DOI: 10.15514/ISPRAS-2024-36(5)-9

Так ли безопасна интерпретируемость ИИ: взаимосвязь интерпретируемости и защищенности моделей машинного обучения

[1,2] *Г.В. Сазонов, ORCID: 0009-0003-0905-534X <sazonovg@ispras.ru>*
[1,3,4] *К.С. Лукьянов, ORCID: 0009-0009-5235-2175 <lukyanov.k@ispras.ru>*
[5] *С.К. Боярский, ORCID: 0000-0003-0093-3719 <serafimboyarsky@yandex.ru>*
[4,6] *И.А. Макаров, ORCID: 0000-0002-3308-8825 <iamakarov@hse.ru>*

[1] *Институт системного программирования им. В.П. Иванникова РАН, Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.*
[2] *Московский государственный университет имени М.В. Ломоносова, Россия, 119991, Москва, Ленинские горы, д. 1.*
[3] *Московский физико-технический институт (НИУ), Россия 117303, Москва, ул. Керченская, д.1 А, корп. 1.*
[4] *Исследовательский центр доверенного искусственного интеллекта ИСП РАН, Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.*
[5] *Школа анализа данных Яндекса, Россия, 119021, Москва, ул. Тимура Фрунзе, 11, корп. 2.*
[6] *Институт искусственного интеллекта AIRI, Россия, 105064, г. Москва, Нижний Сусальный пер., 5 стр. 19.*

Аннотация. В условиях растущего применения интерпретируемых моделей искусственного интеллекта (ИИ) всё больше внимания уделяется вопросам доверия и безопасности для всех типов данных. В этой работе мы сосредотачиваемся на задаче классификации вершин графов, выделяя ее как одну из самых сложных. Эта работа является первой, насколько нам известно, в которой комплексно исследуется взаимосвязь интерпретируемости и защищенности. Наши эксперименты проводятся на наборах данных: цитирования и графов покупок. Мы предлагаем методики построения атак черного ящика графовых моделей на основании результатов интерпретации, показываем, как добавление защиты влияет на интерпретируемость моделей ИИ.

Ключевые слова: интерпретируемость; защищенность; атаки на модели искусственного интеллекта; атаки черного ящика; классификация вершин графов; доверенный искусственный интеллект.

Для цитирования: Сазонов Г.В., Лукьянов К.С., Боярский С.К., Макаров И.А. Так ли безопасна интерпретируемость ИИ: взаимосвязь интерпретируемости и защищенности моделей машинного обучения. Труды ИСП РАН, том 36, вып. 5, 2024 г., стр. 127–142. DOI: 10.15514/ISPRAS–2024–36(5)–9.

Is AI Interpretability Safe: the Relationship between Interpretability and Security of Machine Learning Models

[1, 2] *G.V. Sazonov, ORCID: 0009-0003-0905-534X <sazonovg@ispras.ru>*
[1, 3, 4] *K.S. Lukyanov, ORCID: 0009-0009-5235-2175 <lukyanov.k@ispras.ru>*
[5] *S.K. Boyarsky, ORCID: 0000-0003-0093-3719 <serafimboyarsky@yandex.ru>*
[4,6] *I.A. Makarov, ORCID: 0000-0002-3308-8825 <iamakarov@hse.ru>*

[1] *Ivannikov Institute of System Programming of the Russian Academy of Sciences, 25, A. Solzhenitsyn st., Moscow, 109004, Russia.*
[2] *Lomonosov Moscow State University, Leninskie Gory, 1, Moscow, 119991, Russia.*
[3] *Moscow Institute of Physics and Technology (National Research University), 1 A, building 1, Kerchenskaya st., Moscow, 117303, Russia.*
[4] *Research Center for Trusted Artificial Intelligence ISP RAS, 25, A. Solzhenitsyn st., Moscow, 109004, Russia.*
[5] *Yandex School of Data Analysis, bld. 2, 11, Timur Frunze st., Moscow, 119021, Russia.*
[6] *AIRI, bld. 19, 5, Nizhnij Susal'nij per., Moscow, 105064, Russia.*

Abstract. With the growing application of interpretable artificial intelligence (AI) models, increasing attention is being paid to issues of trust and security across all types of data. In this work, we focus on the task of graph node classification, highlighting it as one of the most challenging. To the best of our knowledge, this is the first study to comprehensively explore the relationship between interpretability and robustness. Our experiments are conducted on datasets of citation and purchase graphs. We propose methodologies for constructing black-box attacks on graph models based on interpretation results and demonstrate how adding protection impacts the interpretability of AI models.

Keywords: interpretability; robustness; attacks on AI models; Black-box attacks; graph node classification; trusted AI.

For citation: Sazonov G.V., Lukyanov K.S., Boyarsky S.K., Makarov I.A. Is AI interpretability safe: the relationship between interpretability and security of machine learning models. *Trudy ISP RAN/Proc. ISP RAS,* vol. 36, issue 5, 2024. pp. 127-142 (in Russian). DOI: 10.15514/ISPRAS-2024-36(5)-9.

1. Введение

Устойчивость глубоких нейронных сетей к входным возмущениям является важнейшим свойством для их интеграции в различные области машинного обучения, требующие безопасности, такие как беспилотные автомобили, медицинская диагностика и финансы. Хотя нейронные сети должны выдавать схожие результаты для схожих входных данных, давно известно, что они уязвимы для состязательных возмущений [1] – небольших, тщательно продуманных преобразований входных данных, которые не изменяют семантику входного объекта, но заставляют модель выдавать предопределенное решение. Большинство методов изучения состязательной устойчивости нейронных сетей направлены на создание состязательных возмущений, которые указывают на то, что в целом прогнозы нейронной сети ненадежны. Наиболее эффективные и скрытные атаки требуют доступа к градиентам модели и, следовательно, сами по себе имеют мало практического применения [2-4].

Также в последние годы возросло использование интерпретируемых моделей искусственного интеллекта (ИИ) [5-7], что сопровождается возрастающим вниманием к вопросам доверия и безопасности. Методы интерпретации, направленные на повышение прозрачности и объяснимость решений моделей, считаются ключевыми компонентами, укрепляющими доверие пользователей и регуляторов.

Однако, наряду с очевидными преимуществами, интерпретируемость может привести к непреднамеренному раскрытию уязвимостей модели, что открывает возможности для

злоумышленников реализовывать атаки, эксплуатирующие эти слабые места, в том числе в сценарии черного ящика. С другой стороны, попытки защитить модели от таких атак могут нарушить корректность или снизить качество интерпретации, создавая проблему одновременного выполнения требований интерпретации и безопасности. Настоящее исследование фокусируется на анализе этой взаимосвязи, рассматривая, как методы интерпретации могут быть использованы для выявления слабых мест модели, и как добавление механизмов защиты сказывается на качестве интерпретаций. Кроме того, обсуждаются потенциальные направления для разработки систем, сбалансированных по критериям прозрачности и устойчивости к атакам.

Наш вклад можно представить следующим образом:

1. Насколько нам известно, мы первые кто комплексно исследует взаимосвязь интерпретируемости и защищенности моделей ИИ.
2. Мы предлагаем методики построения атак черного ящика основываясь на результатах работы методов интерпретации.
3. Мы предлагаем методику оценки совместимости методов интерпретации с защищенными методами.
4. Мы экспериментально демонстрируем эффективность предлагаемых методик на двух графовых доменах.

Структура статьи организована следующим образом. Во второй главе представлен обзор литературы, в котором рассматриваются основные подходы и результаты, достигнутые в данной области исследования. Третья глава посвящена постановке задач, где формулируются основные определения и цели исследования, которые оно призвано решить. В четвертой главе описывается методология, включающая разработанный методы и методики, применяемые для достижения поставленных целей. Пятая глава содержит описание проведенных экспериментов и анализ полученных результатов. В шестой главе рассматриваются ограничения предложенных подходов и их потенциальное влияние на результаты. Наконец, седьмая глава содержит заключение, в котором подводятся итоги работы, а также намечаются возможные направления для будущих исследований.

2. Обзор литературы

В этом разделе приводится краткий обзор существующих методов интерпретации, атак черного ящика и защит от состязательных атак применительно к моделям, работающим с графовыми данными.

Существуют разные подходы к интерпретации данных. Наиболее распространенным в литературе является апостериорный подход к интерпретации, когда модель сначала обучается, а уже только после этого применяются методы интерпретации. Методы апостериорной интерпретации для моделей, работающих с графовыми данными, можно разделить на три основные группы: методы, основанные на внимании, распространении значимости и градиентные методы. Каждый подход по-своему объясняет предсказания графовых моделей, уделяя особое внимание топологии и структуре графов.

Примеры методы на основе внимания: PGExplainer [8], GraphMask [9]. PGExplainer – метод, который строит вероятностные маски для рёбер, указывая значимость каждого соединения в графе. GraphMask – это метод, который использует обучаемые маски для оценивания значимости ребер и узлов в графе. Примеры градиентных методах: Integrated Gradients [10], GNNExplainer [7]. Integrated Gradients для GNN – это метод, который оценивает вклад каждого узла и ребра через интегрирование градиентов. Он позволяет вычислить вклад каждого элемента графа в итоговое предсказание, расширяя базовый метод Integrated Gradients на графовые структуры. GNNExplainer – работает, оптимизируя маску для рёбер и признаков узлов, чтобы минимизировать разницу между предсказанием модели на полном

графе и на выделенном подграфе. Это позволяет определить, какие рёбра и признаки наиболее значимы для конкретного предсказания. При этом GNNExplainer можно применять для интерпретации на разных уровнях (узла, рёбра или графа). Примеры основанные на распространении значимости: GraphLIME [11], SubgraphX [12].

Эти методы могут помогать находить неточности в работе моделей. Но также эти методы могут, с одной стороны, использоваться для построения атак на модели машинного обучения. С другой стороны, их можно использовать, чтобы оценить уязвимости графовых моделей, закладывая основу для анализа защиты от атак, которые используют уязвимые узлы и связи в графе.

Состязательные атаки на модели обработки графов можно классифицировать по степени доступа к модели (атаки белого и черного ящика), цели атаки (на узлы, подграфы, граф в целом) и переносимости. Основные категории включают целевые и нецелевые атаки, атакующие конкретные узлы или подграфы, и переносимые атаки, которые остаются эффективными даже при изменении модели. Эти подходы направлены на манипулирование предсказаниями модели посредством малозаметных изменений в графе. В данной работе основной акцент будет на целевые атаки черного ящика.

Целевые и нецелевые атаки на узлы и подграфы – эти методы нацелены на изменение конкретных узлов и рёбер, чтобы повлиять на предсказания. Nettack [13] – пример целевой атаки, которая изменяет рёбра и признаки отдельных узлов для нарушения предсказаний. FGSM (Fast Gradient Sign Method) [14], адаптированный для графов, является примером целевой атаки, применяющей градиентные оценки для создания искажений на уровне узлов. Переносимые атаки – атаки, направленные на создание искажений, которые будут эффективны для нескольких моделей одновременно. Mettack [15], основанный на метаобучении, позволяет создавать искажения, сохраняющие свою силу против различных архитектур. Атаки черного ящика – атаки, где злоумышленник не имеет доступа к параметрам модели и полагается на изменения в предсказаниях, чтобы провести атаку. RL-S2V [16] – это метод, использующий обучение с подкреплением для атаки на графовые модели без знания их внутренних параметров, полагаясь только на доступ к предсказаниям модели.

Эти подходы к построению различных типов атак показывают, как изменения в графовой структуре могут нарушить предсказания, подчёркивая необходимость в разработке надёжных защитных механизмов, способных противостоять таким атакам.

Методы защиты от атак на графы можно классифицировать по их подходам к устойчивости: защита на уровне данных, включающая фильтрацию и корректировку графа, архитектурные методы, повышающие устойчивость моделей к искажениям, и регуляризация, предотвращающая зависимость от отдельных узлов или рёбер. Основные стратегии включают фильтрацию, добавление шума и устойчивые к атакам архитектуры. Основное внимание в этой работе будет уделено методам, позволяющим противостоять состязательным атакам.

Фильтрация и корректировка данных – метод, который на уровне данных предотвращает добавление искажений, фильтруя подозрительные ребра и узлы. Jaccard Similarity Filtering [17] выполняет фильтрацию рёбер на основе схожести признаков узлов, препятствуя внедрению вредоносных рёбер. Архитектурные изменения – изменение модели для повышения устойчивости к атакам. Robust GCN [18] добавляет регуляризацию и шум на уровне узлов, снижая чувствительность модели к мелким искажениям. Регуляризация и шум – методы, предотвращающие переобучение на отдельных элементах графа, что делает модель менее восприимчивой к атакам. Также к регуляризации можно отнести Adversarial training [19] и Gradient Regularization [20], который меняют процесс обучения добавляя данные применением атаки во время обучения и ограничивая градиенты. GNNGuard [21] использует

механизмы для обнаружения подозрительных рёбер и снижает их влияние на предсказания, что защищает от атак.

Эти защитные подходы позволяют значительно снизить уязвимость графовых моделей, сохраняя высокую точность предсказаний и надежность модели при использовании графов в критических приложениях. Однако, нет исследований, которые оценивали бы насколько меняется качество интерпретации применительно к защищенным моделям.

Хочется также отметить некоторые постановки, которые встречаются в литературе, где одновременно встречаются вопросы, связанные с парами интерпретация + атаки; интерпретация + защита и сказать, чем они отличаются от постановок, предложенных в этой работе. В работе "Adversarial Detection with Model Interpretation"[22] авторы показывают, как результаты работы метода интерпретации могут использоваться для повышения безопасности за счет того, что интерпретация предоставляет больше информации о процессе принятия решений. Результаты работы демонстрируют эффективность при отсутствии других факторов, но не учитываю, что из-за атак сама интерпретация может ухудшаться. В частности, есть направление атак на ухудшение интерпретации [23-24]. Отличие этой постановки в том, что цель построить атаку в первую очередь на методы интерпретации, например, сделав интерпретацию нестабильной, а не ухудшить прогноз модели на тестовых данных как в постановке атак уклонения. И наиболее близкая постановка представлена в статье "Adversarial Attack on Graph Neural Networks as An Influence Maximization Problem" [24], где авторы анализируют данные основываясь на различных показателях распространения информации в графе и на основании этого формируют атаку. Однако, для анализа распространения информации в графах требуется доступ ко всему графу сразу, что не отвечает сценарию атаки черного ящика, по этой причине сравнения с результатами этой работой не представлено в разделе экспериментов.

На основании обзора литература видно, что существует множество методов интерпретации, атак и защит. Однако, на данный момент не так много вопросов изучено на границе пересечения исследований по обеспечению требований безопасности и интерпретируемости в области доверенного ИИ. Далее подробнее описаны предлагаемые новые постановки задач на стыке обеспечения двух требований доверия и их взаимодействия.

3. Постановка задач

В этом разделе формально вводится постановка задач, вводятся обозначения, используемые в статье, и формулируются вопросы исследования.

3.1 Состязательная атака на нейронную сеть решающую задачу классификации

Предположим, что $f: R^d \to \Delta^K$ – это нейронная сеть, решающая задачу классификации, которая сопоставляет входной объект $X \in R^d$ с вектором $f(X) \in \Delta^K$ вероятностей для K классов, а

$$h(f, X) = arg \max_{i \in [1,\dots,K]} f(X)_i$$

соответствующее правило классификации. Начнем с формального определения состязательного примера для данной нейронной сети и переносимости состязательного примера между двумя сетями.

Definition 3.1 (Состязательный пример). Предположим, что $X \in R^d$ – это входной объект, правильно сопоставленный классу $y \in [1, \dots, K]$ сетью f, а именно, $h(f, X) = y$. Пусть $\delta > 0$ – фиксированная константа. Тогда объект $X' \in R^d: \| X - X' \|_p \leq \delta$ – это нецелевой состязательный пример для f в точке X, если $h(f, X') \neq h(f, X)$, где p, может быть, например, равно $p = 2$ или $p = \infty$.

Если $h(f, X') = t$ для некоторого предопределенного индекса класса t, то X' называется targeted состязательным примером.

Definition 3.2 (Переносимый состязательный пример). Пусть X' будет состязательным примером, вычисленным для сети f в точке X, и пусть $b: R^d \to \Delta^K$ будет некоторой другой сетью. Тогда X' переносится с f на b, если $h(f, X) = h(b, X)$ и $h(f, X') = h(b, X')$.

3.2 Интерпретация моделей

Definition 3.3 (Результат интерпретации). Пусть X – входные данные для модели f, а $M(X) \in [0,1]^d$ – маска, определяющая важность каждого признака X_i во входе X. Здесь $M(X)_i = 1$ указывает на полный вклад i-го признака, $M(X)_i = 0$ означает, что признак не вносит вклада, а значения между 0 и 1 соответствуют частичной значимости признака. Тогда важное подмножество X определяется как $X^{int} = X \odot M(X)$, где $\odot$ обозначает поэлементное умножение. Подмножество X^{int} сохраняет только значимые признаки исходного входа X.

Метрики интерпретации:

- Корректность объяснения (Fidelity). Fidelity измеряет, насколько точно результаты метода интерпретации отражают поведение исходной модели.

$$Fidelity = 1/N \sum\_(i = 1)\^N |f(X\_i\^int) - f(X\_i)|$$

- Разреженность (Sparsity). Sparsity оценивает, насколько просто объяснение, то есть какой процент признаков не участвуют в интерпретации прогноза:

$$Sparsity = (\sum\_(j = 1)\^m 1\{M(X)\_j \neq 0\})/m$$

  где $M(X)_j$ – значение маски интерпретации показывающую вклад j-го элемента, а m – общее количество признаков.

- Стабильность объяснений (Stability). Stability оценивает, насколько похожи объяснения для схожих входных данных. Изменения в объяснениях измеряется посредством добавляя к исходным данным небольшой шум и вычислениями отклонений:

$$Stability = 1/n \sum\_(i = 1)\^n \| M(X\_i) - M(X\_i\^noise) \| \_2$$

  где X_i – исходные данные, X_i^{noise} – данные с небольшими изменениями, $\|\cdot\|_2$ – евклидова норма.

- Согласованность объяснений (Consistency). Consistency измеряет, насколько похожи объяснения для одного и того же примера при разных запусках модели или методов интерпретации:

$$Consistency = 1/n \sum\_(i = 1)\^n cos(M(X)\_i, M(X)\_(i + 1))$$

  где $M(X)_i$ и $M(X)_{i+1}$ – объяснения для одного и того же примера, полученные при разных запусках; cos – косинусное расстояние.

Стоит отметить, что помимо рассматриваемых метрик интерпретации существуют и другие. Однако, выбраны были именно эти метрики на основании рекомендаций и исследований, представленных в литературе. Согласно исследованиям [25], объективные количественные метрики позволяют избежать зависимости от субъективной оценки экспертов и повышают воспроизводимость результатов. Выбранные метрики широко используются и дают количественную оценку интерпретируемости, поскольку они не требуют субъективной оценки и могут быть автоматически рассчитаны на основе выходных данных модели. В литературе отмечено, что вовлечение доменных экспертов усложняет оценку интерпретируемости, так как мнение экспертов может варьироваться в зависимости от их

опыта и восприятия [26]. Выбранные метрики не требуют построения дополнительных интерпретируемых моделей или обучения дополнительных классификаторов. Это существенно упрощает процесс и снижает вычислительную сложность. В исследовании [27] отмечается, что такие метрики полезны для оперативной интерпретации, особенно в приложениях с высокими вычислительными затратами. Выбранные метрики охватывают разные аспекты интерпретируемости: Fidelity позволяет измерить, насколько корректно объяснение отражает работу модели, Stability и Consistency оценивают устойчивость и согласованность объяснений, а Sparsity снижает когнитивную нагрузку, минимизируя количество признаков в интерпретации. Это соотносится с рекомендациями по многомерному анализу интерпретируемости, предложенными в работах [28, 29].

3.3 Защищенная модель

Definition 4 (Защищенная модель). Пусть f – исходная модель, а f' – её защищённая версия. Обозначим $X \in R^d$ как чистые входные данные, а X' – атакованные данные. Модель f' называется защищенной, если выполняются следующие условия:

$$(i) \quad Q(f, X) - Q(f', X) \leq \alpha,$$
$$(ii) \quad Q(f', X') - Q(f, X') \geq \beta,$$

где $Q(\cdot,\cdot)$ – функция качества модели на входных данных, α – максимальное допустимое снижение качества на чистых данных, а β – минимально допустимое увеличение качества на атакованных данных относительно незащищенной модели. Эти условия обеспечивают устойчивость защищенной модели к атакам, сохраняя приемлемый уровень производительности.

Вопросы исследования

- Как, используя результаты работы методов интерпретации, построить состязательный пример в сценарии черного ящика? Для оценки эффективности атак оценивается среднее количество запросов к модели и средний процент успешности атаки.
- Как добавление защитных механизмов влияет на метрики интерпретируемости, указанные в этом разделе?

В следующем разделе мы отвечаем на эти вопросы и предлагаем соответствующие методики.

4. Методология

В данном разделе представлена методика исследования, направленная на достижение поставленных целей и решение заявленных задач. Вначале описана методика построения атак основываясь на результатах работы методов интерпретации. Далее показано как можно оценить влияние методов, обеспечивающих защиту на интерпретируемость моделей.

4.1 Методика построения атаки черного ящика на базе интерпретации

Так как работа делает основной фокус на работу с графовыми данными, то принцип построения атаки может быть направлен на изменения признаков или структуры, или на обе составляющие данных сразу. Также стоит отметить, что вектор направления атаки зависит от того какой метод интерпретации применяется и какой результат этот метод выдает. Так, например, GNNExplainer выдает ранжированную важность ребер и признаков, SubgraphX выдает только важные ребра, а Zorro [30] выдает ранжированную важность вершин и признаков.

4.1.1 Атака на признаки

При атаке на признаки рассмотрим вариант действия: изменение признаков.

Необходимо выбрать некоторый процент признаков относительно общего количество $p\_feature$, который можно модифицировать. Этот параметр играет роль бюджета атаки на признаки. Если важных признаков меньше, чем заданный процент от общего числа, то выбираем исключительно важные, а остальные не модифицируются.

В случае изменения признаков необходимо добавить случайное возмущение ϵ, если признак категориальный, то просто поменять его класс. ϵ играет роль максимального допустимого возмущения.

4.1.2 Атака на структуру

При атаке на признаки рассмотрим три варианта действий: переключение ребер, удаление и добавление.

Аналогично атаке на признаки необходимо выбрать некоторый процент ребер относительно общего количество p_edges, который можно модифицировать. Этот параметр играет роль бюджета атаки на структуру.

В случае удаления важные ребра удаляются. Стоит обратить внимание, что чем ближе ребро к целевой вершине (вершина, которую атакуют, и цель поменять класс именно этой вершины), тем больше информации будет потеряно и атака будет более заметной. В этом смысле вариант с переключением является более предпочтительным с точки зрения незаметности атаки. В случае добавления ребер необходимо выбрать вершину не являющуюся важной (или не являющийся вершиной, из которой выходит или входит важное ребро) и связать ее с важной вершиной (или вершиной, из которой выходит важное ребро). В случае переключения выбираются важное ребро, которое удаляется и добавляется ребро связывающая вершину, из которой выходило важное ребро и связывается с целевой вершиной.

Основная цель методик при построении атакующих возмущений, в том, чтобы удалить/поменять важные элементы в окрестности целевой вершины, так как по определению интерпретации выделяет элементы данных оказавший наибольшее влияние на конечный прогноз модели, при условии высокого значения метрики корректности объяснения. Либо, в случае атаки на структуру, спровоцировать размытие информации при агрегации данных по структуре графа, так как в научных работах было показано это является значительной уязвимостью графовых нейросетей [31-32].

4.2 Методика оценки влияние добавления защиты к модели на качество интерпретации

Основная задача методики, описанной в этом подразделе, оценить, насколько добавление защиты влияет на интерпретируемость моделей.

4.2.1 Метод

1. Построить модель;
2. Обучить;
3. Оценить значения метрик интерпретации для этой модели (Корректность, Стабильность, Согласованность и Разреженность);
4. Добавить защиту применительно к модели;
5. Обучить защищенную модель;
6. Оценить значения метрик интерпретации для защищенного аналога;
7. С помощью метода доверительных интервалов оценить значимость изменений метрик интерпретации.

Также, отметим, что основной задачей предложенной методики является показать возможный принцип качественного анализа взаимосвязи различных аспектов доверия. И при

необходимости методика может быть дополнена новыми метриками оценки интерпретируемости моделей машинного обучения.

5. Эксперименты

В этом разделе описаны эксперименты и все необходимое для их воспроизведения. Первый эксперимент проводился для оценки эффективности атак, основанных на интерпретации, и результаты сравниваются с базовыми методами. Второй эксперимент показывает, как добавление методов защиты влияет на метрики интерпретации.

5.1 Методология экспериментов

5.1.1 Наборы данных и обучение моделей

В экспериментах использовались наборы данных: Cora, CiteSeer, PubMed – относящиеся к домену цитирования [33] и представленных в библиотеке torch-geometric в группе наборов данных Planetoid и Computers, Photo – домена графов покупок [34], также представленные в torch-geometric в группе наборов данных Amazon.

Статистика наборов данных:

- Cora: 2708 вершин, 10556 ребер, 7 классов, 1433 признаков;
- CiteSeer: 3327 вершин, 9104 ребер, 6 классов, 3703 признаков;
- PubMed: 19717 вершин, 88648 ребер, 3 классов, 500 признаков;
- Computers: 13752 вершин, 491722 ребер, 10 классов, 767 признаков;
- Photo: 7650 вершин, 238162 ребер, 7 классов, 745 признаков.

В качестве модели обучалась двухслойная модель GCN (GCN-2l).

GCN-2l

```
(
  Sequential(
    (0):  GCNConv(input_size, 16)
    (1):  ReLU(inplace)
    (3):  GCNConv(16, output_size)
    (4):  LogSoftmax(inplace)
  )
)
```

Для обучения использовался оптимизатор Adam с параметрами по умолчанию и функция ошибки NLLLoss из библиотеки PyTorch, разделение на пакеты данных не используется. Количество эпох обучения равно 200 для достижения высокой точности классификации на всех наборах данных.

5.1.2 Параметры проведения экспериментов

Для всех экспериментов использовался метод интерпретации GnnExplainer на базе реализации библиотеки torch-geometric. Все настройки метода интерпретации взяты по умолчанию из библиотеки, кроме node_mask_type, который был взят со значением "attributes", чтобы метод интерпретации возвращал не только важные ребра, но и важные признаки. Основные эксперименты усредняются на основании 100 запусков на одном и том же наборе вершин. Набор вершин для каждого набора данных и эксперимента выбирался различный из тестового набора вершин. Для расчета метрик интерпретации, требующие добавление шума или требующих внесения дополнительных случайных величин в эксперимент, проводилось дополнительные внутренние 20 запусков для каждого из 100 запусков на фиксированном наборе вершин. Так, например, для оценки метрики Stability на каждой итерации метрика считалась не только на 100 вершинах, но и для каждой отдельной вершины считалось на 20 различных добавлениях случайного шума в данные.

Для методов атаки максимально допустимая величина варьировалась и бралась равной 5, 10 и 15 % соответственно для того, чтобы оценить не только качество атаки, но и изучить как изменения допустимого возмущения влияет на ее эффективность.

В качестве методов защиты использовались методы: Adversarial Training в качестве метода атаки в Adversarial Training использовался метод FGSM с параметром возмущения $\epsilon = 0.01$, GnnGuard и метод Jaccard Defense. Последние два метода использовали все настройки по умолчанию в соответствии с приведенными в оригинальных статьях. GNNGuard: $lr = 0.01, attention = True, drop = True, train\_iters = 50$; Jaccard Defense: $threshold = 0.35$; Методы защиты Jaccard и GNNGuard относятся к защитам от атак отравления, а метод AdvTraining – к защитам от атак уклонения.

5.1.3 Базовый метод и методы для сравнения

По причине, что подход построения атаки основываясь на результатах интерпретации, насколько нам известно, является абсолютно новым, то сравнение предложенных методов проводится с базовым методом вариантами, основанными на случайных изменениях. Также, стоит отметить, что предложенный подход атаки работает за один запрос к модели черного ящика получая только класс объекта в результате работы модели и результаты интерпретации модели на заданном объекте, что не позволяет сравниться с другими подходами атак на модели черного ящика. Это происходит, так как часть подходов изначально предполагает доступ к обучающим данным, чего нет в нашем случае и, что значительно упрощает проведение атаки посредством обучения качественной суррогатной модели. Другие же подходы атакуют черный ящик, которые работают без какой-либо дополнительной информации, что ставит их в заведомо проигрышную позицию по метрике среднего количества запросов (AQN), так как в нашем случае она всегда равна 1.

Все базовые методы использовали аналогичный предлагаемым атакам бюджет в рамках экспериментов, то есть 5, 10 и 15 % соответственно. Базовые методы были случайные атаки, которые переключали, удаляли или добавляли заданный процент ребер или в случае атаки на признаки добавляли случайный шум на заданный процент признаков.

5.1.4 Протокол оценки

Для иллюстрации эффективности предлагаемого подхода построения атак мы приводим значение метрики ASR (средний показатель успешности атаки) и эффективности атаки, которая считается как снижение точности на оригинальных и атакованных данных. Также, отдельно сравниваются между собой варианты построения атак.

Для методики оценки влияние добавления защиты к модели на качество интерпретации считались метрики интерпретации для незащищенной и защищенных моделей. После с помощью методов доверительных интервалов проводится оценка значимости влияния добавления методов защиты на качество интерпретации.

5.2 Результаты экспериментов

В табл. 1 представлено сравнение метрик интерпретации для модели без защиты (Unprotected) и для моделей с использованием 3 защит (Jaccard, AdvTraining и GNNGuard соответственно). Строки представляют наборы данных и метрики интерпретации, наборы данных скомпонованы по доменам. Направление стрелок рядом с метриками соответствует в каком направлении улучшается значение метрики, стрелка вверх означает больше лучше, стрелка вниз – меньше лучше. На основании метода доверительных интервалов с уровнем значимости 5% было выявлено, что на наборе данных PubMed произошло значимое улучшение всех метрик, на наборе данных CiteSeer значимо ухудшилась метрика стабильности для моделей с защитой AdvTraining и GNNGuard и для набора данных Photo произошло значимое улучшение стабильности для всех методов защиты. Также, стоит отметить, что при использовании метода защиты Jaccard в среднем получаются лучше

значение метрик стабильности и разреженности, чем при использовании других методов. Основной причиной может быть то, что этот метод защиты удаляет ребра, которые могут приводить к некорректному распространению информации и могли быть добавлены во время атаки, что положительно сказывается на указанные выше метрики.

Табл. 1. Сравнение метрик интерпретации для незащищенной модели и для модели с различными защитами. Стрелка вверх рядом с метриками показывает, что большее значение соответствует лучшему значению метрики, стрелка вниз – показывает, что меньшее значение соответствует лучшему значению метрики.
Table 1. Comparison of interpretation metrics for the unprotected model and for the model with various defenses. An upward arrow next to the metrics indicates that a higher value corresponds to a better metric value, while a downward arrow indicates that a lower value corresponds to a better metric value.

Набор данных	Метрика	Без защиты	Jaccard	AdvTraining	GNNGuard
Cora	Корректность (↑)	0.97 ± 0.03	0.90 ± 0.09	0.97 ± 0.03	0.97 ± 0.03
	Согласованность (↑)	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
	Стабильность (↓)	1.55 ± 0.87	1.55 ± 0.82	1.90 ± 0.49	1.88 ± 0.46
	Разреженность (↓)	0.04 ± 0.00	0.03 ± 0.00	0.04 ± 0.00	0.04 ± 0.00
CiteSeer	Корректность (↑)	0.87 ± 0.12	0.90 ± 0.09	0.90 ± 0.09	0.90 ± 0.09
	Согласованность (↑)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Стабильность (↓)	0.75 ± 0.39	0.68 ± 0.27	2.38 ± 1.11	2.39 ± 1.14
	Разреженность (↓)	0.03 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	0.03 ± 0.00
PubMed	Корректность (↑)	0.83 ± 0.14	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Согласованность (↑)	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Стабильность (↓)	2.04 ± 1.19	0.15 ± 0.04	0.22 ± 0.07	0.21 ± 0.07
	Разреженность (↓)	0.03 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
Photo	Корректность (↑)	0.87 ± 0.12	0.80 ± 0.16	0.83 ± 0.14	0.83 ± 0.14
	Согласованность (↑)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Стабильность (↓)	1.37 ± 0.55	0.65 ± 0.38	0.56 ± 0.26	0.56 ± 0.29
	Разреженность (↓)	0.60 ± 0.69	0.12 ± 0.06	0.15 ± 0.08	0.15 ± 0.08
Computers	Корректность (↑)	0.77 ± 0.18	0.80 ± 0.16	0.83 ± 0.14	0.83 ± 0.14
	Согласованность (↑)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Стабильность (↓)	1.31 ± 0.52	1.33 ± 0.51	1.18 ± 0.42	1.16 ± 0.42
	Разреженность (↓)	0.22 ± 0.12	0.2 ± 0.1	0.22 ± 0.12	0.22 ± 0.12

В табл. 2 и табл. 3 представлены результаты предложенной атаки построенной на основании работы метода интерпретации в сравнении со случайными атаками, которые были выбраны в качестве базовых методов по причине отсутствия рассматриваемой подхода атаки уклонения в сценарии черного ящика. В таблицах представлено значение метрики среднего процента успешности атаки (ASR), то есть больше значение метрики соответствует большей вероятности успеха атаки.

Табл. 2. Сравнение результатов работы предложенных атак, построенных на основании работы метода интерпретации в сравнении со случайными атаками, которые были выбраны в качестве базовых методов по причине отсутствия рассматриваемой атаки уклонения в сценарии черного ящика. В качестве допустимых модификаций данных рассматривается добавление ребер и изменение признаков. В таблице представлено значение метрики среднего процента успешности атаки (ASR) в процентах, больше значение метрики соответствует большей вероятности успеха атаки.
Table 2. Comparison of the results of the proposed attacks based on the interpretation method against random attacks, which were chosen as baselines due to the absence of the considered evasion attack approach in the black-box scenario. Acceptable data modifications include the addition of edges and changes to features. The table presents the value of the average success rate of the attack (ASR) in percentage, where a higher metric value corresponds to a greater probability of attack success.

Набор данных	Тип атаки	Изменение признаков	Добавление ребер
		$\epsilon = 5\%, 10\%, 15\%$	$\epsilon = 5\%, 10\%, 15\%$
Cora	На основе интерпретации	57±2, 66±2, 69±3	**7±1, 8±1,** 6±1
	Случайная атака	56±3, 66±3, 69±4	2±1, 5±3, 7±1
CiteSeer	На основе интерпретации	**71±7**, 51±5, 45±5	**6±2, 7±2, 8±0**
	Случайная атака	46±4, 56±5, 62±6	2±1, 2±1, 3±2
PubMed	На основе интерпретации	45±1, 44±2, 49±1	**2±1**, 3±1, 3±2
	Случайная атака	68±1, 61±2, 76±2	0±1, 4±1, 3±1
Photo	На основе интерпретации	1±1, 3±2, 7±3	15±2, 16±2, 15±1
	Случайная атака	10±1, 10±2, 14±2	16±1, 17±1, 17±1
Computers	На основе интерпретации	3±1, 5±2, 4±3	**20±1, 22±1, 22±1**
	Случайная атака	7±1, 6±1, 11±2	14±, 15±2, 16±2

Отметим, что метрика среднее число запросов не была проведена, так как во всех случаях она равнялась 1, что лучше любой существующей атаки черного ящика, но прямое сравнение не корректно, так как другие методы не используют результаты интерпретации из-за отличной от рассматриваемой в этой статье постановки.

Из таблиц можно сделать выводы, что атака за счет удаления ребер оказывается неэффективной, атака за счет переключения ребер оказалась значимо эффективнее только на наборе данных Computers, атака за счет добавления ребер оказалась наиболее эффективной и побила базовый метод или была сопоставима на всех наборах данных и атака на основе изменения признаков оказалась эффективнее только на наборе данных CiteSeer.

6. Ограничения

Среди четырех предложенный атак только на одному удалось добиться значимой эффективности. Основной причиной можно выделить то, что подходы на основе переключения и удаления ребер слабо влияют на результаты работы, так как на всех графах при удалении даже всех ребер достигается качество на уровне 60%. В тоже время, добавление ребер является более эффективной стратегией, так как на основании интерпретации. Еще стоит отметить, что увеличение бюджета не приводит к эффективности атаки, а в некоторых случаях приводит к ухудшению.

Табл. 3. Сравнение результатов работы предложенных атак, построенных на основании работы метода интерпретации в сравнении со случайными атаками, которые были выбраны в качестве базовых методов по причине отсутствия рассматриваемой атаки уклонения в сценарии черного ящика. В качестве допустимых модификаций данных рассматривается удаление и переключение ребер. В таблице представлено значение метрики среднего процента успешности атаки (ASR) в процентах, больше значение метрики соответствует большей вероятности успеха атаки.
Table 3. Comparison of the results of the proposed attacks based on the interpretation method against random attacks, which were chosen as baselines due to the absence of the considered evasion attack approach in the black-box scenario. Acceptable data modifications include the deleting and switching edges. The table presents the value of the average success rate of the attack (ASR) in percentage, where a higher metric value corresponds to a greater probability of attack success.

Набор данных	Тип атаки	Удаление ребер	Переключение ребер
		$\epsilon = 5\%, 10\%, 15\%$	$\epsilon = 5\%, 10\%, 15\%$
Cora	На основе интерпретации	1±1, 1±1, 1±1	0±0, 4±1, 6±2
	Случайная атака	0±0, 0±1, 1±1	3±2, 4±1, 6±1
CiteSeer	На основе интерпретации	3±1, 1±1, 3±1	0±0, 1±0, 0±0
	Случайная атака	0±0, 0±0, 0±1	1±0, 3±1, 2±1
PubMed	На основе интерпретации	0±0, 0±1, 0±0	2±1, 1±1, 1±1
	Случайная атака	0±1, 0±1, 0±1	1±1, 3±1, 3±1
Photo	На основе интерпретации	1±1, 0±1, 1±1	12±1, 13±0, 12±1
	Случайная атака	9±1, 8±1, 9±1	17±1, 16±1, 17±1
Computers	На основе интерпретации	0±0, 0±0, 0±0	18±1, 19±2, 19±3
	Случайная атака	0±1, 0±1, 0±1	8±1, 16±2, 12±2

С одной стороны, это говорит об ограниченных возможностях атаки, с другой, о том, что атаке для достижения своей пиковой эффективности требуется очень небольшой бюджет.

Основным ограничением в методики оценки влияния на качество интерпретации методов защиты можно выделить высокую с2тоимость этих оценок. В то же время, подобное сравнение можно позиционировать как бенчмарк, для которого всегда требуются большие вычислительные ресурсы.

Также, хочется отметить, что проводились эксперименты с другими моделями такие как двухслойные модели со сверточными слоями GAT и SAGE, а также, трехслойные модели с этими же типами сверточных слоев. К сожалению, в рамках ограничений на объем работы все эксперименты нет возможности добавить, но все сформулированные выводы корректны и для других рассмотренных архитектур моделей.

7. Заключение и будущая работа

В этой статье был предложен новый подход и соответствующая методика построения атак черного ящика на основании результатов работы методов интерпретации в приложении к графовым данным. Была показана эффективность атаки при добавлении ребер на наборах данных различных доменов: цитирования, графа покупок. При это предложенный подход

требует буквально единичное обращение к модели черного ящика. Также, была предложена методика оценки влияния добавления защиты на качество интерпретации. Эксперименты на популярных методах защиты графовых нейросетей от атак отравления и уклонения показали, что использование методов защиты чаще повышает качество интерпретации, чем ухудшает его.

В качестве направлений будущей работы можно выделить улучшение переносимости атак (ASR) за счет увеличения количества запросов (AQN), так как на практике не обязательно добиваться снижение запросов до 1. Современные атаки в сценариях черного ящика требуют порядка 300-500 запросов для формирования одной атаки и снижения количества запросов даже до 50 будет значительным прорывом, но в тоже время это может позволить значительно улучшить показатель успешной переносимости.

Методику оценки влияния добавление методов защит на качество интерпретации можно развивать до полноценного бенчмарка и оценить влияние методов защиты не только использую другие методы интерпретации, но и на других данных. В будущем мы планируем провести комплексное сравнение значительно большего количества различных методов защит, изучить причины этих результатов и дать практические рекомендации выделяя наиболее эффективные комбинации методов защит и интерпретаций, которые позволят эффективнее внедрять модели ИИ в критическую инфраструктуру.

Список литературы / References

[1]. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, Rob Fergus. Intriguing properties of neural networks. 2nd International Conference on Learning Representations, 2014.

[2]. Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. Explaining and Harnessing Adversarial Examples. CoRR, 2014, abs/1412.6572.

[3]. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[4]. Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 2574–2582.

[5]. Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, Cheekong Lee. Protgnn: Towards self-explaining graph neural networks. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(8): 9127–9135.

[6]. Han Xuanyuan, Pietro Barbiero, Dobrik Georgiev, Lucie Charlotte Magister, Pietro Liò. Global concept-based interpretability for graph neural networks via neuron analysis. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(9): 10675–10683

[7]. Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. Advances in Neural Information Processing Systems, 2019.

[8]. Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, Xiang Zhang. Parameterized explainer for graph neural network. Advances in Neural Information Processing Systems, 2020, 33: 19620–19631.

[9]. Michael Sejr Schlichtkrull, Nicola De Cao, Ivan Titov. Interpreting graph neural networks for NLP with differentiable edge masking. arXiv preprint arXiv:2010.00577, 2020.

[10]. Thomas Schnake, Oliver Eberle, Jonas Lederer, Shinichi Nakajima, Kristof T. Schütt, Klaus-Robert Müller, Grégoire Montavon. Higher-order explanations of graph neural networks via relevant walks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 7581–7596.

[11]. Qiang Huang, Makoto Yamada, Yuan Tian, Dinesh Singh, Yi Chang. Graphlime: Local interpretable model explanations for graph neural networks. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(7): 6968–6972.

[12]. Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. International Conference on Machine Learning, 2021, 12241–12252.

[13]. Daniel Zügner, Amir Akbarnejad, Stephan Günnemann. Adversarial attacks on neural networks for graph data. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, 2847–2856.

[14]. Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. Explaining and Harnessing Adversarial Examples. CoRR, 2014, abs/1412.6572.
[15]. Daniel Zügner, Stephan Günnemann. Adversarial Attacks on Graph Neural Networks via Meta Learning. International Conference on Learning Representations, Workshop Track, 2019, https://arxiv.org/abs/1902.08412.
[16]. Xiang Zhang, Marinka Zitnik. Gnnguard: Defending graph neural networks against adversarial attacks. Advances in Neural Information Processing Systems, 2020, 33: 9263–9275.
[17]. Huijun Wu, Chen Wang, Yuriy Tyshetskiy, Andrew Docherty, Kai Lu, Liming Zhu. Adversarial examples on graph data: Deep insights into attack and defense. arXiv preprint arXiv:1903.01610, 2019.
[18]. Dingyuan Zhu, Ziwei Zhang, Peng Cui, Wenwu Zhu. Robust graph convolutional networks against adversarial attacks. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, 1399–1407.
[19]. Fuli Feng, Xiangnan He, Jie Tang, Tat-Seng Chua. Graph adversarial training: Dynamically regularizing based on graph structure. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(6): 2493–2504.
[20]. Chris Finlay, Adam M. Oberman. Scaleable input gradient regularization for adversarial robustness. arXiv preprint arXiv:1905.11468, 2019.
[21]. Xiang Zhang, Marinka Zitnik. Gnnguard: Defending graph neural networks against adversarial attacks. Advances in Neural Information Processing Systems, 2020, 33: 9263–9275.
[22]. Ninghao Liu, Hongxia Yang, Xia Hu. Adversarial Detection with Model Interpretation. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, 1803–1811, https://doi.org/10.1145/3219819.3220027.
[23]. Shen Wang, Yuxin Gong. Adversarial example detection based on saliency map features. Applied Intelligence, 2022, 52(6): 6262–6275.
[24]. Jiaqi Ma, Junwei Deng, Qiaozhu Mei. Adversarial attack on graph neural networks as an influence maximization problem. Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, 675–685.
[25]. Finale Doshi-Velez, Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608, 2017.
[26]. Zachary C. Lipton. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 2018, 16(3): 31–57
[27]. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016, 1135–1144.
[28]. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys (CSUR), 2018, 51(5): 1–42.
[29]. Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence, 2019, 267: 1–38.
[30]. Thorben Funke, Megha Khosla, Mandeep Rathee, Avishek Anand. Zorro: Valid, sparse, and stable explanations in graph neural networks. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(8): 8687–8698.
[31]. Yiqing Xie, Sha Li, Carl Yang, Raymond Chi-Wing Wong, Jiawei Han. When do GNNs work: Understanding and improving neighborhood aggregation. IJCAI'20: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 2020.
[32]. Xie Y. et al. When do gnns work: Understanding and improving neighborhood aggregation //IJCAI'20: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, {IJCAI} 2020. – 2020. – Т. 2020. – №. 1.
[33]. Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, Tina Eliassi-Rad. Collective classification in network data. AI Magazine, 2008, 29(3): 93–93.
[34]. Julian McAuley, Christopher Targett, Qinfeng Shi, Anton Van Den Hengel. Image-based recommendations on styles and substitutes. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015, 43–52.

Информация об авторах / Information about authors

Георгий Владимирович САЗОНОВ – сотрудник отдела информационных систем института

системного программирования им. В.П. Иванникова Российской академии наук; студент магистратуры МГУ.

Georgii Vladimirovich SAZONOV – employee of the Information Systems Department of the Ivannikov Institute for System Programming of the Russian Academy of Sciences; master's student at Moscow State University.

Кирилл Сергеевич ЛУКЬЯНОВ – исследователь центра доверенного искусственного интеллекта ИСП РАН; аспирант МФТИ.

Kirill Sergeevich LUKYANOV – Researcher at the Center for Trusted Artificial Intelligence of the Ivannikov Institute for System Programming of the Russian Academy of Sciences; postgraduate student at Moscow Institute of Physics and Technology.

Серафим Константинович БОЯРСКИЙ – студент школы анализа данных Яндекса; студент университета ИТМО.

Serafim Konstantinovich BOYARSKY – student of Yandex School of Data Analysis, student of ITMO University.

Илья Андреевич МАКАРОВ – старший научный сотрудник научно-исследовательского института искусственного интеллекта (AIRI), Москва, Россия, где руководит исследованиями в области промышленного ИИ.

Ilya Andreevich MAKAROV – Senior Research Fellow position at Artificial Intelligence Research Institute (AIRI), Moscow, Russia, where he leads the research in Industrial AI.