Локальные объяснения для больших языковых моделей: краткий обзор методов

Е. Н. Волков

Федеральный исследовательский центр «Информатика и управление» Российской Академии Наук

envolkoff@gmail.com

А. Н. Аверкин

Федеральный исследовательский центр «Информатика и управление» Российской Академии Наук

averkin2003@inbox.ru

Аннотация. **Большие языковые модели (БЯМ) в последние несколько лет продемонстрировали выдающиеся показатели в обработке естественного языка. Доля использования БЯМ в цифровых приложениях в различных сферах жизни растёт экспоненциально. Однако, механизмы получения результатов БЯМ, как и любых других типов искусственных нейронных сетей, непрозрачны. Отсутствие прозрачности в принятии решения создает риски для дальнейшего использования технологии. В работе представлен краткий обзор подходов к получению объяснений результатов работы БЯМ. Рассмотрены основные методы объяснительного искусственного интеллекта, применяемые для получения объяснений предсказаний БЯМ.**

Ключевые слова: искусственный интеллект, объяснительный искусственный интеллект, объяснимый искусственный интеллект, большие языковые модели, интерпретируемость, объяснимость, прозрачность, доверие

I. Введение

В эпоху возрастающей цифровизации и ускоренного развития искусственного интеллекта (ИИ), большие языковые модели (БЯМ) становятся неотъемлемой частью множества приложений и сервисов, от автоматизированных систем обслуживания до сложных задач обработки естественного языка (NLP), таких как машинный перевод, автоматическое реферирование и создание контента. Прогресс в этой области позволяет создавать всё более мощные и эффективные модели, способные обрабатывать огромные объёмы данных с беспрецедентной точностью и глубиной понимания текста. Однако вместе с повышением способностей ИИ возникают новые вызовы, в частности, связанные с пониманием и интерпретацией того, как именно большие языковые модели приходят к своим выводам или выбирают определённые варианты ответов. Этот аспект критически важен не только для разработчиков и исследователей, но и для конечных пользователей, требующих прозрачности и объяснимости систем на основе технологий искусственного интеллекта [1].

В контексте данной работы необходимо разграничить ключевые понятия «интерпретируемость» (от англ. Interpretability) и «объяснимость» (от англ. Explainability). Существует множество подходов к их определению. Так, например, Belle в работе [2] определяет первое понятие как пассивную интерпретируемость устройства модели или предсказания на объекте, а второе как активную генерацию объяснений как дополнительных выходных данных для объекта. Иными словами, интерпретируемость относится к внутренней структуре и логике модели, а объяснимость – к способности модели предоставлять понятные пояснения своих решений. Рассматриваемые понятия взаимосвязаны между собой, однако, интерпретируемость более важна для разработчиков и исследователей, а объяснимость – для конечных пользователей, которым необходимо понимать, как модель принимает решения.

Локальные объяснения (от англ. local explanations) направлены на выявление механизмов принятия решений моделями ИИ в конкретных случаях использования, делая акцент на понимании поведения и предсказаний модели в отношении отдельных входных данных или наборов данных [3]. В контексте БЯМ, где каждый запрос и каждый ответ являются уникальными, локальные объяснения могут играть ключевую роль в оценке достоверности, нейтральности и общей релевантности сгенерированных текстов, тем самым обеспечивая необходимый уровень доверия и прозрачности для пользователей.

Необходимость использования методов объяснительного ИИ, и получения локальных объяснений, во многом также исходить из двух основных проблем всех моделей БЯМ таких как «предвзятость» (от англ. bias) и «галлюцинации» (от англ. hallucinations). Предвзятость в БЯМ возникает из-за систематических ошибок, заложенных в данных, используемых для их обучения. Эти данные могут отражать социальные и культурные предубеждения, дискриминационные практики, а также недостаточную представленность или искажение определенных групп населения. Как следствие, модели могут демонстрировать предвзятость в отношении расы, пола, возраста, профессии и других характеристик [4].

Галлюцинации в контексте БЯМ относятся к ситуациям, когда модель генерирует ответ, который выглядит правдоподобным, но на самом деле не соответствует действительности. Обычно такой текст включать в себя вымышленные факты, события, цитаты или ссылки, которые модель ошибочно генерирует вместо того, чтобы вывести сообщение о невозможности вывода ответа. Борьба с галлюцинациями является одним из наиболее актуальных направлений в развитии БЯМ, связанным с фундаментальными основами в

вопросах архитектуры и принципов работы моделей данного класса [5].

Проблемы галлюцинаций и предвзятости особенно заметны в приложениях, где требуется достоверность и фактическая точность, таких как медицинские консультации, финансовые рекомендации или юридические консультации. Использование моделей, подверженных галлюцинациям, в таких критически важных областях может привести к серьезным последствиям для здоровья, благосостояния и безопасности людей. Таким образом, исследование возможностей применения методов получения локального объяснения результатов работы БЯМ является актуальной задачей, решение которой позволяет увеличить уверенность человека-пользователя в технологии тем самым расширив границы её применимости.

II. Классификация методов

В исследовании [6] авторы предлагают классификацию методов локального объяснения, основанную на двух ключевых аспектах: способ генерации объяснений и степень изменения входных данных. Способ генерации объяснений – методы могут быть основаны на градиентах, на изменении входных данных, или на внутренних представлениях модели. Степень изменения входных данных – методы могут оперировать с исходными входными данными или с их модификациями. В тоже время, в работе [7] предлагается другая классификация, сфокусированная на: типе и уровне объяснения. Тип объяснений – методы могут предоставлять атрибуции, интерпретации, или генерировать контрфактические примеры. Уровень объяснений – методы могут объяснять на уровне токенов, предложений, или на более высоком семантическом уровне. Таким образом, первая работа делает акцент на технических аспектах методов объяснения, в то время как вторая уделяет больше внимания типам и уровням получаемых объяснений, что отражает различные фокусы и подходы авторов к систематизации методов локального объяснения для БЯМ.

В свою очередь, в данной работе, предлагается разделить методы локального объяснения на методы, основанные на атрибуции признаков и на методы, основанные на особенностях архитектуры БЯМ (рис. 1).

Рис. 1. Группы методов объяснения БЯМ (авторское разделение)

III. Объяснения на основе атрибуции признаков

Наиболее часто используемыми, из всех групп методов объяснительного ИИ, являются методы, основанные на вычислении атрибуции признаков, то есть значимости каждого входного токена для прогноза модели. В контексте подходов к объяснимости БЯМ суть работы данной группы методов можно формально представить следующим образом: при заданном входном запросе x, являющемся некоторой упорядоченной последовательностью из n токенов $\{x_1, x_2, x_3, ..., x_n\}$ предварительно обученная БЯМ f получает предсказание $f(x)$, при этом, методы атрибуции присваивают каждому токену x_i оценку значимости $R(x_i)$, отражающую вклад конкретного токена в полученное предсказание модели, что также относится к пост-фактум объяснениям (от англ. post-hoc explanations). К данной группе относится большое количество методов объяснения, однако не все они могут быть применены для работы БЯМ, рассмотрим лишь некоторые из них.

A. Суррогатные методы

Суррогатные методы объяснительного ИИ – это подход, основанный на построении более простой, интерпретируемой модели (линейные модели, деревья решений) для объяснения исходной сложной. Наиболее популярными методами из этой группы являются LIME SHAP. Однако, с усложнением интерпретируемой модели вычисления на основе данных алгоритмов становятся неоправданно затратными [8].

Несмотря на сложность реализации подобных методов, SHAP (SHapley Additive exPlanations) остается популярным и широко используемым. Для адаптации SHAP к трансформер-основанным языковым моделям были предложены подходы, такие как TransSHAP [9,10]. TransSHAP в основном сосредоточен на адаптации SHAP к вводу текста на уровне токенов и предоставлении последовательных визуализационных объяснений, которые хорошо подходят для понимания того, БЯМ делают предсказания.

B. Градиентные методы

Integrated Gradients (IG) был предложен Sundararajan M. в 2017 году [11]. Основная идея IG заключается в том, чтобы вычислить интеграл градиентов вдоль прямой линии между базовым и фактическим входными значениями, представляющим собой нейтральное и фактическое значения признаков соответственно. Для этого метода требуется задать базовое входное значение, которое служит отправной точкой для вычисления интеграла. IG вычисляет вклад каждого признака, интегрируя градиенты модели по отношению к каждому признаку вдоль прямой линии от базового входного значения до фактического входного значения.

$$IG_i(x) ::= (x_i - x_i^{'}) \times \int_{\alpha=0}^{1} \frac{\partial F(x^{'} + \alpha \times (x - x^{'}))}{\partial x_i} d\alpha$$

Discretized Integrated Gradients (DIG) был предложен Sanyal S. в 2021 году и является модификацией IG,

которая использует дискретизацию для вычисления интеграла градиентов. Вместо вычисления интеграла вдоль прямой линии, DIG делит пространство входных данных на дискретные сегменты и вычисляет интеграл градиентов для каждого сегмента. Затем эти значения интегрируются для получения общего вклада каждого признака. DIG может быть более точным, чем IG, так как он использует дискретизацию для вычисления интеграла [12].

$$DIG_i(x) = \int_{x_i^k = x_i'}^{x_i} \frac{\partial F(x^k)}{\partial x_i} dx_i^k$$

Sequential Integrated Gradients (SIG) был предложен Enguehard J. в 2021 году как переосмысление стандартного IG и его модификации DIG. SIG вычисляет вклад каждого признака последовательно, используя интеграл градиентов вдоль пути от базового входного значения до фактического входного значения. В отличие от IG, SIG не требует задавать базовое входное значение, так как он использует последовательность изменений признаков в качестве отправной точки для вычисления интеграла. SIG вычисляет вклад каждого признака, интегрируя градиенты модели по отношению к каждому признаку вдоль пути, определенного последовательностью изменений признаков.

$$SIG_{ij}(\mathbf{x}) := (x_{ij} - \bar{x}_{ij}) \times \int_0^1 \frac{\partial \mathbf{F}(\bar{\mathbf{x}}^i + \alpha \times (\mathbf{x} - \bar{\mathbf{x}}^i))}{\partial x_{ij}} d\alpha$$

Основное преимущество SIG заключается в том, что он учитывает взаимодействия между признаками, так как он вычисляет вклад каждого признака последовательно, а не независимо, как это делает IG. Кроме того, SIG не требует задавать базовое входное значение, что может быть неочевидным и сложным для некоторых типов данных.

Однако SIG также имеет некоторые ограничения. Во-первых, он предполагает, что поведение модели линейно вдоль пути от базового входного значения до фактического входного значения, что не всегда верно. Во-вторых, он может быть чувствителен к порядку изменений признаков, что может влиять на результаты объяснения. В-третьих, он может быть вычислительно более дорогим, чем IG и DIG, так как требует вычисления интеграла градиентов для каждого шага [13].

IV. Объяснения основанные на структуре модели

Поскольку в основе всех БЯМ лежит архитектура ИНС типа «трансформер», то можно выделить ряд методов, использующих элементы данной архитектуры как основу для вывода локального объяснения. Выделяют два пути получения объяснений: через блоки внимания (attention) или через блоки многослойного перцептрона (MLP). Формально данную группу методов объяснения можно представить в виде x_i^l – последовательности для каждого поступившего токена i по блоку l:

$$x_i^l = x_i^{l-1} + a_i^l + m_i^l,$$

где a_i^l и m_i^l – выходы l - блока внимания и многослойного перцептрона, соответственно. Представленная формализация связана с наличием как блока внимания, так и MLP – блока в любой архитектуре из данного семейства.

A. *Объяснение через блоки внимания*

Использования блоков внимания для получения объяснений является одной из стандартных практик, используемых в ИНС архитектуры «трансформер». Метод имеет высокую эффективность несмотря на тип, входящий данных или вариацию архитектуры блока внимания – головы внимания трансформера. Механизм внимания позволяет модели фокусировать свое «внимание» на наиболее релевантных частях входных данных при генерации выходных результатов (рис. 2). Блоки внимания в нейронной сети модели отвечают за вычисление весов внимания, которые показывают степень важности каждого входного элемента.

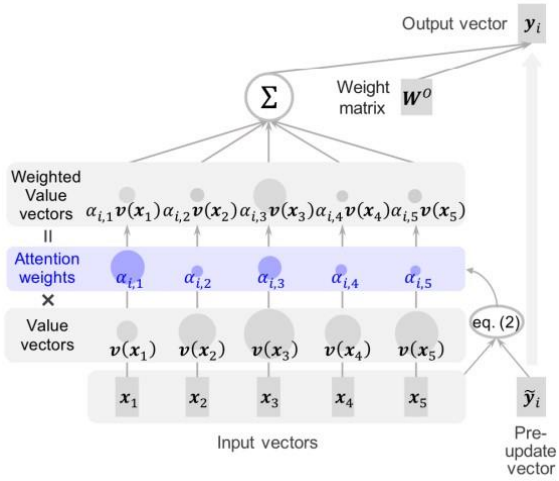

Рис. 2. Визуальное представление механизма внимания [14]

Анализ весов внимания дает возможность получить объяснения в виде атрибуций – то есть, определить, какие токены оказали наибольшее влияние на выходной результат модели. Данный подход к получению объяснения считается одним из наиболее естественных и интуитивно понятных для БЯМ, поскольку он напрямую связан с их архитектурой и механизмом функционирования. Он дает возможность получать локальные, токен-уровневые объяснения, что полезно для многих практических приложений [7].

B. *Объяснение через блок многослойного перцептрона*

В архитектуре БЯМ, как правило, используются блоки MLP, которые трансформируют входные представления в более абстрактные, семантически богатые выходные представления. Эти промежуточные представления в блоках MLP отражают различные уровни семантической иерархии, закодированной в модели (рис. 3). Анализ активаций нейронов в слоях MLP, а также изучение влияния изменения входных данных на эти активации, позволяет получать объяснения на более высоком семантическом уровне. Такие объяснения могут быть в форме интерпретаций, которые связывают входные и выходные данные с

промежуточными семантическими концептами, закодированными в модели. Этот подход дает возможность получать более глубокие, холистические объяснения, выходящие за рамки локальных, токен-уровневых атрибуций. Позволяет лучше понять, как БЯМ формируют свои внутренние представления и как они соотносятся с высокоуровневыми семантическими понятиями [7].

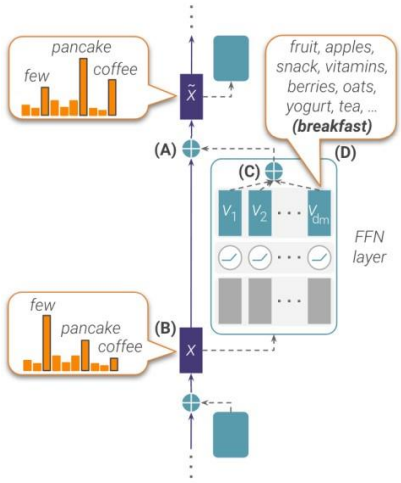

Рис. 3. Визуальное представление многослойного перцептрона [15]

На рис. 4 представлен пример визуализации объяснения БЯМ, полученного с помощью анализа MLP-блока (a) и с помощью извлечения признаков из механизма внимания (b). В качестве средства визуализации объяснения выбрана тепловая карты со шкалой градации. Интенсивность каждой ячейки сетки представляет собой среднее косвенное причинное влияние скрытого состояния на выражение фактической ассоциации [16].

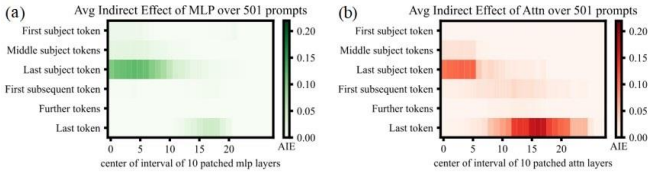

Рис. 4. Примеры визуализации объяснения БЯМ: a) объяснение через MLP – блок, b) объяснение через блок внимания [16]

V. Заключение

Несмотря на значительный прогресс в разработке и применении методов локальных объяснений для БЯМ, существуют важные проблемы и вызовы, которые требуют дальнейшего анализа и решения. Одна из ключевых сложностей заключается в балансе между достаточной информативностью объяснений и их интерпретируемостью для конечных пользователей. Кроме того, важным аспектом является универсальность методов объяснений, их способность адаптироваться к различным типам и конфигурациям языковых моделей, а также их эффективность в рамках разнообразных задач обработки естественного языка. Локальные объяснения для БЯМ представляют собой критически важный инструмент, необходимый для обеспечения прозрачности, доверия и ответственности в области искусственного интеллекта.

Список литературы

[1] Liao Q.V., Vaughan J.W. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap //Harvard Data Science Review. 2024. DOI: 10.1162/99608f92.8036d03b.

[2] Belle V., Papantonis I. Principles and practice of explainable machine learning //Frontiers in big Data. 2021. Vol. 4. P. 688969. DOI: 10.3389/fdata.2021.688969.

[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. DOI: 10.1016/j.inffus.2019.12.012.

[4] Gallegos I.O., Rossi R.A., Barrow J. et al. Bias and fairness in large language models: A survey //arXiv preprint arXiv:2309.00770. 2023.

[5] Xu Z., Jain S., Kankanhalli M. Hallucination is inevitable: An innate limitation of large language models //arXiv preprint arXiv:2401.11817. 2024.

[6] Zhao H., Chen H., Yang F. et al. Explainability for large language models: A survey //ACM Transactions on Intelligent Systems and Technology. 2024. Vol. 15. №. 2. P. 1-38. DOI: 10.1145/3639372

[7] Luo H., Specia L. From Understanding to Utilization: A Survey on Explainability for Large Language Models //arXiv preprint arXiv:2401.12874. 2024.

[8] Аверкин А.Н., Ярушев С.А. Объяснительный искусственный интеллект в моделях поддержки принятия решений для Здравоохранения 5.0 //Компьютерные инструменты в образовании. 2023. №. 2. С. 41-61. DOI: 10.32603/2071-2340-2023-2-41-61.

[9] Kokalj E., Skrlj B., Lavrac N. et al. BERT meets shapley: Extending SHAP explanations to transformer-based classifiers //Proceedings of the EACL hack a shop on news media content analysis and automated report generation. 2021. P. 16-21.

[10] Chen H., Covert I.C., Lundberg S.M. et al. Algorithms to estimate Shapley value feature attributions //Nature Machine Intelligence. 2023. Vol. 5. №. 6. P. 590-601. DOI: 10.1038/s42256-023-00657-x

[11] Sundararajan M., Taly A., Yan Q. Axiomatic attribution for deep networks //International conference on machine learning. PMLR, 2017. P. 3319-3328.

[12] Sanyal S., Ren X. Discretized Integrated Gradients for Explaining Language Models //Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. P. 10285-10299. DOI: 10.18653/v1/2021.emnlp-main.805.

[13] Enguehard J. Sequential Integrated Gradients: a simple but effective method for explaining language models //Findings of the Association for Computational Linguistics: ACL 2023. 2023. P. 7555-7565. DOI: 10.18653/v1/2023.findings-acl.477.

[14] Kobayashi G., Kuribayashi T., Yokoi S. et al. Attention is Not Only a Weight: Analyzing Transformers with Vector Norms //Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020.

[15] Geva M., Caciularu A., Wang K. et al. Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space //Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022.

[16] Meng K., Bau D., Andonian A. et al. Locating and editing factual associations in GPT // Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 17359-17372.