УДК 004.89

РАЗРАБОТКА ИНСТРУМЕНТОВ ЛОКАЛЬНОЙ ИНТЕРПРЕТАЦИИ РЕШЕНИЙ НЕЙРОННЫХ СЕТЕЙ

Буров Т.В., студент гр. М3216, I курс
Университет ИТМО, г. Санкт-Петербург

В условиях широкого распространения нейронных сетей в критически значимых сферах – от медицины и юриспруденции до кибербезопасности и промышленного контроля – потребность в объяснимости и доверии к принимаемым решениям становится не просто желательной, а обязательной [1].

Особенно это касается ситуаций, когда пользователь (врач, инженер, аналитик, юрист) сталкивается не с обобщённой моделью, а с конкретным её выводом: диагнозом пациента, отказом в кредите, предупреждением о неисправности или классификацией текста. В этих случаях недостаточно понимания общей логики модели или её средних показателей на выборке – необходима локальная интерпретация, то есть объяснение конкретного предсказания на конкретном входе. Разработка инструментов, обеспечивающих такую интерпретацию, является сегодня одним из центральных направлений в области доверенного искусственного интеллекта и интерпретируемого машинного обучения.

Локальная интерпретация в контексте нейросетей означает возможность ответить на вопрос: какие элементы входных данных (пиксели, токены, признаки) повлияли на конкретное решение модели и каким образом. При этом важным требованием является не только техническая точность объяснения, но и его когнитивная понятность для человека, проверяемость, воспроизводимость и устойчивость к манипуляциям. Современные нейронные сети, особенно глубокие и трансформерные архитектуры, обладают высокой вычислительной мощностью и способностью выявлять сложные паттерны, но именно эта сложность делает их внутреннюю логику непрозрачной. Поэтому локальная интерпретация требует разработки особых методов, которые могут «пролить свет» на скрытые процессы в модели без вмешательства в её архитектуру.

Наиболее известными подходами к локальной интерпретации являются методы на основе оценки важности признаков, включая:

- LIME (Local Interpretable Model-agnostic Explanations) – метод, создающий локально линейную аппроксимацию модели вокруг интересующего примера, и определяющий вклад каждого признака в результат. Применим к любым моделям, но чувствителен к параметрам генерации интерпретации [2].

- SHAP (Shapley Additive Explanations) – основан на теории игр, предлагает справедливую и математически обоснованную оценку вклада каждого признака. Является более устойчивым, но требует значительных ресурсов [3].
- Integrated Gradients – работает для дифференцируемых моделей, вычисляет интеграл градиентов по пути от базового значения до реального входа, отражая вклад каждого признака [4].
- Grad-CAM / Guided Backpropagation – применимы к сверточным сетям, позволяют визуализировать, какие участки изображения активируют определённые фильтры [5].

Однако за пределами этих классических методов развивается целая экосистема инструментов, ориентированных на гибкую, контекстную и адаптируемую интерпретацию решений. Например, для языковых моделей (BERT, GPT и др.) применяются визуализации attention-механизмов, а также методы анализа скрытых представлений и генерации контрпримеров. В компьютерном зрении всё большее внимание уделяется методам, способным сочетать визуальную и семантическую информацию – от суперпиксельной интерпретации до мультимодальных тепловых карт, объединяющих картинку с текстом пояснения.

Важным направлением становится интеграция локальных интерпретаций в пользовательский интерфейс. Это позволяет сделать объяснение неотъемлемой частью взаимодействия с моделью. Например, в медицинском приложении классификация изображения МРТ сопровождается подсвечиванием участков, повлиявших на диагноз, и текстовым описанием паттернов. В задачах обработки текста отображаются слова, на которые модель опиралась при решении. Такое представление объяснений повышает доверие пользователей, особенно если они могут наглядно сопоставить вывод модели со своими профессиональными знаниями.

Однако существует и ряд проблем и ограничений, которые необходимо учитывать при разработке инструментов локальной интерпретации. Во-первых, интерпретация может быть нестабильной: при небольших изменениях входа объяснение может кардинально измениться. Это снижает доверие к интерпретации как таковой. Во-вторых, существует риск ложной уверенности: даже если объяснение выглядит правдоподобным, оно может не отражать реальную логику модели. В-третьих, интерпретации могут быть подвержены атакам – так называемым adversarial explanations, которые изменяют видимое объяснение без изменения вывода.

Для преодоления этих проблем разрабатываются метрики качества интерпретации, включая: стабильность, воспроизводимость, согласованность с другими объяснениями, чувствительность к важным признакам, интуитивная

понятность. Кроме того, важным направлением является включение доменной информации в процесс интерпретации: например, возможность отображения признаков в терминах предметной области, генерация объяснений на естественном языке, адаптация визуализации под уровень подготовки пользователя.

С инженерной точки зрения, локальные интерпретации должны быть встроены в инфраструктуру предсказания: логироваться, версионироваться, сопровождать предсказания в API, включаться в отчёты. Такие системы как Captum (для PyTorch), Eli5, InterpretML, What-If Tool позволяют реализовать такую интеграцию. Они поддерживают генерацию объяснений, их визуализацию, экспорт в форматы документации, сравнение между версиями модели и автоматическое выявление подозрительных выводов [6].

В будущем ожидается развитие гибридных систем интерпретации, сочетающих нейросетевые механизмы, правила и онтологии, а также самоинтерпретируемых моделей, в которых способность к объяснению вшита в архитектуру. Разработка инструментов локальной интерпретации в этом контексте становится не вспомогательной задачей, а центральным элементом архитектуры доверенного ИИ.

Она обеспечивает не только прозрачность и подотчётность, но и возможность для человека контролировать, улучшать и корректировать поведение модели в индивидуальных случаях – что особенно важно в задачах, где на карту поставлены здоровье, права или безопасность людей.

Список литературы:

1. Molnar, C. Interpretable Machine Learning : A Guide for Making Black Box Models Explainable / C. Molnar. – 2nd ed. – [S. l.] : Christoph Molnar, 2022. – URL: https://christophm.github.io/interpretable-ml-book/ (дата обращения: 09.06.2025).

2. Ribeiro, M. T. «Why Should I Trust You?» : Explaining the Predictions of Any Classifier / M. T. Ribeiro, S. Singh, C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – New York : ACM, 2016. – P. 1135–1144. – DOI 10.1145/2939672.2939778.

3. Lundberg, S. M. A Unified Approach to Interpreting Model Predictions / S. M. Lundberg, S.-I. Lee // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.

4. Sundararajan, M. Axiomatic Attribution for Deep Networks / M. Sundararajan, A. Taly, Q. Yan // Proceedings of the 34th International Conference on Machine Learning. – 2017. – Vol. 70. – P. 3319–3328.

5. Selvaraju, R. R. Grad-CAM : Visual Explanations from Deep Networks via Gradient-Based Localization / R. R. Selvaraju, M. Cogswell, A. Das [et al.] // 2017

IEEE International Conference on Computer Vision. – Venice : IEEE, 2017. – P. 618–626. – DOI 10.1109/ICCV.2017.74.

6. Kokhlikyan, N. Captum : A Unified and Generic Model Interpretability Library for PyTorch / N. Kokhlikyan, V. Miglani, M. Martin [et al.] // arXiv. – 2020. – arXiv:2009.07896.