ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ

УДК 004.054 Научная статья
DOI: 10.35330/1991-6639-2024-26-4-54-61
EDN: KZKDOT

Метод оценки степени доверия к само-объяснениям GPT-моделей

А. Н. Лукьянов, А. М. Трамова

Российский экономический университет им. Г. В. Плеханова
117997, Россия, Москва, Стремянный переулок, 36

Аннотация. Со стремительным ростом использования генеративных нейросетевых моделей для решения практических задач все более остро встает проблема объяснения их решений. По мере ввода решений на основе нейросетей в медицинскую практику, государственное управление и сферу обороны требования к таким системам в плане их интерпретируемости однозначно будут расти. В данной работе предложен метод проверки достоверности само-объяснений, которые модели дают постфактум, посредством сравнения распределения внимания модели во время генерации ответа и его объяснения. Авторами предложены и разработаны методы для численной оценки степени достоверности ответов генеративных предобученных трансформеров. Предлагается использовать расхождение Кульбака – Лейблера над распределениями внимания модели во время выдачи ответа и следующего за этим объяснения. Также предлагается вычислять отношение внимания модели между изначальным запросом и сгенерированным объяснением с целью понять, насколько само-объяснение было обусловлено собственным ответом. Для получения данных величин предлагается алгоритм для рекурсивного вычисления внимания модели по шагам генерации. В результате исследования была продемонстрирована работа предложенных методов, найдены значения метрик, соответствующие корректным и некорректным объяснениям и ответам. Был проведен анализ существующих в настоящий момент методов определения достоверности ответов генеративных моделей, причем подавляющее большинство из них сложно интерпретируемые обычным пользователем. В связи с этим мы выдвинули собственные методы, проверив их на наиболее широко используемых на момент написания генеративных моделях, находящихся в открытом доступе. В результате мы получили типичные значения для предложенных метрик, алгоритм их вычисления и визуализации.

Ключевые слова: нейронные сети, метрики, языковые модели, интерпретируемость, LLM, GPT, XAI

Поступила 24.06.2024, одобрена после рецензирования 01.08.2024, принята к публикации 07.08.2024

Для цитирования. Лукьянов А. Н., Трамова А. М. Метод оценки степени доверия к само-объяснениям GPT-моделей // Известия Кабардино-Балкарского научного центра РАН. 2024. Т. 26. № 4. С. 54–61. DOI: 10.35330/1991-6639-2024-26-4-54-61

MSC: 68T09 Original article

A method for assessing the degree of confidence in the self-explanations of GPT models

A.N. Lukyanov, A.M. Tramova

Plekhanov Russian University of Economics
117997, Russia, Moscow, 36 Stremyanny Lane

Abstract. With the rapid growth in the use of generative neural network models for practical tasks, the problem of explaining their decisions is becoming increasingly acute. As neural network-based solutions are being introduced into medical practice, government administration, and defense, the demands for interpretability of such systems will undoubtedly increase. In this study, we aim to propose a method for verifying the reliability of self-explanations provided by models post factum by comparing the attention distribution of the model during the generation of the response and its explanation. The authors propose and develop methods for numerical evaluation of answers reliability provided by generative pre-trained transformers. It is proposed to use the Kullback-Leibler divergence over the attention distributions of the model during the issuance of the response and the subsequent explanation. Additionally, it is proposed to compute the ratio of the model's attention between the original query and the generated explanation to understand how much the self-explanation was influenced by its own response. An algorithm for recursively computing the model's attention across the generation steps is proposed to obtain these values. The study demonstrated the effectiveness of the proposed methods, identifying metric values corresponding to correct and incorrect explanations and responses. We analyzed the currently existing methods for determining the reliability of generative model responses, noting that the overwhelming majority of them are challenging for an ordinary user to interpret. In this regard, we proposed our own methods, testing them on the most widely used generative models available at the time of writing. As a result, we obtained typical values for the proposed metrics, an algorithm for their computation, and visualization.

Keywords*:* neural networks, metrics, language models, interpretability, GPT, LLM, XAI

Submitted 24.06.2024, approved after reviewing 01.08.2024, accepted for publication 07.08.2024

For citation. Lukyanov A.N., Tramova A.M. A method for assessing the degree of confidence in the self-explanations of GPT models. *News of the Kabardino-Balkarian Scientific Center of RAS.* 2024. Vol. 26. No. 4. Pp. 54–61. DOI: 10.35330/1991-6639-2024-26-4-54-61

Введение

За последние 7 лет трансформеры перевернули сферу глубокого обучения [1, 2]. Однако вопрос интерпретируемости нейронных сетей стоит довольно давно. Главным предметом исследований в сфере объяснительного искусственного интеллекта (XAI) долгое время являлись сверточные нейронные сети (CNN). В частности, довольно долгое время в сфере компьютерного зрения (CV) главным механизмом объяснения был метод CAM (class activation maps) [3]. В связи с совместным использованием глобального пулинга и линейного классификатора эти две операции можно было представить в виде единой матрицы коэффициентов, которая при совмещении с конечными картами активации давала метод оценки значимости зон изображения по отношению к выводу сети. Важным моментом является то, что данный подход не привязан только к сфере компьютерного зрения, но может быть применен везде, где результаты работы модуля получаются путем линейной комбинации входного вектора с коэффициентами.

В последующие годы появились новые методы объяснения работы нейронных сетей во всевозможных модальностях. Так, метод LIME [4] позволяет обучить меньшую нейросеть и аппроксимировать изучаемую, он был применен во многих работах, включая и интересующую нас обработку естественных языков (NLP). То же самое касается и метода SHAP [5].

Большой революцией в глубоком обучении стало изобретение, распространение и улучшение сетей на архитектуре трансформера. По природе работы механизма внимания эти сети уже включают в себя некую степень интерпретируемости. Так, особенно сильно выделяется крайне глубоко проработанный инструмент BertViz [6]. И хоть из-за

присутствия многослойного перцептрона в трансформерах объяснения на основании карт внимания не являются на 100% достоверными, они наименее инвазивные и требовательные к архитектуре, что крайне важно для объяснения уже существующих моделей. Объяснения такого рода принадлежат к сфере механистических объяснений [7].

Постановка проблемы галлюцинаций и обмана генеративных моделей, цели и задачи

На данный момент галлюцинации являются одной из ключевых проблем в области объяснимого искусственного интеллекта. Они происходят тогда, когда нейронная сеть «выдумывает» факты или связи, которых на самом деле не существует. Одним из методов решения данной проблемы является RAG [8] (retrieval augmented generation). Однако инвестируя ресурсы в непараметрическую память, мы теряем ресурсы, которые могли бы быть потрачены на саму модель; более того, чтобы RAG показывал хорошие результаты, требуется довольно дорогое обучение второй модели, которая будет находить связанные с запросом документы.

Еще большими угрозами являются сокрытие и ложь. Так как нейросетевые модели – это черные ящики, то мы никогда не можем быть полностью уверены, что модель отвечает на вопросы честно или что ее объяснения достоверны. Сокрытие частично решается методом CoT [9] (chain of thought prompting), который, хоть и был введен для повышения эффективности работы генеративных моделей, также позволяет нам заглянуть в их мыслительный процесс. Однако даже в этом случае уже было показано, что GPT (generative pretrained transformer) модели могут делегировать часть обработки информации на малоинформативные части текста [10] (например, пунктуацию), так что нам необходимо отслеживать, куда именно смотрит модель.

Метод исследования. Описание предложенного решения

Как первый шаг на пути к ИИ, ответы которого мы можем проверять на достоверность, мы предлагаем использовать attention rollout [11] для получения карт внимания модели во время генерации своих ответов и само-объяснения. В отличие от предыдущих имплементаций мы собираемся отслеживать внимание для генерации каждого токена, поэтому методу мы даем название AROT (attention rollout over time). Алгоритмы для получения карт внимания приведены ниже. На основе данных карт мы вводим две численные меры – отношение контекста к ответу во время генерации само-объяснения и отношение распределения внимания до и после генерации ответа. Первая величина позволяет нам примерно понять, не проявляла ли модель избыточное внимание к своему же ответу или к не связанным с задачей элементам запроса. Вторая величина показывает, был ли мыслительный процесс модели во время генерации объяснения схож с тем, который она испытывала во время генерации ответа. Первая величина является пропорцией сумм коэффициентов внимания до и после генерации ответа. Вторая вычисляется с помощью расхождения Кульбака – Лейблера между распределениями внимания во время объяснения относительно распределения во время составления изначального ответа. Так как расхождение Кульбака – Лейблера не ограничено сверху, то пороговые значения будут приближенно найдены эвристически.

Attention rollout over time для авторегрессивной сети с проходом по каждому сгенерированному токену (базовый алгоритм, не адаптирован для работы во время генерации, адаптированная версия ниже)

Вход: список A, содержащий *t* тензоров размером (*l*, *h*, *p*, *p*), где $p = ctx + i$
ctx – длина контекста
t – число сгенерированных токенов
i – номер генерируемого токена от *0* до *t*
l – число слоев модели
h – число голов внимания
p – длина последовательности на шаге *i*

Листинг 1. Алгоритм вычисления внимания модели на протяжении шагов генерации.

```
  def TimeRollout(A, ctx):
     past = []

     for i, token in enumerate(A):
       attention = torch.zeros(ctx+i)
       attention[-1] = 1

       for layer in reversed(token):
           map = layer.mean(0)
           # из-за остаточной связи мы предлагаем суммировать
результат внимания на шаге t с вниманием на шаге (t-1)
           attention += attention @ map

       out = attention[:ctx]
       for j in range(0, i):
           out += attention[ctx-1+i] * past[j]
       out /= out.sum()

       past.append(out)

     return past[-1]
```

Листинг 2. Совместная генерация и вычисление AROT

```
  def GenAndAtt(text):
     amap = [1] * len(text)
     ctx = text.index(1217)

     for i in range(100):
       attention = torch.zeros(len(text))
       attention[-1] = 1

       inp = {'input_ids': torch.tensor([text]), 'atten-
tion_mask': torch.tensor([amap+[1]*i])}
         amap.append(1)

      out = model(**inp)
      pred = out.logits[0, -1]
```

```
    layerAttentions = torch.cat(out.attentions, 0)

    for layer in reversed(layerAttentions):
        map = layer.mean(0)

        attention += attention @ map

     out = attention[ctx:ctx+seqLen]

     for j in range(0, i):
         out += attention[ctx-1+i] * past[j]
     out /= out.sum()

     past.append(out)

     id = pred.argmax().item()
     text.append(id)
   print(tokenizer.convert_ids_to_tokens(id).replace('Ġ', ''))
     if id == 128256:
         break

 return text, past[-1]
```

Расхождение Кульбака – Лейблера

$$D(p' \,||\, p) = \sum_i p(x_i) \cdot log(\frac{p'(x_i)}{p(x_i)})$$

Листинг 3. Расхождение Кульбака – Лейблера

```
def KLDivergence (P, Q):
     return (P * (P.log() - Q.log())).sum()
```

Эксперимент и результаты

Для проведения эксперимента были взяты open source модели Llama 3 8B [12] – Dolphin 2.9, Mistral 7-B [13], Zephyr 7-B-β [14]. Моделям были заданы несколько вопросов и для ответов вычислены предложенные метрики. Для получения и представления информации о распределении внимания, а также вычисления метрик был написан свой программный код, взаимодействующий с API huggingface. Примеры полученных результатов с dolphin-2.9 приложены ниже:

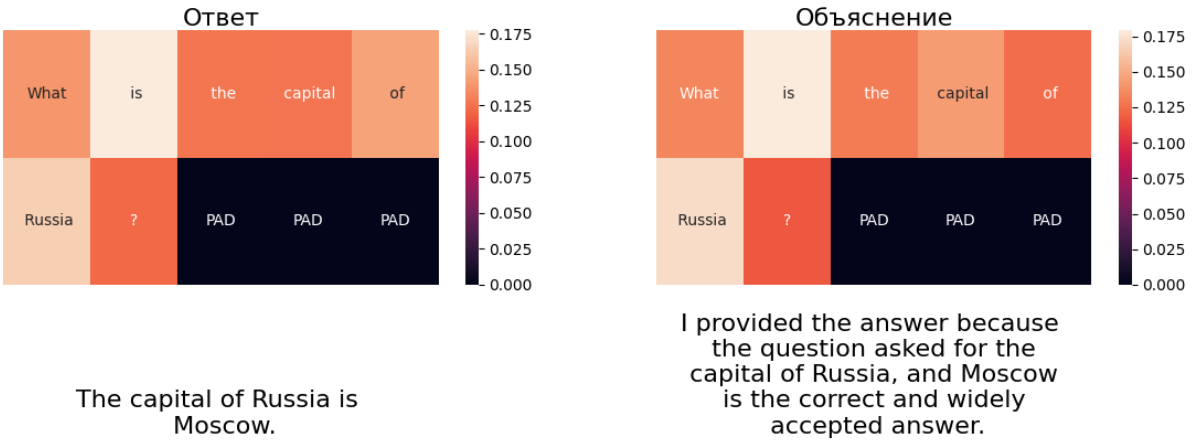
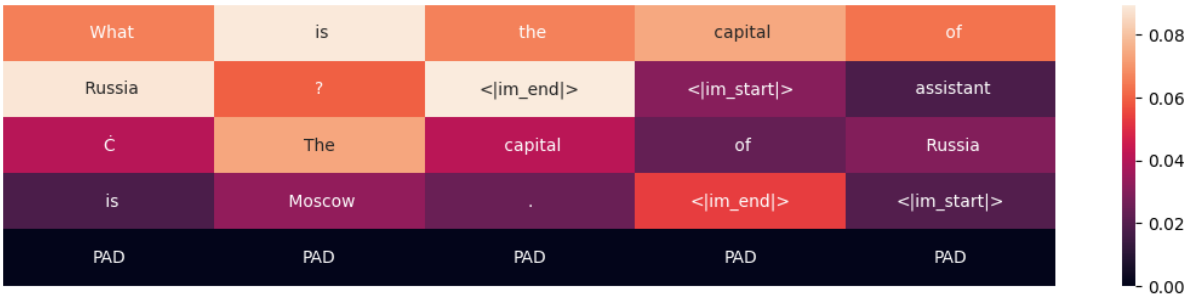


Рис. 1. *Распределение внимания к контексту для ответа и объяснения с расхождением КЛ между ними*

Fig. 1. *Distribution of attention to context for answer and explanation and the divergence of CL between them*



Отношение между вниманием к контексту и ответу =1.355078935623169

Рис. 2. *Распределение внимания к контексту и ответу для объяснения с коэффициентом отношения внимания*

Fig. 2. *Distribution of attention to context and response for explanation with attention ratio coefficient*

Каждой модели задавалось 10 идентичных вопросов. На основании проведенных экспериментов было замечено, что когда модель выдает правильный ответ, расхождение КЛ принимает малые числа (<0,1), а отношение внимания к контексту числа больше единицы (полученные значения – среднее по всем моделям и запросам). Наоборот, при ошибке расхождение КЛ заметно выше (>1), а отношение внимания ниже единицы вплоть до 0,26.

Выводы

Введено и протестировано два метода для оценки достоверности само-объяснений модели и ее уверенности в ответах. Данное исследование послужит фундаментом для будущих исследований в сфере объяснительного интеллекта, а это в свою очередь даст нам больше гарантий безопасности используемых нами генеративных нейросетей. Несмотря на применение англоязычных моделей в проведенной работе, следует сделать вывод, что она приведет к повышению интереса к разработке русскоязычных аналогов. Также в связи с распространением моделей на основе Mamba [15], их экономичностью по сравнению с

моделями на основе трансформеров и возможностью интерпретации с помощью механизма внимания [16] они являются идеальным кандидатом для дальнейшего изучения измерения степени доверия к само-объяснениям генеративных моделей.

REFERENCES

1. Vaswani A., Shazeer N., Parmar N. et al. Attention is all you need. *Advances in neural information processing systems.* 2017. No. 3. URL: https://arxiv.org/abs/1706.03762

2. Dosovitskiy A., Beyer L., Kolesnikov A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2020. URL: https://arxiv.org/abs/2010.11929

3. Selvaraju R.R., Cogswell M., Das A. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. URL: https://arxiv.org/abs/1610.02391

4. Ribeiro M.T., Singh S., Guestrin C. "Why should I trust you?": Explaining the Predictions of Any Classifier. URL: https://arxiv.org/abs/1602.04938

5. Lundberg S., Lee S. A unified approach to interpreting model predictions. URL: https://arxiv.org/abs/1705.07874

6. Jesse Vig. Visualizing attention in transformer-based language representation models. URL: https://arxiv.org/abs/1904.02679

7. Bereska L., Gavves E. Mechanistic interpretability for AI Safety – A review. URL: https://arxiv.org/abs/2404.14082

8. Lewis P., Perez E., Piktus A. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. URL: https://arxiv.org/abs/2005.11401

9. Wei J., Wang X., Schuurmans D. et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. URL: https://arxiv.org/abs/2201.11903

10. Pfau J., Merrill W., Bowman S.R. Let's think dot by dot: Hidden computation in transformer language models. URL: https://arxiv.org/abs/2404.15758

11. Abnar S., Zuidema W. Quantifying attention flow in transformers. URL: https://arxiv.org/abs/2005.00928

12. Touvron H., Lavril T., Izacard G. et al. LLaMA: Open and efficient foundation language models. URL: https://arxiv.org/abs/2302.13971

13. Jiang A.Q., Sablayrolles A., Mensch A. et al. Mistral 7B. URL: https://arxiv.org/abs/2310.06825

14. Tunstall L., Beeching E., Lambert N. et al. Zephyr: Direct distillation of LM alignment. URL: https://arxiv.org/abs/2310.16944

15. Gu A., Dao T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. URL: https://arxiv.org/abs/2312.00752

16. Ali A., Zimerman I., Wolf L. The Hidden Attention of Mamba Models. URL: https://arxiv.org/abs/2403.01590

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Финансирование. Исследование проведено без спонсорской поддержки.

Funding. The study was performed without external funding.

Информация об авторах

Лукьянов Андрей Николаевич, студент, лаборант-исследователь, Центр перспективных исследований в искусственном интеллекте, Российский экономический университет им. Г. В. Плеханова;
117997, Россия, Москва, Стремянный переулок, 36;
andreylukianovai@gmail.com

Трамова Азиза Мухамадияевна, д-р экон. наук, профессор, профессор кафедры информатики, Российский экономический университет им. Г. В. Плеханова;
117997, Россия, Москва, Стремянный переулок, 36;
Tramova.AM@rea.ru, ORCID: https://orcid.org/0000-0002-4089-6580, SPIN-код: 8583-3592

Information about the authors

Andrey N. Lukyanov, Student, Research Assistant, Center for Advanced Studies in Artificial Intelligence, Plekhanov Russian University of Economics,
117997, Russia, Moscow, 36 Stremyanny Lane;
andreylukianovai@gmail.com

Aziza M. Tramova, Doctor of Economic Sciences, Professor, Professor of the Department of Informatics, Plekhanov Russian University of Economics;
117997, Russia, Moscow, 36 Stremyanny Lane;
Tramova.AM@rea.ru, ORCID: https://orcid.org/0000-0002-4089-6580, SPIN-code: 8583-3592