

УДК 004.89

МАЙТАК Р. В., студент гр. ИИМ-231 (КузГТУ)
Научный руководитель БАУМГАРТЭН М.И., к.ф.-м.н., доцент (КузГТУ)
г. Кемерово

БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ И ИХ МЕСТО В СОВРЕМЕННЫХ ИНТЕЛЛЕКТУАЛЬНЫХ ТЕХНОЛОГИЯХ

Искусственный интеллект стремительно интегрируется в повседневную жизнь людей. Разработчики интеллектуальных систем в ходе своей работы пытаются реализовать такую систему, которая была бы универсальной и позволяла бы решать большинство задач, не автоматизированных человеком прежде [1].

Однако объять все направления человеческого знания, которые были накоплены за тысячелетия опыта, в одной интеллектуальной системе не представляется возможным даже современным суперкомпьютерам [2]. Поэтому ученые из области анализа данных приняли решение декомпозировать искусственный интеллект на разделы. Наиболее популярные и востребованные из них представлены на рисунке 1.



Рисунок 1. Разделы искусственного интеллекта

В рамках данной статьи мы детальнее рассмотрим «двигатель развития» одного из разделов искусственного интеллекта, связанного с пониманием и анализом текстов на естественном языке, которым является группа методов больших языковых моделей [3].

Большая языковая модель (от английского «Large Language Model» или LLM) – это нейронная сеть, для обучения которой использовались большие фрагменты текста, написанного на естественном (то есть собственном для человека) языке.

Немногие из людей, далеких от области искусственного интеллекта, могли бы дать определение обозначенному термину, однако каждый из нас так или иначе уже столкнулся с присутствием этих моделей в повседневной жизни.

- Так, в браузерах и «поисковиках» интернета информация стала находиться гораздо быстрее. Теперь мы с большей скоростью находим именно то, что нам нужно, даже если сформулированный текстовый запрос не описывает искомый предмет напрямую [4].

- Порой на просторах сети или в реальной жизни встречается изображение, о котором хочется узнать больше. В современных условиях уже стало привычным делом сфотографировать объект и прикрепить картинку к строке поиска: по изображению мы сможем очень быстро получить соответствующую информацию и текст, характеризующий картинку. Несмотря на то, что картинка и текст представляют собой разные формы представления информации, они, тем не менее, могут быть связаны воедино с помощью языковых моделей [5].

- Голосовые помощники и онлайн-переводчики стали работать значительно лучше за последние 5 лет [6]. При этом секрет их успеха кроется не в совершенствовании аппаратных составляющих персональных компьютеров, а именно в интеграции больших языковых моделей во внутреннюю структуру работы приложений. Такие модели позволили ИИ понимать контекст введенного пользователем предложения, что многократно увеличило качество интерпретации данных компьютерами.

Одной из наиболее известных архитектур больших языковых моделей является «ChatGPT» от компании OpenAI [7]. Рост её популярности во многом доказывает успех применения больших языковых моделей для автоматизации ряда повседневных человеческих задач (см. рисунок 2).



Рисунок 2. Рост популярности большой языковой модели «ChatGPT» на основе запросов в поиске «Яндекс»

Несмотря на то, что популярность сервиса «ChatGPT» может со временем уменьшаться, это не означает, что подобные языковые модели теряют свою популярность. Лидирующие позиции «ChatGPT» смещены чаще всего другими моделями, которые принадлежат к всё тому же классу методов языковых моделей.

Например, по состоянию на ноябрь 2023 года значительную популярность приобрела языковая модель LLAMA (от английского акронима «Large Language Model Meta AI») [8], которая характеризуется большим числом обучаемых параметров и, соответственно, более высокой скоростью выдачи ответов, чем «ChatGPT» [9]. В январе 2024 года начала набирать популярность другая языковая модель – МАМБА [10]. Эта программная архитектура, в свою очередь, характеризуется усовершенствованным способом обучения линейно-временных последовательностей, что позволяет модели выборочно прогнозировать более релевантный для пользователя результат.

Принимая во внимание вышеуказанные факты, стоит отметить, что все перечисленные модели и алгоритмы, которые так или иначе поднимались на волне популярности, объединены в один общий класс языковых моделей. Если проследить за цикличностью улучшений модельных разработок в этом разделе искусственного интеллекта, то можно обозначить небезосновательное предположение о том, что в недалеком будущем тенденция развития больших языковых моделей приведет человечество к универсальной модели прикладного искусственного интеллекта, которая будет автоматизировать все задачи, связанные с обработкой информационных потоков. Объединение усилий ученых по созданию обобщенной языковой модели позволит человечеству существенно ускорить прогресс в разных областях научного знания за счет инструментария искусственного интеллекта.

Список литературы:

1. Кормен, Т., Лейзерсон, Ч. Алгоритмы: построение и анализ / Т. Кормен. – М.: ООО И.Д. Вильямс. 2013. – 1328 с.
2. Жерон, О. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow / О. Жерон. – М.: O'Reilly Media. 2018. – 662 с.
3. Большие языковые модели: [сайт]. – Амстердам, 2024 –URL: <https://www.elastic.co/what-is/large-language-models> (дата обращения: 31.01.2024). – Текст: электронный.
4. Колец, Д. Классические задачи Computer Science на языке Python / Д. Колец. – СПб.: Питер. 2020 – 256 с.
5. Яндекс Справка. Поиск по картинке: [сайт]. – Москва, 2024 –URL: <https://yandex.ru/support/images/loaded-image.html> (дата обращения: 31.01.2024). – Текст: электронный.
6. «Спросите Siri», Диктовка: [сайт]. – Купертино, 2024 –URL: <https://www.apple.com/ru/legal/privacy/data/ru/ask-siri-dictation/> (дата обращения: 31.01.2024). – Текст: электронный.
7. OpenAI: [сайт]. – Сан-Франциско, 2024 –URL: <https://openai.com> (дата обращения: 31.01.2024). – Текст: электронный.
8. ChatGPT: [сайт]. – Сан-Франциско, 2024 –URL: <https://chat.openai.com> (дата обращения: 31.01.2024). – Текст: электронный.
9. LLaMa: [сайт]. – Кембридж, 2024 –URL: <https://ai.meta.com/llama/> (дата обращения: 31.01.2024). – Текст: электронный.

10. Mamba: Linear-Time Sequence Modeling with Selective State Spaces: [сайт]. – Принстон, 2024 –URL: <https://arxiv.org/abs/2312.00752> (дата обращения: 31.01.2024). – Текст: электронный.