УДК 004.932

Векторизация изображений: обзор. Дзюба М., Ярский И., Ефимова В., Фильченков А. — В кн.: Исследования по прикладной математике и информатике. II_2. (Зап. научн. семин. ПОМИ, т. 530), СПб., 2023, с. 6–23.

В настоящее время существует множество диффузионных и авторегрессионных моделей, которые показывают впечатляющие результаты для порождения изображений из текста и других входных областей. Однако эти методы не предназначены для синтеза изображений сверхвысокого разрешения. Векторная графика лишена этого недостатка, поэтому создание изображений в этом формате представляется весьма перспективным направлением. Вместо непосредственного создания векторных изображений можно сначала синтезировать растровое изображение, а затем применить векторизацию. Векторизация — это процесс преобразования растрового изображения в аналогичное векторное изображение с использованием примитивных форм. Помимо схожести, сгенерированное векторное изображение также должно содержать минимальное количество фигур для рендеринга. В этой работе мы фокусируемся конкретно на методах векторизации, совместимых с машинным обучением. Мы рассматриваем модели Mang2Vec, Deep Vectorization of Technical Drawings, DiffVG и LIVE. Мы также даем краткий обзор существующих решений, доступных онлайн. Мы также рассматриваем другие алгоритмические методы — модели Im2Vec и ClipGEN — но они не участвуют в сравнении, так как либо открытой реализации этих методов нет, либо их официальные реализации работают некорректно. Наши исследования показывают, что, несмотря на возможность напрямую указывать число и тип фигур, существующие методы машинного обучения работают очень долго и не воссоздают точно исходное изображение. Мы считаем, что не существует быстрого универсального автоматического подхода, и для каждого метода необходим человеческий контроль.

Библ. – 38 назв.

УДК 004.932

Порождение векторной графики большими языковыми моделями: подходы и модели. Тимофеенко Б., Ефимова В., Фильченков А. — В кн.:

Исследования по прикладной математике и информатике. II$_2$. (Зап. научн. семин. ПОМИ, т. 530), СПб., 2023, с. 24–37.

Задача создания векторной графики с помощью искусственного интеллекта недостаточно исследована. В последнее время большие языковые модели (large language models, LLM) успешно применяются для решения многих задач. Например, современные LLM достигают отличного качества в задачах порождения кода и открыты для публичного доступа. В этом исследовании сравниваются подходы к созданию векторной графики с помощью LLM, а именно ChatGPT (GPT-3.5) и GPT-4. GPT-4 имеет заметные улучшения по сравнению с ChatGPT. Обе модели легко генерируют геометрические примитивы, но с трудом справляются даже с простыми объектами. Результаты, полученные с помощью GPT-4, визуально напоминают запросы, но являются неточными. GPT-4 умеет корректировать вывод по инструкции. Кроме того, обеим моделям сложно распознать объект по изображению SVG. Обе модели правильно распознают только примитивные объекты.

Библ. – 20 назв.

УДК 004.852

Фаззинг Python для надежных фреймворков машинного обучения. Егоров И., Кобрин Э., Парыгина Д., Вишняков А., Федотов А. — В кн.: Исследования по прикладной математике и информатике. II$_2$. (Зап. научн. семин. ПОМИ, т. 530), СПб., 2023, с. 38–50.

Обеспечение безопасности и надежности сред машинного обучения имеет решающее значение для создания надежных систем на базе искусственного интеллекта. Фаззинг – популярная техника в жизненном цикле разработки безопасного программного обеспечения (SSDLC), которая может использоваться для разработки безопасного и надежного программного обеспечения. Популярные платформы машинного обучения, такие как PyTorch и TensorFlow, сложны и написаны на нескольких языках программирования, включая C/C++ и Python. Мы предлагаем конвейер динамического анализа для проектов Python с помощью набора инструментов Sydr-Fuzz. В нашем конвейере есть фаззинг, минимизация корпуса, сортировка сбоев и сбор покрытия. Классификация сбоев и оценка их серьезности являются важными шагами, гарантирующими, что наиболее критические уязвимости будут

устранены оперативно. Кроме того, предлагаемый конвейер интегрирован в GitLab CI. Чтобы выявить наиболее уязвимые части фреймворков машинного обучения, мы проанализировали их потенциальные поверхности атаки и разработали цели фаззинга для PyTorch, TensorFlow и связанных с ними проектов, таких как h5py. Применяя наш пайплайна динамического анализа к этим целям, мы смогли обнаружить 3 новые ошибки и предложить исправления для них.

Библ. – 35 назв.

УДК 004.852

Максимизация покрытия нейронов для эффективного построения тестового набора относительно модели. Кущук Д., Рындин М. — В кн.: Исследования по прикладной математике и информатике. II$_2$. (Зап. научн. семин. ПОМИ, т. 530), СПб., 2023, с. 51–67.

Реальные данные не являются стационарными, поэтому модели необходимо отслеживать во время использования. Один из способов убедиться в работоспособности модели — регулярное тестирование. В случае отсутствия размеченных данных можно сформулировать задачу минимизации стоимости разметки. В этой работе мы исследуем и разрабатываем различные способы построения минимального набора тестов для данной обученной модели таким образом, чтобы точность модели, рассчитанной на выбранном подмножестве, была максимально приближена к реальной. Мы фокусируемся на сценарии белого ящика (white box) и предлагаем новый подход, который использует покрытие нейронов в качестве наблюдаемого функционала, который нужно максимизировать для минимизации числа примеров. Мы оцениваем предложенный подход и сравниваем его с байесовскими методами и алгоритмами стратификации, которые являются основными подходами к решению этой задачи в литературе. Разработанный метод показывает примерно такой же уровень производительности, но имеет ряд преимуществ перед конкурентами. Он детерминирован, что исключает разброс результатов. Кроме того, этот метод может дать информацию об оптимальном бюджете.

Библ. – 16 назв.

УДК 81.322.2

Автоматическая оценка методов интерпретации в категоризации текстов. Рогов А., Лукашевич Н. — В кн.: Исследования по прикладной

математике и информатике. II_2. (Зап. научн. семин. ПОМИ, т. 530), СПб., 2023, с. 68–79.

Искусственные нейронные сети начали все больше захватывать повседневную жизнь человека, а сложность нейронных сетей только возрастает. При тестировании на собранных тестовых данных модель может показать вполне приличную производительность, но при использовании в реальных условиях может дать совершенно неожиданные результаты. Чтобы определить причину ошибки, важно знать, как модель принимает решения. В данной работе мы рассматриваем различные методы интерпретации модели BERT в задачах классификации, а также рассматриваем метод оценки методов интерпретации с использованием векторных представлений fastText и GloVe.

Библ. – 15 назв.

УДК 81.322.2

Состязательные атаки на языковые модели: фильтрация WordPiece и синонимы ChatGPT. Тер-Ованесян Т., Алексанян Х., Аветисян К. — В кн.: Исследования по прикладной математике и информатике. II_2. (Зап. научн. семин. ПОМИ, т. 530), СПб., 2023, с. 80–95.

В последние годы состязательные атаки на текст привлекли значительное внимание из-за их потенциальной возможности подорвать надежность моделей обработки естественного языка. Мы представляем новые подходы к созданию состязательных примеров на уровне символов и слов в виде черного ящика, применимые к моделям на основе BERT. Подход на уровне символов основан на идее добавления естественных опечаток в слово в соответствии с его токенизацией WordPiece. В рамках подходов на уровне слов мы представляем три метода, которые используют синонимичные слова-заменители, созданные ChatGPT и затем скорректированные для приведения их в соответствующую грамматическую форму для данного контекста. Кроме того, мы пытаемся минимизировать частоту возмущений, принимая во внимание ущерб, который каждое возмущение наносит модели. Комбинируя подходы на уровне символов, подходы на уровне слов и технику минимизации частоты возмущений, мы достигаем наилучшей производительности атаки. Наш лучший подход работает на 30–65% быстрее,

чем лучший ранее метод Tampers, и имеет сопоставимую частоту возмущений. В то же время предлагаемые возмущения сохраняют семантическое сходство исходного и состязательного примеров и достигают относительно низкого значения расстояния Левенштейна.

Библ. – 22 назв.

УДК 81.322.4

Переведите свою тарабарщину: состязательная атака в модели черного ящика на системы машинного перевода. Чертков А., Цымбой О., Паутов М., Оселедец И. — В кн.: Исследования по прикладной математике и информатике. II_2. (Зап. научн. семин. ПОМИ, т. 530), СПб., 2023, с. 96–112.

Нейронные сети широко применяются в задачах обработки естественного языка в промышленных масштабах и, возможно, чаще всего они используются в составе систем автоматического машинного перевода. В этой работе мы представляем простой способ обмануть современные инструменты машинного перевода при переводе с русского языка на английский и наоборот. Используя новый безградиентный тензорный оптимизатор в модели черного ящика, мы показываем, что многие инструменты онлайн-перевода, в частности Google, DeepL и Яндекс, могут как производить неправильные или оскорбительные переводы для бессмысленных состязательных входных запросов, так и отказываться переводить, казалось бы, безобидные фразы. Эта уязвимость может помешать пониманию нового языка и просто ухудшить опыт пользователя при использовании систем машинного перевода, и, следовательно, для улучшения перевода необходимы дополнительные улучшения этих инструментов.

Библ. – 33 назв.

УДК 004.852

Исследование графовых нейронных сетей для прогнозирования ссылок на уязвимость к атакам на членство. Шайхелисламов Д., Лукьянов К., Северин Н., Дробышевский М., Макаров И., Турдаков Д. — В кн.: Исследования по прикладной математике и информатике. II_2. (Зап. научн. семин. ПОМИ, т. 530), СПб., 2023, с. 113–127.

Графовые нейронные сети (GNN) демонстрируют большие перспективы в решении множества задач, связанных с графовыми данными,

включая системы рекомендаций. Однако по мере того, как GNN получают более широкое распространение в практических приложениях, возникают опасения по поводу их уязвимости к состязательным атакам. Эти атаки могут привести к предвзятым рекомендациям, что потенциально может привести к экономическим потерям и рискам для безопасности. В этой работе мы рассматриваем промышленное применение рекомендательных систем для транспортной логистики и изучаем их уязвимость к атакам на членство (membership attacks). Набор данных представляет собой реальные потоки поездов в России, опубликованные в проекте ETIS. Эксперименты с тремя популярными архитектурами GNN показывают, что все они могут быть успешно атакованы, даже если у противника есть лишь минимальные знания о контексте. В частности, злоумышленник, имеющий доступ только к 1-2% фактических данных, может успешно обучить свою собственную модель GNN, чтобы сделать вывод о наличии связи грузоотправитель-грузополучатель в обучающем наборе с точностью более 94%. Наше исследование также подтверждает, что оверфиттинг является основным фактором, влияющим на эффективность атак на рекомендательные системы.

Библ. – 35 назв.

УДК 004.852

Реалистичные состязательные атаки на детекторы объектов с использованием порождающих моделей. Шелепнева Д., Архипенко К. — В кн.: Исследования по прикладной математике и информатике. II$_1$. (Зап. научн. семин. ПОМИ, т. 530), СПб., 2023, с. 128–140.

Важным ограничением существующих состязательных атак на детекторы реальных объектов является их модель угроз: состязательные методы, основанные на исправлениях, часто создают подозрительные изображения, в то время как подходы с порождением изображений не ограничивают возможности злоумышленника по изменению исходной сцены. Мы разрабатываем модель угроз, в которой злоумышленник изменяет отдельные сегменты изображения и должен создавать *реалистичные* изображения. Мы также разрабатываем и оцениваем атаку

в модели белого ящика (white box), которая использует порождающие состязательные сети и диффузионные модели в качестве генератора вредоносных изображений. Наша атака способна создавать изображения высокой точности, измеренные с помощью расстояния Фреше (FID), и уменьшает mAP модели Faster R-CNN на >0.2 в наборах данных Cityscapes и COCO-Stuff. Реализация нашей атаки на PyTorch доступна по адресу `https://github.com/DariaShel/gan-attack`.

Библ. – 32 назв.

УДК 004.85

Обзор систем моделирования пользовательского отклика в рекомендательных системах. Широких М., Шенбин И., Алексеев А., Володкевич А., Васильев А., Николенко С. И. — В кн.: Исследования по прикладной математике и информатике. II$_2$. (Зап. научн. семин. ПОМИ, т. 530), СПб., 2023, с. 141–190.

За последние несколько десятилетий рекомендательные системы стали неотъемлемой частью как повседневной жизни, так и переднего края исследований в области машинного обучения. В этом обзоре мы исследуем различные подходы к разработке симуляторов рекомендательных систем, особенно подходы к моделированию функции отклика пользователя. Мы рассматриваем вероятностные модели, подходы, основанные на порождающих состязательных сетях, и полномасштабные симуляторы, а также рассматриваем наборы данных, доступные исследовательскому сообществу.

Библ. – 133 назв.