УДК 004.89

СОЗДАНИЕ ИНСТРУМЕНТОВ ВИЗУАЛИЗАЦИИ ПРИНЯТИЯ РЕШЕНИЙ ПРЕДОБУЧЕННЫМИ МОДЕЛЯМИ

Абдрашитов М.С., Оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

С ростом масштабов и применения предобученных моделей в реальных, чувствительных к ошибкам задачах – таких как медицинская диагностика, кредитный скоринг, отбор кандидатов на работу, предсказание правонарушений и другие – всё более остро встаёт вопрос: на каком основании модель приняла то или иное решение?

Когда пользователи – будь то врачи, аналитики, юристы или сами разработчики – не могут понять, почему модель классифицировала объект определённым образом или почему дала именно такой прогноз, теряется не только доверие к системе, но и возможность её отладки, адаптации, правовой и этической ответственности. В этом контексте визуализация процесса принятия решений становится неотъемлемым компонентом инженерии доверенного искусственного интеллекта. Она позволяет превратить «чёрный ящик» в прозрачный, проверяемый и объяснимый механизм, особенно в тех случаях, когда поведение модели основано на сложных нелинейных и многомерных зависимостях, непонятных на уровне числовых весов или логов [1].

Создание инструментов визуализации принятия решений предобученными моделями – это не просто задача интерфейса или представления данных. Это область на стыке вычислительной интерпретируемости, когнитивной психологии, графической семиотики и программной инженерии. Целью таких инструментов является не просто показать, «что произошло внутри модели», а сделать это таким образом, чтобы человек понял, принял и – при необходимости – поставил под сомнение логику модели [2].

Первым шагом при проектировании визуализации становится определение уровня, на котором будет предоставлено объяснение. Можно выделить несколько ключевых уровней: [3]

1. Локальное объяснение (instance-level) – пояснение того, почему модель выдала конкретное предсказание для конкретного входа. Это, например, тепловая карта признаков, повлиявших на классификацию, или шкала значимости слов в тексте.

2. Глобальное объяснение (model-level) – визуализация логики всей модели: какие признаки наиболее значимы, какие классы путаются, как выглядят границы решений.
3. Интерактивное сравнение (comparison-level) – сравнение разных моделей, версий или решений на схожих данных [4].
4. Признаковое пространство (latent-level) – проекция скрытых представлений модели (эмбеддингов, attention-матриц и т.д.) в человекочитаемом виде.

Среди наиболее широко применяемых методов визуализации можно выделить: [5]

- SHAP-графики (Shapley Additive Explanations) – показывают вклад каждого признака в итоговое решение, основаны на кооперативной теории игр. Используются как для отдельных предсказаний, так и для сводной картины по всей выборке. Часто отображаются в виде горизонтальных баров с цветовой шкалой.
- LIME (Local Interpretable Model-agnostic Explanations) – строит локальные интерпретируемые модели (обычно линейные), визуализируя вклад признаков для конкретного случая.
- Grad-CAM и Score-CAM – визуализация значимых областей на изображениях для моделей компьютерного зрения, показывающая, какие участки активировали определённые фильтры сети.
- Attention Maps и Attention Flow – применимы к трансформерным архитектурам, позволяют проследить, на какие токены модель «смотрела» при генерации ответа или классификации.
- Feature Attribution Graphs – визуализируют связи между признаками, особенно полезны при обнаружении взаимодействий и зависимости.
- Projection Methods (UMAP, t-SNE, PCA) – позволяют визуализировать скрытые пространства признаков или эмбеддингов, часто применяются для анализа кластеров, outlier-объектов и смещений.

Ключевой задачей при разработке таких инструментов является обеспечение когнитивной доступностивизуализации: пользователю должно быть понятно не только «что» отображено, но и как это связано с его предметной областью. В медицинских системах, например, визуализация не должна просто подсвечивать пиксели на изображении МРТ, а должна сопровождаться пояснением: «повышенная значимость области, связанной с лобной корой, может указывать на риск патологии класса X» [6].

В юридических приложениях объяснение должно следовать структуре доказательства: «предложение классифицировано как угрожающее, поскольку

содержит модальный глагол, выражающий намерение, и лексему, связанную с причинением вреда».

Современные визуализационные фреймворки – такие как What-If Tool от Google, Eli5, ExplainerDashboard, Lucid, Captum, Netron, Facets и другие – предоставляют разработчикам широкие возможности по интеграции визуальных объяснений в пользовательский интерфейс или MLOps-инфраструктуру. Они позволяют настраивать интерактивные панели, переключать слои модели, анализировать метки внимания, сравнивать поведение между срезами данных и выявлять зоны «технического долга» в логике модели. Некоторые из них (например, Captum) тесно интегрированы с фреймворками обучения (PyTorch, TensorFlow), что позволяет создавать кастомные визуализации даже для нестандартных архитектур.

Существуют также перспективные разработки в области пользовательской адаптации визуализации. Это подходы, при которых система не просто показывает объяснение, но и подстраивает его под уровень компетенции пользователя. Например, один и тот же вывод может быть визуализирован для врача и для пациента по-разному: первый получит клинические метки и параметры, второй – графическое отображение риска. Такие адаптивные визуализации основаны на метаинформации о пользователе, его предпочтениях, истории взаимодействий и задачах, что делает систему более гибкой и доверительной.

Особое направление – это визуализация отклонений от ожидаемого поведения, когда инструменты фиксируют и объясняют аномалии в предсказаниях. Это может быть полезно для мониторинга дрейфа данных, ошибок модели, adversarial-атак или изменений в распределении пользователей. В таких системах визуализация становится не просто элементом объяснения, а инструментом диагностики и принятия решений, как для инженеров, так и для регуляторов.

Наконец, важным аспектом является документирование визуальных объяснений: в ряде случаев (например, в рамках законодательства ЕС) требуется хранение не только предсказания, но и объяснения, данное пользователю. Это означает, что визуализация должна быть воспроизводимой, сериализуемой и проверяемой, что требует стандартизации форматов, версионности и контроля изменений визуальных модулей.

Таким образом, создание инструментов визуализации принятия решений предобученными моделями – это не вспомогательная задача, а стратегический компонент доверенного ИИ, формирующий прозрачность, подотчётность и взаимодействие человека и машины.

Такие инструменты позволяют не только увидеть поведение модели, но и понять, принять или оспорить его – что в условиях роста влияния ИИ на общество становится вопросом не удобства, а принципа технологической этики.

Список литературы:

1. Yosinski, J. Understanding Neural Networks through Deep Visualization / J. Yosinski, J. Clune, A. Nguyen [et al.] // Deep Learning Workshop, International Conference on Machine Learning. – 2015.

2. Hohman, F. Summit : Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations / F. Hohman, H. Park, C. Robinson, D. H. Chau // IEEE Transactions on Visualization and Computer Graphics. – 2020. – Vol. 26, № 1. – P. 1096–1106. – DOI 10.1109/TVCG.2019.2934659.

3. Strobelt, H. LSTMVis : A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks / H. Strobelt, S. Gehrmann, H. Pfister, A. M. Rush // IEEE Transactions on Visualization and Computer Graphics. – 2018. – Vol. 24, № 1. – P. 667–676. – DOI 10.1109/TVCG.2017.2744158.

4. Kahng, M. GAN Lab : Understanding Complex Deep Generative Models using Interactive Visual Experimentation / M. Kahng, N. Thorat, D. H. Chau [et al.] // IEEE Transactions on Visualization and Computer Graphics. – 2019. – Vol. 25, № 1. – P. 310–320. – DOI 10.1109/TVCG.2018.2864500.

5. Smilkov, D. Direct-Manipulation Visualization of Deep Networks / D. Smilkov, N. Thorat, C. Nicholson [et al.] // arXiv. – 2017. – arXiv:1708.03788.

6. Свидетельство о государственной регистрации программы для ЭВМ № 2025662611 Российская Федерация. Программа для создания визуализации кластерной структуры моделей глубокого обучения : заявл. 24.04.2025 : опубл. 22.05.2025 / Е. А. Николаева, П. А. Пылов, Р. В. Майтак ; заявитель Федеральное государственное бюджетное образовательное учреждение высшего образования «Кузбасский государственный технический университет имени Т.Ф. Горбачева». – EDN UVVDYS.