



ОБЪЯСНИМЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ (ХАИ): МЕТОДОЛОГИЧЕСКИЕ ПОДХОДЫ К ИНТЕРПРЕТАЦИИ РЕШЕНИЙ СЛОЖНЫХ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ

Савельев Артем Николаевич

Аспирант кафедры системного анализа и управления Национальный
исследовательский университет ИТМО
г. Санкт-Петербург, Россия

Аннотация.

В данной статье рассматривается одна из наиболее актуальных проблем современного машинного обучения — отсутствие прозрачности в процессе принятия решений нейронными сетями, известное как проблема «черного ящика». Автор проводит детальный анализ концепции объяснимого искусственного интеллекта (Explainable AI, XAI), исследуя математические и алгоритмические методы, позволяющие интерпретировать логику работы моделей. В работе систематизированы основные подходы к объяснимости: методы локальной интерпретации (LIME), аддитивные объяснения Шапли (SHAP), а также карты внимания и градиентные методы визуализации для сверточных нейронных сетей. Рассматривается критическая значимость объяснимости в таких областях, как медицина, финансовый сектор и автономный транспорт, где цена ошибки крайне высока. В статье аргументируется необходимость поиска баланса между точностью предсказаний модели и доступностью её внутренней логики для человеческого понимания. Предложены критерии оценки качества объяснений и описаны перспективы развития самообъясняющих архитектур нейронных сетей.

Ключевые слова: объяснимый искусственный интеллект, XAI, интерпретируемость моделей, глубокое обучение, значения Шапли, визуализация признаков, этика ИИ, доверие к моделям.

Введение

Стремительное развитие технологий глубокого обучения привело к созданию моделей, способных превосходить человеческие возможности в задачах классификации изображений, распознавания речи и стратегического планирования. Однако по мере усложнения архитектур — от простых линейных регрессий до трансформеров с миллиардами параметров — возникла критическая проблема: человек перестал понимать, на основе каких конкретных признаков модель принимает то или иное решение.

Этот феномен, получивший название проблемы «черного ящика» (Black Box), создает барьеры для внедрения искусственного интеллекта в критически важные сферы жизни общества.

Объяснимый искусственный интеллект (Explainable AI или XAI) — это направление, целью которого является создание методов и инструментов, делающих результаты работы ИИ прозрачными и понятными для экспертов и конечных пользователей. Понимание внутренней логики модели необходимо не только для исправления ошибок и отладки алгоритмов, но и для обеспечения этичности, отсутствия предвзятости и соблюдения законодательных норм. В данной статье мы подробно разберем, как современные исследователи «заглядывают под капот» нейронных сетей и трансформируют сложные математические абстракции в интерпретируемые человеком выводы.

Классификация методов объяснимости и интерпретируемости

В области XAI принято разделять понятия интерпретируемости и объяснимости. Интерпретируемость относится к моделям, которые прозрачны сами по себе в силу своей структуры (например, деревья решений малой глубины или линейные модели). Объяснимость же касается методов, которые применяются к уже обученным сложным моделям (post-hoc методы).

Методы объяснимости классифицируются по нескольким признакам. По охвату они делятся на глобальные (описывают поведение модели в целом на всем наборе данных) и локальные (объясняют конкретное предсказание для одного объекта). По применимости методы бывают модельно-агностическими (подходят для любого алгоритма, от случайного леса до нейросети) и модельно-специфичными (разработанными под конкретную архитектуру, например, методы визуализации весов сверточных слоев). Понимание этой классификации позволяет исследователям выбирать наиболее адекватный инструмент в зависимости от типа данных и требуемого уровня детализации отчета.

Аддитивные объяснения Шапли (SHAP) как математический стандарт

Одним из наиболее теоретически обоснованных подходов в XAI является использование значений Шапли (Shapley Values), заимствованных из теории кооперативных игр. В контексте машинного обучения каждый входной признак (например, возраст пациента или уровень дохода клиента) рассматривается как «игрок», а предсказание модели — как «выигрыш» или общий результат игры. Задача состоит в том, чтобы справедливо распределить этот выигрыш между игроками в зависимости от их вклада.

Математически значение Шапли для признака i рассчитывается как средневзвешенный предельный вклад этого признака во все возможные комбинации (субсеты) других признаков:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

Где n — общее число признаков, $v(S)$ — предсказание модели на подмножестве признаков S . Метод SHAP гарантирует соблюдение важных свойств: локальной точности, отсутствия вклада для незначущих признаков и аддитивности. Это делает его золотым стандартом для объяснения табличных данных, позволяя четко сказать, насколько каждый фактор сдвинул вероятность предсказания относительно среднего значения.

Локальные суррогатные модели и алгоритм LIME

В то время как SHAP опирается на теорию игр, метод LIME (Local Interpretable Model-agnostic Explanations) использует концепцию локальной аппроксимации. Основная идея заключается в том, что даже если глобально модель ведет себя крайне сложно и нелинейно, в окрестности конкретной точки её поведение можно описать простой линейной функцией.

Процесс работы LIME включает создание возмущений вокруг объясняемого объекта (например, случайное удаление слов из текста или замена пикселей изображения на серый цвет). Модель делает предсказания для этих новых синтетических объектов, и на полученных данных обучается простая интерпретируемая модель (например, линейная регрессия), взвешенная по близости к оригиналу. Коэффициенты этой суррогатной модели и служат объяснением. LIME особенно эффективен для текстовых данных, позволяя визуализировать, какие именно слова в письме заставили классификатор пометить его как «спам».

Интерпретация глубоких нейронных сетей: внимание и градиенты

Для компьютерного зрения и обработки естественного языка наиболее эффективны методы, учитывающие внутреннюю структуру нейросетей. В сверточных нейронных сетях (CNN) широко применяется метод Grad-CAM (Gradient-weighted Class Activation Mapping). Он использует информацию о градиентах, поступающих на последний сверточный слой, чтобы построить тепловую карту (heat map), указывающую на области изображения, которые были наиболее важны для принятия решения.

В современных архитектурах на базе трансформеров (Transformer) ключевым механизмом является «внимание» (Attention). Карты внимания позволяют увидеть, на какие слова в предложении ориентируется модель при генерации ответа или переводе. Например, при переводе местоимения модель может «смотреть» на существительное, к которому оно относится. Анализ паттернов внимания помогает исследователям выявлять случаи, когда модель полагается на ложные корреляции в данных (например, когда ИИ узнает собаку на фото только

по наличию травы на заднем плане), что является критически важным этапом верификации системы.

Проблема доверия и необходимость объяснимости в медицине и праве

Внедрение ИИ в системы поддержки принятия врачебных решений требует не только высокой точности, но и возможности обоснования диагноза. Врач не может доверять системе, которая выдает вердикт «рак» без указания на конкретные патологические изменения на снимке МРТ. Здесь ХАИ выполняет роль моста между статистической мощностью алгоритмов и клиническим опытом специалиста.

Аналогичная ситуация наблюдается в юриспруденции и финансовом скоринге. В соответствии с регламентами (такими как GDPR в ЕС), граждане имеют «право на объяснение» в случаях, когда автоматизированное решение существенно влияет на их жизнь (например, при отказе в кредите). Без инструментов объяснимости алгоритмы могут стать источником системной дискриминации, скрыто используя защищенные признаки (расу, пол, религию) через косвенные корреляции. ХАИ позволяет аудиторам проверять модели на предвзятость, обеспечивая социальную справедливость и прозрачность цифрового правосудия.

Перспективы: от объяснения к самообъясняющим системам

Современные исследования в области ХАИ направлены на создание архитектур, которые по своей природе являются объяснимыми. Вместо того чтобы пытаться интерпретировать обученный «черный ящик», ученые разрабатывают модели, которые генерируют объяснение одновременно с предсказанием. К таким подходам относятся нейронные сети на базе прототипов, которые классифицируют объект, находя его сходство с эталонными примерами из обучающей выборки («это птица, потому что её крылья похожи на крылья вот этого зяблика»).

Еще одним трендом является разработка интерактивных интерфейсов для ХАИ. Объяснение не должно быть статичным; пользователь должен иметь возможность задавать уточняющие вопросы («Что изменится, если я увеличу доход на 10%?») и проводить контрфактический анализ. Это превращает искусственный интеллект из директивного оракула в полноценного интеллектуального помощника, чьи выводы можно проверять и оспаривать.

Заключение

Объяснимый искусственный интеллект — это не просто дополнительная надстройка над существующими алгоритмами, а фундаментальное условие безопасного и этичного развития технологий ИИ. Мы находимся в точке, где точность предсказаний перестает быть единственной метрикой успеха. Способность модели быть понятной человеку становится критическим требованием для её интеграции в реальную экономику и социальные институты.

Методы SHAP, LIME и визуализация карт внимания уже сегодня позволяют значительно повысить прозрачность глубокого обучения. Однако впереди еще много работы по стандартизации метрик объяснимости и созданию инструментов, понятных не только разработчикам, но и непрофильным специалистам. Развитие ХАИ позволит преодолеть страх перед «восстанием машин» и создать системы, работающие в симбиозе с человеческим интеллектом, дополняя его мощью вычислений, не лишая при этом контроля и понимания происходящих процессов. Будущее ИИ лежит в области прозрачности, и именно объяснимость станет тем фундаментом, на котором будет выстроено доверие между человеком и алгоритмом.

Литература

1. Абрамова А. В., Савельев А. И. Интерпретируемость моделей глубокого обучения: состояние проблемы и перспективы // Искусственный интеллект и принятие решений. 2023. Т. 15. № 2. С. 45–58.
2. Ручкин В. Н. Объяснимый ИИ: современные методы и архитектуры. СПб.: ИТМО, 2024. 210 с.
3. Arrieta A. B., Díaz-Rodríguez N., Del Ser J. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI // Information Fusion. 2020. Vol. 58. P. 82–115.
4. Lundberg S. M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems (NIPS). 2017. Vol. 30. P. 4765–4774.
5. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2nd ed. 2022. 354 p.
6. Ribeiro M. T., Singh S., Guestrin C. "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. 2016. P. 1135–1144.
7. Samek W., Montavon G., Vedaldi A. et al. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Berlin: Springer, 2019. 438 p.
8. Gunning D., Aha D. W. DARPA's Explainable Artificial Intelligence (XAI) Program // AI Magazine. 2019. Vol. 40. No. 2. P. 44–58.