УДК 004.89

МЕТОДЫ ОБЪЯСНЕНИЯ РЕШЕНИЙ ПРЕДОБУЧЕННЫХ МОДЕЛЕЙ ДЛЯ ПОВЫШЕНИЯ ДОВЕРИЯ ПОЛЬЗОВАТЕЛЕЙ

Князева О.С., студент гр. ИУ12-41М, II курс
Московский государственный технический университет имени Н. Э. Баумана,
г. Москва

С увеличением масштабов применения машинного обучения в прикладных задачах, особенно в высокорисковых областях – таких как здравоохранение, финансовое прогнозирование, судебно-правовая экспертиза и управление инфраструктурой – вопрос доверия пользователей к моделям искусственного интеллекта становится центральным. Предобученные модели, в особенности крупные языковые и мультимодальные архитектуры, продемонстрировали впечатляющие результаты, однако их высокая сложность и непрозрачность решений вызывают закономерное недоверие со стороны конечных пользователей, регуляторов и профессиональных экспертов. В ответ на этот вызов развивается отдельное направление исследований – методы объяснения решений предобученных моделей, направленное на повышение интерпретируемости, прозрачности и, как следствие, доверия к системам искусственного интеллекта [1].

Объяснение решений (explainability) подразумевает создание таких представлений о процессе принятия решения моделью, которые доступны для восприятия и оценки человеком. В контексте предобученных моделей, обладающих миллионами или миллиардами параметров, задача объяснения становится особенно сложной, так как сама модель зачастую обучена на данных, к которым пользователь не имеет прямого доступа, и использует внутренние представления, не имеющие очевидного семантического соответствия.

Методы объяснения решений условно можно разделить на пост-хок объяснители (post hoc explainers) и встроенные интерпретируемые механизмы. Первые строят объяснения на уже обученной модели, не вмешиваясь в её архитектуру; вторые требуют модификации или проектирования модели с учётом интерпретируемости [2].

Среди пост-хок подходов наибольшее распространение получили методы локальной интерпретации, такие как LIME (Local Interpretable Model-agnostic Explanations) и SHAP (SHapley Additive exPlanations). Эти алгоритмы оценивают вклад отдельных признаков (input features) в результат конкретного предсказания. Для предобученных моделей они применяются, как правило, на уровне дообученной головы (head) или к выходным векторным

представлениям. Их преимущество – независимость от конкретной архитектуры; их слабость – ограниченность интерпретации на уровне локального линейного приближения и потенциальная нестабильность на высокоразмерных входах.

Для сложных архитектур, таких как трансформеры, особенно языковые модели, получили развитие внимательные методы объяснения, использующие матрицы внимания (attention maps) как основу для визуализации и анализа. Например, в задачах генерации текста можно визуализировать, какие токены входа оказали наибольшее влияние на конкретное слово вывода. Однако внимание не всегда коррелирует с причинностью, что порождает дискуссии в научном сообществе о границах применимости таких методов.

Альтернативой являются градиентные методы, такие как Integrated Gradients, Grad-CAM и DeepLIFT, оценивающие чувствительность выхода модели к изменениям на входе. В контексте предобученных моделей их применение требует доступа к внутренним слоям, что делает их трудноприменимыми в условиях ограниченного API (например, при использовании облачных моделей без доступа к параметрам).

Особую роль играют семантически осмысленные визуализации скрытых представлений. В случае предобученных языковых моделей это может быть редуцированное до двух или трёх измерений пространство эмбеддингов (с помощью t-SNE, UMAP), где пользователь может наблюдать кластеризацию по лексическим или тематическим признакам. Такие подходы позволяют оценивать степень семантической согласованности модели и её обученных представлений.

Наряду с техническими методами объяснения развивается направление человекоцентрированных интерфейсов интерпретации, в которых объяснение подаётся в виде текстовых, графических или интерактивных модальностей. Так, в клинических системах модель не просто выводит диагноз, но сопровождает его объяснением на естественном языке, указывая ключевые симптомы или отклонения в данных пациента. Такой подход снижает когнитивную нагрузку на пользователя и способствует формированию доверия, особенно у непрофильных специалистов.

Важно отметить, что объяснение должно соответствовать контексту применения и профилю пользователя. Для инженера по данным релевантны матрицы внимания и значения градиентов, а для врача или финансового аудитора – простые и логически осмысленные причинно-следственные связи между входом и решением. Это требует разработки адаптивных систем объяснения, способных переключаться между уровнями детализации и типами интерпретации.

На практике объяснения выполняют не только информативную, но и нормативную и юридическую функции. В системах принятия решений,

регулируемых законодательством (например, кредитное скорингование, автоматизированное трудоустройство, медицинская диагностика), объяснение может использоваться как доказательство обоснованности решения и средство защиты от дискриминации. В этом контексте особое значение приобретают формальные методы верифицируемой интерпретируемости, включающие логико-алгебраические проверки соответствия предсказаний модели утверждённым правилам или требованиям справедливости.

Повышение доверия пользователей к предобученным моделям посредством объяснения их решений требует не только разработки технических средств, но и интеграции этих решений в процесс проектирования ИИ-систем. Это включает обязательное тестирование объяснений на понятность, согласованность и влияние на поведение пользователя. Недостаточно лишь «объяснить» модель – важно убедиться, что объяснение действительно улучшает понимание и способствует ответственному принятию решений [3].

В заключение, можно отметить, что методы объяснения решений предобученных моделей являются необходимым компонентом инфраструктуры доверенного ИИ. Их развитие позволяет не только повысить прозрачность и интерпретируемость, но и создать условия для включения человека в цикл принятия решений, что критически важно в условиях высокой неопределённости, юридической ответственности и необходимости социальной приемлемости интеллектуальных систем. Продолжение работы в этом направлении является залогом успешной интеграции сложных ИИ-моделей в прикладные, социально значимые и критически ответственные области [4, 5].

Список литературы:

1. Miller, T. Explanation in Artificial Intelligence : Insights from the Social Sciences / T. Miller // Artificial Intelligence. – 2019. – Vol. 267. – P. 1–38. – DOI 10.1016/j.artint.2018.07.007.

2. Gilpin, L. H. Explaining Explanations : An Overview of Interpretability of Machine Learning / L. H. Gilpin, D. Bau, B. Z. Yuan [et al.] // 2018 IEEE 5th International Conference on Data Science and Advanced Analytics. – Turin : IEEE, 2018. – P. 80–89. – DOI 10.1109/DSAA.2018.00018.

3. Mittelstadt, B. Explaining Explanations in AI / B. Mittelstadt, C. Russell, S. Wachter // Proceedings of the Conference on Fairness, Accountability, and Transparency. – New York : ACM, 2019. – P. 279–288. – DOI 10.1145/3287560.3287574.

4. Gunning, D. DARPA's Explainable Artificial Intelligence (XAI) Program / D. Gunning, D. W. Aha // AI Magazine. – 2019. – Vol. 40, № 2. – P. 44–58. – DOI 10.1609/aimag.v40i2.2850.

5. Biran, O. Explanation and Justification in Machine Learning : A Survey / O. Biran, C. Cotton // IJCAI Workshop on Explainable AI. – 2017.