаны специально для RAG-систем. Чтобы преодолеть это ограничение, мы предлагаем метод детектирования галлюцинаций, основанный на оценке расстояний между распределениями эмбеддингов токенов запроса и эмбеддингов токенов ответа языковой

модели. Метод анализирует геометрическую структуру скрытых состояний токенов, чтобы надёжно извлекать сигнал фактологичности текста, оставаясь при этом применимым к длинным последовательностям. Многочисленные эксперименты показывают, что наш метод достигает наилучших или конкурентоспособных результатов. Кроме того, он обладает переносимостью с задачи NLI на задачу детектирования галлюцинаций, что делает его полностью неконтролируемым и эффективным методом с конкурентоспособным качеством на итоговой задаче.
Библ. — 37 назв.

УДК 004.89

Пожалуйста, проголосуйте: ранжирование ответов на Stack Overflow в текстовой постановке. Орлов Е., Малых В. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 350–380.

Stack Overflow (SO) – давний ресурс для обмена знаниями о программировании, однако в последние годы наблюдается снижение активности пользователей и числа задаваемых вопросов. Одним из факторов, обсуждаемых в предшествующих работах и сообщениях сообщества, является социальное напряжение, связанное с заданием вопросов, особенно для новичков, включая опасения по поводу негативной обратной связи и модерации. Совершенствование автоматических вспомогательных инструментов могло бы помочь сократить число бесполезных взаимодействий и сделать участие более доступным. В данной работе мы исследуем одну из ключевых составляющих ответов на вопросы в сообществе SO: ранжирование ответов на заданный вопрос, которое может служить опорой для инструментов модерации и курирования контента. В отличие от предыдущих работ, опирающихся на метаданные постов и пользователей, мы рассматриваем текстовую постановку, использующую только содержание вопроса и ответа. Такая постановка позволяет ранжировать ответы для недавно созданных учётных записей без надёжных признаков профиля и снижает риск усиления смещений, связанных с репутацией. Мы обучаем набор ранжировщиков на основе языковых моделей, предобученных на данных SO, используя два различных подхода к ранжированию и две архитектуры кодировщиков. Мы наблюдаем, что дообученные модели стабильно превосходят свои аналоги, основанные на замороженных

эмбеддингах, по метрике nDCG. Наша лучшая модель превосходит сильную текстовую базовую модель на основе случайного леса из предшествующих работ. Она также превосходит ранжировщики на основе языковых моделей общего назначения сопоставимого размера, что подчёркивает ценность предобучения в предметной области. Наконец, мы анализируем качество в зависимости от тегов вопросов и выделяем группу тегов, для которых различия статистически значимы.
Библ. — 55 назв.

УДК 004.932

Использование стохастических уязвимостей рандомизированного сглаживания в регрессионных моделях: анализ адаптивных атак и эмпирические ограничения. Шумицкая Е. А., Ватолин Д. С., Анциферова А. В. — В кн.: Исследования по прикладной математике и информатике. VI. (Зап. научн. семин. ПОМИ, т. 552), СПб., 2026, с. 381–397.

Рандомизированное сглаживание недавно было распространено с задач классификации на регрессионные задачи, такие как оценка качества изображений (Image Quality Assessment, IQA), обеспечивая формальные гарантии устойчивости к ограниченным возмущениям входных данных. Среди этих методов медианное сглаживание зарекомендовало себя как устойчивая альтернатива традиционному сглаживанию на основе среднего, обещая большую устойчивость к выбросам. Однако его практическая устойчивость к адаптивным противникам остаётся недостаточно изученной. В данной работе исследуется уязвимость IQA-моделей с медианным сглаживанием к атакам на основе градиента. Мы показываем, что, хотя теоретические гарантии устойчивости выполняются в идеализированных условиях бесконечного сэмплирования, реальные реализации подвержены неопределённости конечной выборки и стохастическому смещению. Чтобы выявить эти слабые места, мы расширяем метод проекционного градиентного спуска (Projected Gradient Descent, PGD) математическим ожиданием по преобразованиям (Expectation over Transformation, EOT), что позволяет осуществлять адаптивные атаки, явно использующие стохастическую природу рандомизированного сглаживания. Эксперименты на трёх современных IQA-моделях – KonCept, PaQ-2-PiQ и HyperIQA – показывают, что предложенная атака существенно увеличивает как Adversarial Gain (AdvG), так и Certified Attack Success Rate (cASR),

превосходя базовые методы, такие как FGM и PGD. Мы также показываем, что устойчивость медианного сглаживания сильно зависит от бюджета сэмплирования и уровня шума: недостаточное сэмплирование или неоптимальные параметры шума приводят к полному разрушению сертификации. Эти результаты выявляют существенный разрыв между теоретической и эмпирической устойчивостью и подчёркивают необходимость улучшенных методов защиты на основе сглаживания в регрессионных задачах.
Библ. — 17 назв.