ИССЛЕДОВАНИЕ СИГНАЛЬНЫХ МЕТОДОВ ВЫЯВЛЕНИЯ СОСТЯЗАТЕЛЬНЫХ АТАК НА НЕЙРОННЫЕ МОДЕЛИ ЧЕРЕЗ ОБЪЯСНИТЕЛЬНУЮ ВИЗУАЛИЗАЦИЮ

Чеколаев Д. А.[1], Шевченко А. В.[2], Алексеев А. К.[3], Трофимов Ю. В.[4], Аверкин А. Н.[5]

DOI: 10.21681/2311-3456-2025-6-140-146

Цель исследования: *разработка сигнального метода определения состояния состязательной атаки на графические объекты, обрабатываемые нейронной сетью при классификации.*

Метод(ы) исследования: *использование метода объяснительного искусственного интеллекта Grad-CAM для визуализации состязательной FGSM-атаки, предложен сигнальный метод на основе интегрального расчета поверхности градиента тепловой карты объяснения.*

Результат(ы) исследования: *описаны и опробованы методы объяснительного искусственного интеллекта для выделения важных признаков и способы использования полученных тепловых карт для обнаружения атак. В практической части рассмотрен один из современных подходов: анализ смещения и размытия объяснений с помощью Grad-CAM. Проведен анализ и обзор эффективности в повышении устойчивости модели к атакам. Выделены различные эффекты воздействия в следствии атак на зоны внимания и характер их изменения. Предложен интегральный метод расчета факта наличия состязательной атаки во входном изображении, что применимо для автоматической детекции атаки.*

Научная новизна: *Исследование направлены на повышение информативности о характере атаки, степени воздействия на атакуемое входное изображение, формирование сигнального метода детектирования наличия состязательной атаки.*

Вклад авторов*: Чеколаев Д. А. и Шевченко А. В. – составление и реализация концепции визуализации состязательной атаки, на основе опубликованных исследований, Алексеев А. К. – описание методов состязательных атак, Трофимов Ю. В. и Аверкин А. В. – теоретическое обоснование применения методов объяснений.*

Ключевые слова: *нейросетевые технологии, атаки на системы искусственного интеллекта, атаки на объяснительный искусственный интеллект, информационная безопасность, визуализация атак на нейронную сеть.*

Введение

В последние годы глубокие нейронные сети достигли выдающихся результатов во множестве задач, однако их применение в критически важных областях сдерживается проблемой состязательных атак [1]. Под такой атакой понимаются специально сконструированные малые возмущения входных данных, незаметные для человека, но приводящие модель к неверному решению [2].

Пусть x — исходное изображение, y — правильный класс изображения, $f(x)$ — функция классификации нейронной сети, θ — параметры модели.

Состязательная атака создаёт возмущение δ, такое что: $\|\delta\| \le \epsilon$, где ϵ – максимально допустимая норма возмущения; $f(x + \delta) \neq y$, при этом $x + \delta$ визуально неотличимо от x.

Процесс создания состязательного примера можно описать как решение оптимизационной задачи:

$$\delta^* = \arg\min_\delta L(f(x + \delta), y) + \lambda\|\delta\|^2, \quad (1)$$

где L — функция потерь, λ — коэффициент регуляризации.

Например, добавление слабого шума к изображению может заставить классификатор ошибочно распознать объект, сохраняя при этом высокую уверенность.

Эти атаки обладают рядом характерных свойств:

1. Практически неразличимы визуально:

$$\|x - (x + \delta)\| \approx 0. \quad (2)$$

2. Целенаправленно управляют результатом (ошибкой):

$$P(f(x + \delta) = y') \gg P(f(x) = y'), \text{ где } y' \neq y. \quad (3)$$

3. Часто переносимы между моделями

$$f_1(x + \delta) = f_2(x + \delta) = y' \neq y, \quad (4)$$

где f_1 и f_2 – разные модели.

То есть, успешное возмущение, рассчитанное для одной архитектуры, может влиять и на другую [2].

Параллельно развивается направление объяснимого искусственного интеллекта (XAI), призванное сделать работу сложных моделей более прозрачной

1 Чеколаев Дмитрий Алексеевич, магистр, Государственный университет «Дубна». г. Дубна, Россия. E-mail: D.1369@icloud.com
2 Шевченко Алексей Валерьевич, старший преподаватель, аспирант, Государственный университет «Дубна», г. Дубна, Россия, E-mail: leviathan0909@gmail.com
3 Артем Кириллович Алексеев, бакалавр, Государственный университет «Дубна». г. Дубна, Россия. E-mail: aak.24@uni-dubna.ru
4 Трофимов Юрий Владиславович, инженер-программист, лаборатория информационных технологий им. Мещерякова, Объединённый институт ядерных исследований (ОИЯИ). г. Дубна, Россия. Аспирант, Государственный университет «Дубна». г. Дубна, Россия. E-mail: ura_trofim@bk.ru
5 Аверкин Алексей Николаевич, кандидат технических наук, ведущий научный сотрудник, Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление», Российская академия наук. г. Москва, Россия. доцент, Государственный университет «Дубна». г. Дубна, Россия. E-mail: averkin2003@inbox.ru

для человека. XAI-методы генерируют объяснения к предсказаниям модели, например, выделяя наиболее важные фрагменты входных данных, повлиявшие на решение. В контексте компьютерного зрения широко применяются визуальные объяснения в виде тепловых карт (heatmap), показывающих значимость каждого участка изображения. Так, метод Grad-CAM строит карту активаций на основе градиентов, указывая, какие области изображения «привлекли внимание» сверточной сети при классификации: холодные тона соответствуют несущественным регионам, а теплые (красные) выделяют наиболее значимые фрагменты [2]. Подобные приемы повышают интерпретируемость моделей и доверие пользователей, особенно в критичных приложениях (медицина, автономный транспорт и др.).

Однако недавние работы выявили ограниченности и уязвимости методов объяснения под воздействием атак [3]. Состязательные воздействия могут приводить не только к сбоям в предсказаниях сети, но и искажать сопровождающие их объяснения, ставя под вопрос достоверность выводов XAI [3]. Иными словами, злоумышленник способен вводить в заблуждение интерпретатор модели, «обманывая» механизм объяснения или приукрашивая истинные причины решения (эффект известен как fairwashing – от англ. «отбеливание») [3]. Это представляет особую опасность в высокозначимых областях применения моделей, где на объяснения возлагается ответственность за обоснование решений. В ответ на данную угрозу исследователи начали изучать методы, повышающие робастность интерпретаций и защищающие объяснения от враждебных манипуляций [3]. Кроме того, набирает силу подход, при котором сами средства объяснения используются для активной защиты моделей – например, для выявления факта атаки или для фильтрации входных данных.

Целью данной статьи является обзор и анализ таких подходов защиты нейросетевых моделей от состязательных атак с помощью объяснительной визуализации. Постановка задачи формулируется через проблему нестабильности XAI-пояснений под атакой. В разделе «Методы» описываются основные подходы XAI для визуализации признаков и то, каким образом визуальные объяснения могут помочь обнаружить факт атаки. Практическая часть посвящена методике из современной литературы: метрике на основе Grad-CAM для анализа атак. [4]

Наконец, в выводах обсуждается, как визуальные объяснения могут быть интегрированы в активную систему защиты моделей, повышая их устойчивость и надежность.

Постановка задачи

Проблема нестабильности XAI-пояснений под атакой, состоит в том, что состязательные атаки не только влияют на выход модели (классификацию), но и способны существенно изменять или искажать генерируемые моделью объяснения.

Пусть $S(x)$ – функция генерации объяснений (например, Grad-CAM) для исходного изображения x; $S(x+\delta)$ — функция генерации объяснений для возмущенного изображения; $d(S_1,S_2)$ – функция расстояния (меры различия) между двумя объяснениями.

Тогда нестабильность объяснений выглядит как:

$$d(S(x),S(x+\delta)) \gg \|x-(x+\delta)\|, \quad (5)$$

где левая часть показывает большое изменение в объяснении, а правая часть отражает малое возмущение входных данных.

Под нестабильностью объяснений понимается ситуация, когда небольшое целенаправленное возмущение входных данных приводит к несоразмерно сильным изменениям в карте значимости модели, отражающей «ход мысли» нейросети. В нормальных условиях объяснение для схожих изображений должно быть относительно устойчивым:

$$\|x_1-x_2\| \le \epsilon \Longrightarrow d(S(x_1),S(x_2)) \le \epsilon', \quad (6)$$

где ϵ' – малая величина.

Напротив, под атакой модель может начать фокусироваться на иных, неинформативных деталях или шумовых пикселях:

$$\exists\delta{:}\|\delta\| \le \epsilon, \text{ но } d(S(x),S(x+\delta)) \gg \epsilon'. \quad (7)$$

Например, показано, что при атаке Grad-CAM тепловые карты систематически смещают свое выделение на другие части изображения, отличные от тех, что были важны для изначального (корректного) распознавания [2]. Это свидетельствует о принципиальном изменении внутренних признаков, используемых сетью, под влиянием атаки.

Более того, атакующий может намеренно эксплуатировать уязвимости метода объяснения. Есть возможность манипулировать объяснением модели без явного ухудшения её точности – вплоть до того, что вредоносная модель может скрывать свои истинные критерии решения, предоставляя правдоподобные, но ложные объяснения наблюдателю [3]. Такое «объяснительное мошенничество» крайне опасно [3], так как подрывает доверие к системе: пользователь получает подтверждение якобы корректной работы модели, тогда как на самом деле ее вывод сфабрикован или обусловлен посторонними факторами.

Постановка задачи сводится к обеспечению устойчивости и достоверности объяснений нейросети перед лицом состязательных воздействий. Необходимо разработать подходы, способные: обнаруживать факт атаки по аномалиям в визуальном объяснении и противодействовать искажению объяснений, сохраняя интерпретируемость модели. Иными словами,

требуется научиться распознавать, когда полученная от модели тепловая карта не соответствует реальным сущностям изображения, а затем либо сигнализировать об атаке, либо автоматически корректировать входные данные или модель для восстановления правильного объяснения. Решение этой задачи позволит сделать системы с глубоким обучением более надежными: объяснение станет не пассивным описанием, а частью механизма защиты, повышающего доверие к модельным предсказаниям.

Методы объяснительной визуализации и обнаружения атак

XAI-подходы для визуализации признаков. Современные методы интерпретируемого ИИ предлагают ряд подходов к выделению значимых признаков, лежащих в основе решения модели. В контексте распознавания изображений наибольшее распространение получили методы визуализации на основе обратного распространения градиента и варианты активационных карт. Классическим примером является Grad-CAM (Gradient-weighted Class Activation Mapping) – метод, использующий градиентные признаки выходного класса для построения карты активаций, локализующей вклад каждого участка изображения [2, 5]. Grad-CAM предоставляет наглядное объяснение: тепловая карта накладывается на изображение, подсвечивая регионы, которые сеть учитывала при принятии решения (синие области не влияют, красные – наиболее значимы). Другой подход – методы распределения атрибуции, к которым относятся Layer-wise Relevance Propagation (LRP) и интегральные градиенты. [6] LRP распространяет выходной скор модели назад по сети, распределяя «релевантность» по входным пикселям, что дает карту важности входа. Интегральные градиенты вычисляют усредненный градиент по пути от некоторого базового состояния входа к данному изображению, выявляя вклад каждого пикселя. Схожей идеей руководствуются методы агностичные к модели, например, LIME и SHAP, [7, 8] которые строят приближенные линейные модели локально или рассчитывают ценность характеристик, удаляя или заменяя части входных данных. Все эти техники в итоге формируют визуальную карту или схему, указывающую, какие признаки (области изображения) обусловили конкретный вывод нейросети.

Уязвимости и устойчивость объяснений. Поскольку методы XAI изначально не разрабатывались с учетом противодействия злоумышленнику, они могут быть обмануты или выведены из строя атакой. Существует несколько направлений повышения их устойчивости. Один из подходов – робастное обучение с учетом объяснений: в функцию потерь модели вводятся дополнительные слагаемые, поощряющие стабильность объяснений при малых изменениях входа. Была предложена регуляризация, сглаживающая поверхность решения: минимизируется разница между картами атрибуции для близко расположенных точек в пространстве данных [3]. Это делает объяснения модели менее чувствительными к небольшим возмущениям на входе. В более поздних работах данная идея развита с помощью вторых производных: регуляризация на основе гессиана нейросети ограничивает изменение градиентных атрибуций при шумовых воздействиях, тем самым повышая робастность градиентных объяснений [3]. В случае методов типа LIME/SHAP предлагается улучшать процедуру выборки примеров: отказ от выбросов и генерация тестовых точек в пределах реального манифолда данных позволяют сделать локальные объяснения более устойчивыми к атакующим вмешательствам [3]. Все эти меры направлены на то, чтобы уменьшить вариативность и чувствительность карт значимости, тем самым лишая атакующего возможности легко их исказить.

Использование визуальных карт для обнаружения атак. Другой важный класс методов – это определение факта атаки по аномальному виду объяснений модели. Предполагается, что под воздействием состязательной атаки визуальное объяснение (тепловая карта) будет статистически отличаться от типичных объяснений для корректных входных данных данного класса. Для количественной оценки различий между тепловыми картами в работе [2] была предложена метрика NISSIM (Normalized Inverted SSIM), представляющая собой нормализованный обратный индекс структурного сходства между двумя картами значимости. Значение NISSIM стремится к 0, если две карты практически идентичны, и к 1 при максимальном различии. Вычисляя NISSIM между объяснением для проверяемого изображения и эталонным объяснением (или между картами до и после предполагаемого возмущения), можно количественно измерить степень искажения объяснения. На основе этой меры предлагаются агрегированные показатели: MOD (Mean Observed Dissimilarity) – среднее значение NISSIM по множеству атакованных образцов при фиксированном уровне атаки, и VID (Variation in Dissimilarity) – разброс NISSIM при вариации параметра атаки (например, величины допустимого шума ε). Метрика MOD характеризует общее снижение схожести пояснений под атакой (чем больше MOD, тем сильнее влияет атака на фокус модели), а VID отражает стабильность модели при усилении атаки (низкий VID означает, что даже при росте ε структура карты меняется предсказуемо и умеренно). В идеале для полностью устойчивой модели объяснения не меняются под атакой, давая NISSIM ≈ 0, MOD ≈ 0 и VID ≈ 0 [2].

Методы обнаружения на основе визуализации привлекательны тем, что используют внутреннюю информацию самой модели (её «взгляд» на данные) для повышения безопасности. Они могут дополнять традиционные детекторы атак, работая на ином уровне — уровне интерпретации решения. Важно отметить, что эффективность таких подходов зависит от надежности самих XAI-методов: если атакующий сумеет одновременно исказить и предсказание, и сопутствующую ему карту значимости, то выявить проблему будет значительно сложнее. Поэтому разрабатываются и более устойчивые способы построения самих карт, и интегрированные схемы защиты, комбинирующие анализ объяснений с контролем исходных данных и ответов модели.

Практическая часть: Применение Grad-CAM для детектирования состязательных атак

В практической части исследования для анализа робастности сверточной нейронной сети (CNN) к состязательным воздействиям применён метод Grad-CAM — интерпретатор, визуализирующий пространственное распределение значимых признаков в последнем сверточном слое модели. Данный подход не требует модификации архитектуры и основан на градиентном взвешивании карт активаций, что делает его пригодным как для post-hoc анализа, так и для включения в конвейер онлайн-защиты.

Сравнительный анализ тепловых карт «чистых» и атакованных входов позволяет локализовать:

- критические регионы изображения, к которым модель наименее робастна;
- чувствительные слои, демонстрирующие резкую перестройку активаций под действием пертурбации;
- смещение внимания сети в нерелевантные области изображения, указывающее на потенциально эксплуатируемые слабости.

Для демонстрации методики были сгенерированы состязательные примеры алгоритмами FGSM и PGD. [9, 10] На рисунках 1 и 2 показано смещение теплового пятна при атаке на изображение кошки: площадь высоких активаций существенно сократилась и сместилась, что коррелирует с ошибкой классификации.

Аналогичный эффект наблюдается для изображения вомбата (рис. 3–5); при низкой контрастности тепловая карта не выявляет информативных областей, тогда как повышение контрастности восстанавливает локализацию значимых признаков, подтверждая корректность диагностической процедуры.

Проведённый анализ показал, что внедрение Grad-CAM в процесс оценки доверия к предсказаниям повышает выявляемость атак: визуальные несоответствия тепловых карт служат индикатором внешнего воздействия, что формирует сигнальный признак. К ключевым достоинствам метода относятся интерпретируемость, архитектурная независимость и низкая вычислительная стоимость. Ограничения заключаются в грубой локализации мелких артефактов, чувствительности к шумным градиентам и пониженной эффективности против атак, искажающих глобальные признаки.

Вместе с тем, следует обращать внимание на важное обстоятельство: в результате атаки зона активации атакованного изображения может принять новое значение в интервале состояний от полного размытия (как на примере с вомбатом, рис. 3, 4 и 5) до частичного смещения пятна внимания (как на примере с котом, рис. 1 и 2). Данное обстоятельство зависит от наличия схожих и близости других изображений в обучающем наборе в параметрическом смысле и индивидуально для определяемых изображений. Для сигнального элемента, который мог бы служить оценочным критерием проще всего работать с полностью рассеянной зоной внимания, так как она при интегральном расчете будет сильно отличаться от четко определяемых изображений.

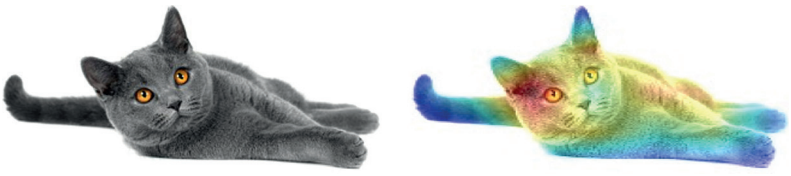

Рис. 1. Тепловая карта чистого изображения

Рис. 2. Тепловая карта атакованного изображения

Рис. 3. Тепловая карта чистого изображения

Рис. 4. Низкоконтрастная тепловая карта атакованного изображения

Рис. 5. Высококонтрастная тепловая карта атакованного изображения

В тоже время отсутствие явного размытия зон активации (смещение ярковыраженнной зоны активации и ее фактическое наличие) требует от нас либо вычисление порогового значения интеграла объема зоны активации, либо ограничивает нас в применимости данного метода детекции факта атаки.

Интеграл объема тепловой карты Grad-CAM представляет собой количественную меру распределения активаций в зоне внимания нейронной сети.

Общая формула интеграла объема может быть представлена следующим образом:

$$V = \iiint_{\Omega} f(x,y,z)\,dV, \quad (8)$$

где: V – интегральный объем зоны активации; Ω – область интегрирования (пространство тепловой карты); $f(x,y,z)$ — функция распределения активаций в точке (x,y,z); dV — элемент объема.

В двумерном случае (для тепловой карты) формула упрощается до:

$$S = \iint_{D} f(x,y)\,dx\,dy, \quad (9)$$

где D — область на плоскости (размер тепловой карты); $f(x,y)$ — интенсивность активации в точке (x,y).

Практическое применение формулы позволяет оценить степень размытия зоны активации, определить смещение фокуса внимания, выявить аномалии в распределении активаций.

При использовании интеграла необходимо установить эталонное пороговые значения для сравнения, учитывать размер и разрешение тепловой карты, принимать во внимание специфику обучающей выборки и анализировать характер распределения активаций.

При анализе атак следует сравнивать полученное значение интеграла с эталонными показателями для корректных изображений, что позволит выявить отклонения, характерные для атакованных образцов.

Заключение

Состязательные атаки ставят под угрозу не только точность глубоких моделей, но и саму концепцию их объяснимости. Выявленная нестабильность

XAI-методов под влиянием атак требует разработки новых подходов к обеспечению надежности как предсказаний, так и пояснений моделей. В данном обзоре рассмотрены современные методы, в которых визуальные объяснения интегрируются в систему защиты нейросети. Ключевые выводы и перспективы этой области можно сформулировать следующим образом:

Во-первых, объяснительная визуализация действительно способна служить индикатором невидимых сбоев в работе модели. Как показывают исследования, малейшие adversarial-воздействия оставляют «след» на внутренних активациях сети, который проявляется в изменении тепловых карт Grad-CAM и других атрибуций. Это открывает возможность использовать XAI для активного мониторинга: непрерывная оценка сходства или аномальности пояснений позволяет в режиме реального времени выявлять атаки до того, как они приведут к нежелательным последствиям. Такой подход особенно ценен в критических приложениях, где доверие к каждому решению модели обязательно должно быть подтверждено стабильным и правдоподобным объяснением.

Во-вторых, отмечается необходимость дальнейших исследований в направлении робастных и безопасных XAI-методов. Уже сейчас предложены подходы к усилению стойкости объяснений (через регуляризацию, улучшение алгоритмов LIME/SHAP и пр.), однако единичные решения не обеспечивают полной безопасности. Требуется разработать стандартизованные протоколы оценки устойчивости объяснений и новые методы, учитывающие возможное противодействие со стороны злоумышленника. Будущее XAI должно сочетать интерпретируемость с надежностью: методы объяснения должны проектироваться с учётом потенциальных атак, а метрики качества моделей – включать показатели безопасности пояснений.

Подводя итог, объяснительная визуализация из пассивного инструмента анализа эволюционирует в активный механизм защиты нейросетевых моделей. Визуальные карты значимости могут выявлять чужеродные вмешательства и выступать триггером защитных процедур. Применение таких подходов повышает общую надежность систем глубокого обучения, поскольку атака должна теперь обмануть не только сам классификатор, но и его «внутренний взор». В дальнейшем совмещение методов XAI и кибербезопасности ИИ обещает создание более прозрачных, доверенных и стойких к атакующим воздействиям моделей. Интеграция объяснимости в защиту не только сохраняет интерпретацию при атаках, но и делает саму интерпретацию щитом, стоящим на страже правильности работы модели.

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 124112200072-2).

Литература

1. Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples //arXiv preprint arXiv:1412.6572. – 2014.
2. Chakraborty T., Trehan U., Mallat K., Dugelay J.-L. Generalizing Adversarial Explanations with Grad-CAM // Proceedings of CVPR Workshop on Art of Robustness, 2022, pp. 186–192. DOI: 10.1109/CVPRW56347.2022.00031.
3. Baniecki H., Biecek P. Adversarial attacks and defenses in explainable artificial intelligence: A survey // Information Fusion, 2024, 107:102303. DOI: 10.1016/j.inffus.2024.102303.
4. Selvaraju R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization //Proceedings of the IEEE international conference on computer vision. – 2017. – С. 618–626.
5. Lucas M. et al. RSI-Grad-CAM: Visual explanations from deep networks via Riemann-Stieltjes integrated gradient-based localization // International Symposium on Visual Computing. – Cham: Springer International Publishing, 2022. – С. 262–274.
6. Bassi P. R. A. S., Dertkigil S. S. J., Cavalli A. Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization //Nature Communications. – 2024. – Т. 15. – №. 1. – С. 291.
7. Gaspar D., Silva P., Silva C. Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron // IEEE Access. – 2024. – Т. 12. – С. 30164–30175.
8. Hariharan S. et al. XAI for intrusion detection system: comparing explanations based on global and local scope //Journal of Computer Virology and Hacking Techniques. – 2023. – Т. 19. – №. 2. – С. 217–239.
9. Huang T. et al. Bridging the performance gap between fgsm and pgd adversarial training //arXiv preprint arXiv:2011.05157. – 2020.
10. Zhong,Z. Improving Model Robustness through Hybrid Adversarial Training: Integrating FGSM and PGD Methods. Applied and Computational Engineering, 109, 57–62. – 2024.

RESEARCH OF SIGNAL METHODS FOR DETECTING ADVERSARIAL ATTACKS ON NEURAL MODELS THROUGH EXPLANATORY VISUALIZATION

Chekolaev D. A.[6], Shevchenko A. V.[7], Alekseev A. K.[8], Trofimov Yu. V.[9], Averkin A. N.[10]

Keywords: *neural network technologies, attacks on artificial intelligence systems, attacks on explainable artificial intelligence, information security, visualization of neural network attacks.*

Purpose of the study: *development of a signal-based method for determining the state of an adversarial attack on graphic objects processed by a neural network during classification.*

Methods of research: *the use of explainable artificial intelligence (Grad-CAM) for visualization of adversarial FGSM attack is employed. A signal-based method relying on integral calculation of the gradient surface of the explanation heatmap has been proposed.*

Result(s): *methods of explainable artificial intelligence for identifying important features and ways to utilize the obtained heatmaps for attack detection have been described and tested. In the practical part, one of the modern approaches is considered: analysis of bias and blurring of explanations using Grad-CAM. An analysis and review of the effectiveness in enhancing model resistance to attacks has been conducted. Various effects of impact resulting from attacks on attention zones and the nature of their changes have been identified. An integral method for calculating the presence of an adversarial attack in the input image has been proposed, which is applicable for automatic attack detection.*

Scientific novelty: *the research is aimed at enhancing the informativeness regarding the nature of the attack, the degree of impact on the attacked input image, and the development of a signal-based detection method for identifying the presence of an adversarial attack.*

References

1. Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples //arXiv preprint arXiv:1412.6572. – 2014.
2. Chakraborty T., Trehan U., Mallat K., Dugelay J.-L. Generalizing Adversarial Explanations with Grad-CAM // Proceedings of CVPR Workshop on Art of Robustness, 2022, pp. 186–192. DOI: 10.1109/CVPRW56347.2022.00031.
3. Baniecki H., Biecek P. Adversarial attacks and defenses in explainable artificial intelligence: A survey // Information Fusion, 2024, 107:102303. DOI: 10.1016/j.inffus.2024.102303.
4. Selvaraju R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization //Proceedings of the IEEE international conference on computer vision. – 2017. – С. 618–626.
5. Lucas M. et al. RSI-Grad-CAM: Visual explanations from deep networks via Riemann-Stieltjes integrated gradient-based localization // International Symposium on Visual Computing. – Cham : Springer International Publishing, 2022. – С. 262–274.
6. Bassi P. R. A. S., Dertkigil S. S. J., Cavalli A. Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization //Nature Communications. – 2024. – Т. 15. – №. 1. – С. 291.
7. Gaspar D., Silva P., Silva C. Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron // IEEE Access. – 2024. – Т. 12. – С. 30164–30175.
8. Hariharan S. et al. XAI for intrusion detection system: comparing explanations based on global and local scope //Journal of Computer Virology and Hacking Techniques. – 2023. – Т. 19. – №. 2. – С. 217–239.
9. Huang T. et al. Bridging the performance gap between fgsm and pgd adversarial training //arXiv preprint arXiv:2011.05157. – 2020.
10. Zhong Z. Improving Model Robustness through Hybrid Adversarial Training: Integrating FGSM and PGD Methods. Applied and Computational Engineering, 109, 57–62. – 2024.

6 Dmitry A. Chekolaev, Master's Degree, State University «Dubna». Dubna, Russia. E-mail: D.1369@icloud.com
7 Alexey V. Shevchenko, Senior Lecturer, Postgraduate student, State University «Dubna», Dubna, Russia, E-mail: Russia leviathan0909@gmail.com
8 Artem K. Alekseev, Bachelor's Degree, State University «Dubna». Dubna, Russia. E-mail: aak.24@uni-dubna.ru
9 Yuri V. Trofimov, Software Engineer, Meshcheryakov Laboratory of Information Technologies, Joint Institute for Nuclear Research (JINR). Dubna, Russia. PhD student, State University «Dubna». Dubna, Russia. E-mail: ura_trofim@bk.ru
10 Alexey N. Averkin, Ph.D., Leading Researcher, Federal Research Center «Informatics and Control», Russian Academy of Sciences. Moscow, Russia. Associate Professor, State University «Dubna». Dubna, Russia. E-mail: averkin2003@inbox.ru