ПРИНЦИП РАБОТЫ И ПРОБЛЕМЫ «GENERATIVE PRE-TRAINED TRANSFORMER ARTIFICIAL INTELLIGENCE»

Зонова Д.Ю.

Зонова Дарья Юрьевна - студент бакалавриата,
кафедра САПР, факультет компьютерных технологий и информатики,
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»,
г. Санкт-Петербург

Аннотация: *генеративный, предварительно обученный, искусственный интеллект (Chat- GPT) завоевал значительный интерес и внимание с момента своего запуска в ноябре 2022 года. Он показал впечатляющие результаты в различных областях, однако проблемы, связанные с предубеждениями и недоверием, сохраняются. В этой работе рассматривается архитектура и принцип работы генеративного ИИ. Главной целью является: раскрыть потенциал ChatGPT в решении реальных задач, а также определить потенциальные направления будущих исследований ChatGPT, предлагая решения текущих проблем. Полностью используя функционал ChatGPT, мы можем раскрыть его возможности в различных областях, что приведет к прогрессу в области разговорного искусственного интеллекта и модернизации общества.*
Ключевые слова: *ChatGPT, архитектурf «Transformer», кодер-декодер, BPE (Byte Pair Encoding), авторские права, обучающие данные.*

THE PRINCIPLE OF OPERATION AND PROBLEMS OF "GENERATIVE PRE-TRAINED TRANSFORMER ARTIFICIAL INTELLIGENCE"

Zonova D.Yu.

Zonova Darya Yuryevna, undergraduate student,
DEPARTMENT OF CAD, FACULTY OF COMPUTER TECHNOLOGY AND INFORMATICS,
ST. PETERSBURG STATE ELECTROTECHNICAL UNIVERSITY "LETI",
ST. PETERSBURG

Abstract: *generative, pre-trained, artificial intelligence (Chat-GPT) has gained considerable interest and attention since its launch in November 2022. He has shown impressive results in various fields, but the problems associated with prejudice and distrust persist. This paper examines the architecture and the principle of operation of generative AI. The main goal is: to unlock the potential of ChatGPT in solving real problems, as well as to identify potential areas for future ChatGPT research, offering solutions to current problems. Fully using the ChatGPT functionality, we can unlock its capabilities in various fields, which will lead to progress in the field of conversational artificial intelligence and modernization of society.*
Keywords: *ChatGPT, architecture "Transformer", encoder-decoder, BPE (Byte Pair Encoding), copyright, training data.*

Архитектура ChatGPT

ChatGPT, разработанный OpenAI, представляет собой языковую модель, которая позволяет создавать разговорные системы искусственного интеллекта, способные понимать вводимые человеком данные и давать осмысленные ответы на них. Функционируя как чат-бот с поддержкой искусственного интеллекта, он использует алгоритмы для обработки пользовательских вводимых данных и генерации соответствующих ответов. ChatGPT обладает способностью генерировать новые ответы или использовать уже существующие. Чтобы улучшить понимание пользовательских запросов и генерировать точные ответы, ChatGPT постоянно совершенствуется с использованием обучения с подкреплением, машинного обучения, и методов обработки естественного языка.

Чтобы постоянно повышать надежность и точность модели, ChatGPT включает в себя обучение с подкреплением на основе обратной связи с человеком, что позволяет ему изучать и понимать предпочтения человека посредством расширенных диалогов. Кроме того, исследователи активно изучают новые технологии для дальнейшего повышения его производительности. ChatGPT использует архитектуру «Transformer», состоящую из уровней кодер-декодер, которые совместно обрабатывают и генерируют текст на естественном языке. Кодировщик получает информацию, собранную в векторную последовательность с позиционной информацией, а декодировщик в свою очередь получает на вход часть этой последовательности и выход кодировщика. Архитектура ChatGPT включает в себя несколько жизненно важных компонентов, таких как токенизатор, который делит необработанный текст на более мелкие блоки, называемые токенами, для облегчения обработки. Токенизация хороший способ каждому слову задать свой токен (число), для дальнейшей обработки. Затем компонент встраивания входных данных преобразует эти токены в многомерные векторные представления. OpenAI использовали byte-level BPE токенизацию. Эта модификация BPE работает не с текстом и не задает каждому символу свой номер, а напрямую с его байтовым

представлением. Использование такого метода позволило сжать словарь до всего лишь 50 тысяч токенов при том, что с его помощью всё ещё можно выразить любое слово на любом языке мира, а так же эмодзи.

Архитектура «Transformer» состоит из двух основных компонентов: кодера и декодера. Кодировщик обрабатывает входной текст иерархически, создавая представления на разных уровнях абстракции. С другой стороны, декодер генерирует выходной текст по одному токену за раз, используя входные представления, сгенерированные кодировщиком. Существенной особенностью архитектуры «Transformer» является механизм внимания, который позволяет модели выборочно фокусироваться на различных частях входного текста при генерации выходных данных. Этот механизм повышает способность модели улавливать релевантную информацию и выдавать согласованные выходные данные. Выходной уровень отвечает за преобразование многомерного векторного представления выходного текста в распределение вероятностей по словарю возможных выходных токенов. Это позволяет ChatGPT генерировать высококачественный и связный текст на естественном языке. Архитектура ChatGPT позволяет ему преуспевать в различных задачах, включая чат-ботов, языковой перевод и заполнение текста, генерируя точные и содержательные ответы.

Несмотря на свою популярность и полезность, ChatGPT вызвал обеспокоенность у исследователей и практиков из-за его способности генерировать контент, который, хотя и кажется разумным, не обладает фактической точностью. Эта проблема может привести к появлению противоречащих действительности или бессмысленных ответов, что создает серьезную угрозу надежности онлайн-контента. Кроме того, ложные сообщения, генерируемые ChatGPT, могут быть легко ошибочно приняты за законные, особенно лицами, которые не знакомы с рассматриваемой темой. Исследователи изучают и выделяют потенциальный вред, связанный с ChatGPT, включая распространение стереотипов, предвзятых ответов и ложной информации. Также были подняты этические проблемы в отношении использования ChatGPT, особенно когда он используется для создания манипулируемого контента, способствующего дезинформации и подстрекает к насилию, потенциально причиняя вред как на индивидуальном, так и на организационном уровнях. Более того, существуют опасения, что контент, созданный с помощью ChatGPT, может нарушать авторские права и интеллектуальную собственность. Кроме того, нельзя игнорировать этические соображения, касающиеся использования этого инструмента для академического и научного письма.

Первоначальная версия ChatGPT, называемая GPT-1, была оснащена 117 миллионами параметров и прошла обучение на значительном массиве текстовых данных. Последующие версии, такие как GPT-2, GPT-3, и последняя версия, GPT-4, претерпели заметные улучшения за счет значительного увеличения числа параметров. Это дополнение облегчило выработку ответов, которые стали еще более точными и похожими на человеческие. GPT-4 получил возможность работать не только с текстом, но и с изображениями.

Важным прорывом в ChatGPT является его способность к обучению с нуля, что позволяет модели генерировать согласованные ответы на запросы, с которыми она никогда ранее не сталкивалась. Эта замечательная способность достигается за счет использования методов обучения без учителя и языкового моделирования.

Несмотря на ограничения ChatGPT, его применение распространилось на различные области, включая здравоохранение, кибербезопасность, экологические исследования, научную литературу, образование. Ожидается, что использование ChatGPT продолжит расширяться в будущем, с потенциальными разработками, направленными на расширение его возможностей. Эти разработки могут включать в себя обучение ChatGPT в режиме реального времени для повышения его производительности и расширение его знаний в конкретной предметной области, чтобы сделать его более адаптированным и персонализированным для конкретных областей, таких как обслуживание клиентов, здравоохранение, бизнес или финансы.

Кроме того, можно предпринять усилия для решения проблемы дезинформации, гарантируя, что ChatGPT предоставляет беспристрастные и справедливые ответы, тем самым повышая его надежность и приводя в соответствие с растущей важностью этики ИИ и соображений справедливости.

Проблемы ChatGPT

Исследователи выявили несколько проблем, касающихся ChatGPT, которые в широком смысле можно разделить на две группы: внутренние ограничения и проблемы, а также связанные с использованием. Эти ограничения затрудняют использование и развертывание ChatGPT в реальных сценариях.

Внутренняя проблема

Внутренние проблемы относятся к ограничениям, присущим ChatGPT, и могут быть преодолены в первую очередь разработчиками инструмента путем усовершенствования алгоритма и/или обновления обучающих данных. Она включает в себя пять основных ограничений, а именно: галлюцинации, предвзятое сообщение, не в режиме реального времени, дезинформацию и необъяснимость. ChatGPT может галлюцинировать, т.е. создавать новые данные, которых не существует. Другой подобной проблемой является дезинформация. Обе проблемы могут привести к созданию противоречащих фактам или бессмысленных ответов, что может серьезно угрожать надежности сгенерированного контента. Ложные сообщения, генерируемые ChatGPT, могут быть легко ошибочно приняты за правдивые, особенно лицами, не вовлеченными в контекст.

Улучшение алгоритма, правильный ввод запросов и проверка сгенерированных ответов могли бы помочь преодолеть эти проблемы. Обучение с подкреплением посредством обратной связи с человеком также

поможет ChatGPT повысить достоверность своих ответов.

Помимо улучшения алгоритмов и обратной связи с людьми, доработка обучающих данных для удаления или пометки «ложного» контента может помочь в этом направлении. Существует много важных приложений

ChatGPT, где требуются здравые рассуждения и объяснение шагов логического вывода. Это включает в себя принятие решений в различных областях, где не допускаются ошибки, таких как финансовые услуги, науки об окружающей среде, здравоохранение и т.д. В таких сценариях ChatGPT должен не только предоставлять точную информацию, которая может быть использована для принятия решений, но и упоминать этапы, связанные с процессом логического вывода.

Проблемы, связанные с использованием

Категория проблем, связанных с использованием, включает неэтичное использование инструмента, контент, нарушающий авторские права, и чрезмерную зависимость от ChatGPT. Возникают этические проблемы, особенно когда инструмент используется для создания контента без подтверждения. Неэтичное использование также включает в себя сознательное создание манипулируемого контента, который может способствовать дезинформации и провоцировать насилие, нанося ущерб на индивидуальном или организационном уровне. Этичное использование инструмента включает в себя упоминание ChatGPT в качестве автора сгенерированной информации. На самом деле, лишь немногие издатели признали этичным использование ChatGPT для академического письма. Кроме того, должны быть разработаны законы и нормативные акты, предусматривающие наказание за неэтичное использование ChatGPT. Существуют также опасения, что контент, созданный с помощью ChatGPT, может привести к нарушениям авторских прав и прав интеллектуальной собственности. Нарушение авторских прав в первую очередь включает в себя создание полной или частичной информации, идентичной уже опубликованным работам, без предварительного согласия владельца. Поскольку ChatGPT не знает о материалах, защищенных авторским правом, проверка сгенерированного контента перед использованием или публикацией, и пометка пользователями в случае нарушения авторских прав может помочь решить эту проблему. Наконец, существует опасение чрезмерной зависимости от инструмента, что может сделать людей ленивыми и апатичными и заставить их всегда полагаться на сгенерированную информацию. Следовательно, мы должны проверять сгенерированный контент и использовать ChatGPT только в качестве инструмента для достижения лучших результатов.

Список литературы / References

1. *Abdel-Messih M.S., Kamel Boulos M.N.* - «ChatGPT in clinical toxicology».
2. *Agathokleous E., Saitanis C.J., Fang C., Yu Z.* «Use of chat GPT: What does it mean for biology and environmental science? ».
3. «Potential Use of Chat GPT in Global Warming» Som S. Biswas.
4. «How Chat GPT Can Transform Autodidactic Experiences and Open Education?» Mehmet Firat.
5. [Электронный ресурс]. Режим доступа: https://openai.com/ (дата обращения: 30.07.2023).