DOI: 10.58168/PPSPM2025_185-191
УДК 004.9

ПРИМЕНЕНИЕ МЕТОДОВ EXPLAINABLE AI (XAI) ДЛЯ ИНТЕРПРЕТАЦИИ РЕШЕНИЙ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

А.М. Тюнина[1], В.В. Кондусова[1], Д.С. Кукуева[1], П.П. Амелин[2]

[1]ФГБОУ ВО «Воронежский государственный лесотехнический университет имени Г.Ф. Морозова»
[2]Ярославское высшее военное училище ПВО имени маршала Советского Союза Говорова Л.А.

Современные глубокие нейронные сети демонстрируют выдающиеся результаты в различных областях, от компьютерного зрения до обработки естественного языка. Однако их сложная внутренняя структура часто делает процесс принятия решений непрозрачным, что порождает проблему «черного ящика». Это является серьезным препятствием для внедрения таких систем в критически важные области, такие как медицина, автономное вождение и юриспруденция, где требуется не только высокая точность, но и понимание логики принятия решений, доверие и возможность аудита. В ответ на этот вызов активно развивается направление Explainable AI (XAI), целью которого является создание методов интерпретации и объяснения решений искусственного интеллекта. В данной статье представлен обзор и сравнительный анализ ключевых методов XAI, включая LIME, SHAP и Grad-CAM. Рассматриваются их теоретические основы, области применения и практическая значимость для обеспечения надежности и безопасности интеллектуальных систем.

Ключевые слова: Explainable AI, интерпретируемость, глубокое обучение, черный ящик, LIME, SHAP, Grad-CAM, компьютерное зрение, доверие к ИИ.

APPLICATION OF EXPLICABLE AI (XAI) METHODS FOR INTERPRETING THE SOLUTIONS OF DEEP NEURAL NETWORKS

A.M. Tyunina[1], V.V. Kondusova[1], D.S. Kukueva[1], P.P. Amelin[2]

[1]Voronezh State University of Forestry and Technologies named after G.F. Morozov
[2]Yaroslavskoye vyssheye voyennoye uchilishche PVO imeni marshala Sovetskogo Soyuza Govorova L.A.

Modern deep neural networks demonstrate outstanding results in various fields, from computer vision to natural language processing. However, their complex internal structure often makes the decision-making process opaque, which creates a "black box" problem. This is a major obstacle to the implementation of such systems in critical areas such as medicine, autonomous driving, and law, which require not only high accuracy, but also an understanding of decision-making logic, trust, and the ability to audit. In response to this challenge, Explicable AI (XAI) is actively developing, which aims to create methods for interpreting and explaining artificial intelligence solutions. This article provides an overview and comparative analysis of key XAI methods, including LIME, SHAP, and Grad-CAM. Their theoretical foundations, fields of application and practical significance for ensuring the reliability and security of intelligent systems are considered.

Keywords: Explicable AI, interpretability, deep learning, black box, LIME, SHAP, Grad-CAM, computer vision, trust in AI.

Стремительное проникновение технологий искусственного интеллекта в повседневную жизнь и критически важные сферы деятельности сопровождается растущим запросом на прозрачность и подотчетность алгоритмов. Глубокие нейронные сети, являясь основным двигателем современного прогресса в ИИ, за счет своей многослойной и нелинейной природы скрывают логику своей работы. Врач не может поставить диагноз, основываясь лишь на непонятном выводе модели; инженер беспилотного автомобиля должен понимать, на что именно система обратила внимание при принятии решения об остановке; заемщик в банке имеет право знать причины отказа в кредите. Таким образом, проблема интерпретируемости перестала быть сугубо академической и превратилась в практическую необходимость, обусловленную этическими, правовыми и социальными требованиями.

Область Explainable AI сформировалась как междисциплинарная научная область, находящаяся на стыке машинного обучения, когнитивной науки и визуализации данных. Среди множества подходов можно выделить три наиболее влиятельных и широко применяемых метода: LIME, SHAP и Grad-CAM. Каждый из них предлагает уникальный способ заглянуть внутрь «черного ящика», отвечая на вопрос «Почему модель приняла именно такое решение?». Цель данной статьи заключается в систематическом обзоре этих методов, описании их принципов работы, сравнительном анализе их сильных и слабых сторон, а также в обсуждении их решающей роли в таких чувствительных к ошибкам областях, как медицина и автономное вождение.

Задача интерпретации модели может быть формализована как поичение для конкретного предсказания функции объяснения, которая является человеко-понятной. Пусть у нас есть модель-классификатор, которая для входного изображения выдает вероятность принадлежности к определенному классу. При этом сама модель представляет собой сложную функцию, mapping входные данные к выходному предсказанию. Проблема заключается в том, что эта функция слишком сложна для непосредственного анализа человеком.

Основные вызовы, связанные с интерпретируемостью, включают в себя установление доверия к системе, обнаружение смещений в данных, обеспечение соблюдения нормативных требований и, что наиболее важно, отладку и улучшение самих моделей. Без понимания того, на каких именно признаках основывается предсказание, невозможно надежно оценить, будет ли модель работать в новых, слегка отличающихся условиях, или же она научилась использовать ложные корреляции, присутствующие в обучающей выборке. Например, модель, предназначенная для диагностики пневмонии по рентгеновским снимкам, может научиться распознавать не медицинские признаки, а метки на снимках, специфичные для определенной больницы, в данных которой она обучалась.

LIME (Local Interpretable Model-agnostic Explanations) предлагает интуитивно понятный подход к объяснению индивидуальных предсказаний. Его ключевая идея заключается в том, что, хотя глобальное поведение сложной модели может быть нелинейным и запутанным, его можно аппроксимировать простой, интерпретируемой моделью в окрестности конкретного рассматриваемого примера. LIME генерирует новые, слегка perturbed версии исходного входного данных, например, зашумленные сегменты изображения, и наблюдает, как меняются предсказания черного ящика для этих новых примеров. Затем он обучает простую модель, такую как линейная регрессия с L1-регуляризацией, чтобы предсказать выход черного ящика на основе этих искусственно созданных samples. Веса этой локальной линейной модели и являются объяснением, показывающим, какие части входных данных наиболее сильно повлияли на исходное предсказание. Важным достоинством LIME является его модель-агностичность, что позволяет применять его к любым алгоритмам, от случайного леса до глубоких нейросетей.

SHAP (SHapley Additive exPlanations) основывается на строгой математической теории из области кооперативных игр. В этой теории ценность каждого игрока в коалиции определяется его средним предельным вкладом во

все возможные коалиции. SHAP адаптирует эту концепцию для машинного обучения, рассматривая признаки как «игроков», а предсказание модели — как «выигрыш» коалиции. Значение SHAP для конкретного признака вычисляется как его средний вклад в предсказание по всем возможным подмножествам остальных признаков. Это обеспечивает теоретически обоснованное и согласованное распределение важности признаков для каждого отдельного предсказания. Хотя точное вычисление значений SHAP является вычислительно сложной задачей, существуют эффективные аппроксимации, такие как KernelSHAP и TreeSHAP, что делает метод применимым на практике. SHAP предоставляет не только локальные объяснения для отдельных экземпляров, но и позволяет агрегировать их для получения глобенного взгляда на поведение модели.

Grad-CAM (Gradient-weighted Class Activation Mapping) был разработан специально для сверточных нейронных сетей, используемых в компьютерном зрении. В отличие от модельно-агностичных методов, Grad-CAM использует внутренние механизмы работы CNN для генерации визуальных объяснений. Его принцип действия основан на анализе градиентов целевого класса, проходящих через последний сверточный слой сети. Эти градиенты используются для взвешивания карт активации этого слоя, которые, по сути, представляют собой абстрактные признаки, извлеченные сетью. Путем взвешенной комбинации этих карт активации Grad-CAM строит грубую тепловую карту, которая накладывается на исходное изображение. Области, выделенные теплыми цветами, показывают те части изображения, которые были наиболее важны для предсказания моделью данного конкретного класса. Например, для задачи классификации пород собак Grad-CAM может визуализировать, что модель сфокусировалась на морде и ушах животного, чтобы принять решение, что повышает доверие к ее корректной работе.

Для демонстрации практической ценности методов XAI рассмотрим их применение в задаче классификации заболеваний по рентгеновским снимкам. Модель глубокого обучения обучается определять наличие пневмонии по грудным снимкам. После достижения высокой точности на тестовой выборке возникает закономерный вопрос: на основе каких признаков модель принимает решение?

Применение Grad-CAM к такому классификатору позволяет получить визуальную карту важности областей. В идеальном случае тепловая карта должна выделять именно легочные поля с инфильтратами, соответствующими

пневмонии. Однако на практике может выясниться, что модель научилась использовать артефакты — например, метки на снимках или особенности конкретного рентгеновского аппарата, что свидетельствует о смещении в данных и ненадежности модели.

Применение LIME к тому же снимку дает иной тип объяснения. Метод может выделить отдельные суперпиксели (сегменты изображения), которые сильнее всего повлияли на предсказание. В отличие от плавной тепловой карты Grad-CAM, объяснение LIME имеет более дискретный характер, но при этом его проще формально проверить, последовательно маскируя эти области и наблюдая за изменением вероятности предсказания.

Расчет вклада признаков с помощью SHAP позволяет количественно оценить важность различных областей. Для пикселя или сегмента изображения вычисляется значение SHAP, показывающее, насколько изменилась бы вероятность предсказания «пневмония», если бы значение этого пикселя было усредненным. Это предоставляет строгую математическую основу для анализа, свободную от эвристик.

Таблица 1 – Сравнительная таблица методов XAI

Критерий	LIME	SHAP	Grad-CAM
Область объяснения	Локальная	Локальная и глобальная	Локальная
Применимость	Модельно-агностичный	Модельно-агностичный	Только для CNN
Вид объяснения	Важность сегментов	Точечные значения вклада	Тепловая карта
Вычислительная сложность	Средняя	Высокая	Низкая
Теоретическая база	Локальная аппроксимация	Теория Шепли	Градиенты и активации

В медицинской диагностике комбинация методов XAI становится стандартом де-факто. Grad-CAM позволяет врачу быстро оценить, на какие анатомические области смотрел алгоритм, в то время как SHAP предоставляет количественную меру уверенности модели в своих выводах. Это не только повышает доверие клиницистов, но и служит инструментом контроля качества — систематический анализ объяснений помогает выявлять случаи, когда модель основывает диагноз на некорректных признаках.

В системах автономного вождения объяснимость является критически важным компонентом безопасности. Методы XAI позволяют инженерам понять, почему система приняла то или иное решение в сложной дорожной

ситуации. В финансовой и правовой сферах требования регулирующих органов делают интерпретируемость обязательной. Такие методы, как LIME и SHAP, позволяют предоставить клиенту понятное объяснение отказа в кредите или страховой выплате, выделив конкретные факторы, повлиявшие на решение — кредитную историю, уровень дохода или другие параметры.

Несмотря на впечатляющие успехи, современные методы XAI имеют ограничения. Локальные объяснения не всегда дают полную картину работы модели, а визуализации могут быть подвержены субъективной интерпретации. Кроме того, существует риск создания «объяснений-софизмов» — когда метод генерирует правдоподобное, но неверное объяснение решения «черного ящика».

Методы объяснимого искусственного интеллекта перестали быть просто академическим интересом и превратились в необходимый компонент промышленных систем машинного обучения. LIME, SHAP и Grad-CAM, каждый со своими особенностями и областью применения, составляют основу современного инструментария для интерпретации решений глубоких нейронных сетей.

Важнейшим выводом является то, что объяснимость следует рассматривать не как опциональную надстройку, а как неотъемлемую часть жизненного цикла интеллектуальной системы — от разработки и валидации до внедрения и мониторинга. Особенно это касается таких ответственных областей, как медицина и автономные системы, где ошибка алгоритма может иметь серьезные последствия.

Перспективы развития XAI видятся в создании стандартизированных фреймворков для оценки качества объяснений, разработке методов, сочетающих достоинства разных подходов, и интеграции принципов интерпретируемости непосредственно в архитектуры нейронных сетей. Дальнейшие исследования должны быть направлены на обеспечение не только локальной, но и глобенной интерпретируемости сложных моделей, что остается одной из самых challenging задач в области искусственного интеллекта.

Список литературы

1. Объединение глубокого обучения и объяснимого ИИ для неинвазивного прогнозирования мутаций EGFR и KRAS в NSCLC: новый радиогеномный подход / Ф. Шариати, В. Павлов, С. Федяшина, Н. Серебренников // V Международная конференция по нейронным сетям и

нейротехнологиям (NeuroNT'2024) : Сборник докладов конференции, Санкт-Петербург, 20 июня 2024 года. – Санкт-Петербург: Федеральное государственное автономное образовательное учреждение высшего образования Санкт-Петербургский государственный электротехнический университет ЛЭТИ им. В.И. Ульянова (Ленина), 2024. – С. 41-44. – EDN RWFMZV.

2. Argumentative XAI: A Survey / K. Cyras, A. Rago, E. Albini [et al.] // IJCAI International Joint Conference on Artificial Intelligence : 30, Virtual, Online, 19–27 августа 2021 года. – Virtual, Online, 2021. – P. 4392-4399. – EDN RDGNIP.

3. Caroprese, L. Argumentation approaches for explanaible AI in medical informatics / L. Caroprese, E. Vocaturo, E. Zumpano // Intelligent Systems with Applications. – 2022. – Vol. 16. – P. 200109. – DOI 10.1016/j.iswa.2022.200109. – EDN TQUGES.

References

1. Combining deep learning and explicable AI for noninvasive prediction of EGFR mutations with Red in NSCLC: a new radiogenomic approach / F. Shariati, V. Pavlov, S. Fedyashina, N. Serebrennikov // X International Conference on Neural Networks and Neurotechnologies (NeuroNT'2024) : Conference proceedings, St. Petersburg, 20 June 2024. – St. Petersburg: Federal State Autonomous Educational Institution of Higher Education St. Petersburg State Electrotechnical University named after V.I. Ulyanov (Lenin), 2024. – pp. 41-44. – ED. RWFMZV.

2. Argumentative XAI: a survey / K. Sairas, A. Rago, E. Albini [et al.] // IJCAI International Joint Conference on Artificial Intelligence : 30, virtual, online, August 19-27, 2021. – Virtual, Online, 2021. – pp. 4392-4399. – PUBLISHING HOUSE of the Russian State Scientific and Technical University.

3. Caroprese, L. Approaches to argumentation for explicable AI in medical informatics / L. Caroprese, E. Vocaturo, E. Zumpano // Intelligent systems with applications. – 2022. – Volume 16. – p. 200109. – DOI 10.1016/j.iswa.2022.200109. – TQUGES PUBLISHING HOUSE.