УДК 004.89

ИНСТРУМЕНТЫ ПОСТХОК-ИНТЕРПРЕТАЦИИ РЕЗУЛЬТАТОВ МОДЕЛЕЙ ГЛУБОКОГО ОБУЧЕНИЯ

Шаповалова К.В., студент гр. 024, I курс
Уральский федеральный университет имени первого Президента России
Б. Н. Ельцина, г. Екатеринбург

Модели глубокого обучения за последние годы стали краеугольным камнем в широком спектре прикладных задач – от медицинской диагностики и автономного вождения до генерации текста, прогнозирования финансовых рисков и автоматической модерации контента. Однако парадокс современной ситуации заключается в том, что по мере роста мощности и универсальности моделей снижается их интерпретируемость.

Глубокие архитектуры – сверточные сети, рекуррентные сети, трансформеры и их производные – функционируют как сложные нелинейные композиции, чья внутренняя логика слабо соотносится с интуитивными представлениями пользователя. В этой связи всё большее значение приобретают инструменты постхок-интерпретации – то есть методы и фреймворки, позволяющие анализировать поведение уже обученной модели ретроспективно, без необходимости встраивания интерпретируемости в саму архитектуру. Эти инструменты становятся необходимым элементом построения доверенных и нормативно соответствующих систем искусственного интеллекта [1].

Под постхок-интерпретацией (от лат. post hoc – «после этого») понимается извлечение осмысленных, человекочитаемых объяснений из уже обученной и функционирующей модели без модификации её структуры или параметров. Это особенно важно в реальных условиях, где разработчик не всегда контролирует обучение (например, при использовании сторонних предобученных моделей) или где архитектура слишком сложна и ресурсоёмка для повторного обучения с ограничениями на интерпретируемость. Кроме того, постхок-интерпретация позволяет анализировать поведение модели на уровне отдельных предсказаний, групп объектов, слоёв сети или даже отдельных нейронов, обеспечивая широкую гибкость и масштабируемость.

Инструменты постхок-интерпретации можно условно разделить на несколько категорий в зависимости от уровня анализа:

1. Локальные методы интерпретации предсказаний. К этой категории относятся подходы, объясняющие конкретное решение модели по конкретному входу. Среди них особое место занимают:

- LIME (Local Interpretable Model-Agnostic Explanations) – обучает локально-аппроксимирующую модель (обычно линейную) на случайно сгенерированных соседствах входа, оценивая вклад каждого признака.
- SHAP (SHapley Additive exPlanations) – вычисляет вклад признаков в предсказание с опорой на теорию игр, обеспечивая математически обоснованную интерпретацию.
- Integrated Gradients – основан на интеграции градиентов по траектории от базового входа к текущему, используется преимущественно в нейросетях.
- DeepLIFT и LRP (Layer-wise Relevance Propagation) – позволяют оценить, как изменения на входе распределяются по слоям и нейронам вглубь модели, выявляя каскад значимости.

Эти методы особенно популярны при анализе текстов и изображений, где результат можно визуализировать в виде тепловых карт, подсвеченных слов или зон активации.

2. Глобальные методы анализа поведения модели. Постхок-интерпретация может охватывать модель в целом, выявляя закономерности в предсказаниях и предпочтениях:
    - Feature Importance Aggregation – агрегирование важности признаков по всей выборке с использованием SHAP, перестановочных тестов или градиентного анализа.
    - Surrogate Modeling – построение простой интерпретируемой модели (например, решающего дерева), аппроксимирующей поведение сложной модели на данных.
    - Partial Dependence Plots (PDP) – визуализация зависимости результата от отдельных признаков при фиксированных значениях остальных.
    - ICE (Individual Conditional Expectation) – обобщение PDP на уровне отдельных примеров.

Эти инструменты особенно важны в табличных задачах, кредитном скоринге, здравоохранении и других областях с нормативной отчётностью, где требуется объяснять не только конкретное предсказание, но и поведение модели в целом.

3. Визуализация внутренних представлений. Модели глубокого обучения строят сложные латентные пространства, в которых признаки преобразуются на каждом слое. Постхок-анализ этих пространств включает:

- Проекции эмбеддингов (t-SNE, UMAP, PCA) – для визуализации кластеров, аномалий, логики разбиений.
- Анализ attention-матриц – для трансформеров, в том числе агрегация по головам, слоям, времени.
- Активационные карты (Activation Maps) – особенно в свёрточных сетях, отображают, на что «смотрит» сеть при обработке изображения.
- Neuron Attribution – выявление нейронов, кодирующих конкретные концепты, признаки или категории.

Эти методы позволяют понять, какие структуры данных «перекодируются» в процессе прохождения по сети и насколько устойчиво формируются признаки.

4. Генерация контрфактических и контрастивных объяснений.  Набирают популярность методы, демонстрирующие, как малое изменение входа может повлиять на решение:
    - Counterfactual Explanations – генерируют минимально отличающиеся примеры, приводящие к другому предсказанию.
    - Contrastive Explanations – формируют объяснение в форме: «эта картинка классифицирована как кот, а не как собака, потому что...».

Такие методы полезны в системах поддержки решений, где требуется обоснованное и actionable-объяснение (ориентированное на действия).

5. Аудит и отслеживание доверия.  Новые инструменты поддерживают мониторинг устойчивости модели, сдвига данных, появления неожиданных зависимостей:
    - Evidently AI, WhyLabs, Fiddler – библиотеки, позволяющие анализировать дрейф признаков, поведение модели на новых данных, аномалии и предвзятость.
    - Model Cards и Explanation Reports – стандарт для представления метаданных модели, включая интерпретации, ограничения, тестовые сценарии.

Инженерная интеграция. Современные инструменты постхок-интерпретации интегрируются в ML-платформы: TensorBoard, PyTorch Captum, Keras Explainability, Google What-If Tool, Microsoft InterpretML и др. Эти фреймворки позволяют реализовать интерфейсы, логгировать объяснения, генерировать отчёты и сравнивать версии моделей. Все чаще интерпретации становятся частью CI/CD-процессов (MLOps), а также используются для верификации, сертификации и соответствия требованиям законодательства (например, GDPR, AI Act) [2].

Вызовы и перспективы. Несмотря на успехи, инструменты постхок-интерпретации сталкиваются с рядом вызовов: отсутствие общих метрик качества объяснений, нестабильность результатов, возможность манипуляции визуализацией, ограниченность объяснений в высокоабстрактных пространствах. Решением может стать интеграция с каузальными моделями, онтологиями, активным обучением и пользовательскими интерфейсами, адаптируемыми к уровню компетентности [3].

В целом, постхок-интерпретация становится критически важным элементом доверенной инженерии ИИ, предоставляя пользователю, регулятору и разработчику инструменты для понимания, контроля и критики модели. Это не просто этап отладки, а элемент этической и социальной инфраструктуры, определяющий, как машинные решения встраиваются в человеческие процессы принятия решений [4, 5].

Список литературы:

1. Nori, H. InterpretML : A Unified Framework for Machine Learning Interpretability / H. Nori, S. Jenkins, P. Koch, R. Caruana // arXiv. – 2019. – arXiv:1909.09223.

2. Wexler, J. The What-If Tool : Interactive Probing of Machine Learning Models / J. Wexler, M. Pushkarna, T. Bolukbasi [et al.] // IEEE Transactions on Visualization and Computer Graphics. – 2020. – Vol. 26, № 1. – P. 56–65. – DOI 10.1109/TVCG.2019.2934619.

3. Arya, V. One Explanation Does Not Fit All : A Toolkit and Taxonomy of AI Explainability Techniques / V. Arya, R. K. E. Bellamy, P.-Y. Chen [et al.] // arXiv. – 2019. – arXiv:1909.03012.

4. Baniecki, H. Dalex : Responsible Machine Learning with Interactive Explainability and Fairness in Python / H. Baniecki, W. Kretowicz, P. Piatyszek [et al.] // Journal of Machine Learning Research. – 2021. – Vol. 22. – P. 1–7.

5. Alibi Explain : Algorithms for Explaining Machine Learning Models / S. Klaise, A. Van Looveren, G. Vacanti, A. Coca // Journal of Machine Learning Research. – 2021. – Vol. 22. – P. 1–7.