ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.89

БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ С ПОИСКОВОЙ РАСШИРЕННОЙ ГЕНЕРАЦИЕЙ: ОБЗОР И ПЕРСПЕКТИВЫ

Федоров Виктор Олегович
кандидат технических наук, доцент кафедры «Системы обработки информации» КФ МГТУ им. Н.Э. Баумана, г. Калуга
E-mail: fedorov_vo@bmstu.ru

Поляков Роман Андреевич
студент КФ МГТУ им. Н.Э. Баумана, г. Калуга
E-mail: polyakovra@student.bmstu.ru

Аннотация

В данной работе представлено описание метода поисковой расширенной генерации (RAG). Описан принцип работы систем с RAG. Приводятся лучшие практики внедрения RAG в платформы машинного обучения, подчеркивая важность качества и релевантности данных. Статья также обсуждает перспективы технологии RAG. В заключение подчеркивается активное развитие инструментов и исследований в области расширенной поисковой генерации, с ожидаемым влиянием на цифровые продукты и бизнес-решения.

Ключевые слова: машинное обучение, искусственный интеллект, большая языковая модель; поисковая расширенная генерация, LLM, RAG.

LARGE LANGUAGE MODELS WITH RETRIEVAL AUGMENTED GENERATION: OVERVIEW AND PERSPECTIVES

Victor O. Fedorov
Ph.D. in Technical Sciences, Associate Professor at the Department of Information Processing Systems, BMSTU (KB), Kaluga
E-mail: fedorov_vo@bmstu.ru

Roman A. Polyakov
Student at the BMSTU (KB), Kaluga
E-mail: polyakovra@student.bmstu.ru

ABSTRACT

The paper provides a description of the Retrieval Augmented Generation (RAG) method. It outlines the working principle of systems utilizing RAG. The best practices for implementing RAG in machine learning platforms are discussed, emphasizing the importance of data quality and

relevance. The article also explores the prospects of RAG technology. In conclusion, there is an emphasis on the active development of tools and research in the field of extended search generation, with anticipated impacts on digital products and business solutions.

Keywords: machine learning, artificial intelligence, large language model; retrieval augmented generation, LLM, RAG.

Сегодня такие инструменты, как ChatGPT, являющиеся частью крупной сферы генеративных моделей искусственного интеллекта, могут автоматизировать до 60-70% задач, которые обычно занимают ценное время специалистов разного уровня. Однако около 60% руководителей крупного и среднего бизнеса по всему миру не решаются внедрять в текущие бизнес-процессы подобные инструменты, ссылаясь на опасения по поводу безопасности и неточностей в результатах, генерируемых ИИ.[1]

Метод поискового расширенного поиска (Retrieval Augmented Generation, RAG) представляет собой новаторский подход, объединяющий возможности поисковых систем с передовыми большими языковыми моделями (Large Language Models, LLM). Эта техника, особенно в области мультимодального языкового моделирования, обеспечивает синергию точности и креативности, повышая производительность языковых моделей путем предоставления модели контекста вместе с запросом. [2] За счет передаваемой актуальной и проверенной информации, конечный пользователь может получить ссылки на данные, которыми оперирует ИИ, что обеспечивает большую объективность получаемых ответов.

Несмотря на свой потенциал, RAG является относительно новой разработкой, пока не получившей широкого распространения.

Путь развития RAG. В 2020 году команда из Meta Research создала так называемые RAG-модели для более точной работы с информацией. Эти модели, как объясняют Патрик Льюис – научный сотрудник по обработке естественного языка, и его команда, — новый способ улучшения того, как компьютеры понимают и используют информацию. Они объединяют два типа систем памяти: одна похожа на долговременную память компьютера, который знает много языков (так называемая параметрическая память), а другая больше напоминает базу данных с возможностью поиска, как например коллекция статей Википедии (непараметрическая память). [3]

Льюис и его коллеги работали над тем, чтобы сделать эти модели RAG лучше. В этих моделях знания языка из первого типа памяти смешиваются с базой данных Википедии. Для поиска в Википедии они используют специальный инструмент (предварительно обученный нейронный ретривер). Существует два вида RAG-моделей: одна использует одни и те же статьи Википедии для всего текста, который она создает, а другая может переключаться между разными статьями для разных частей текста.

Они усовершенствовали эти модели, обучая их на задачах, требующих большого количества знаний. Самое интересное, что эти модели установили новые рекорды в ответах на открытые вопросы. Они были лучше, чем старые модели, которые использовали только языковую память, или те, которые совмещали поиск и извлечение информации. Когда дело дошло до создания текста, модели RAG оказались более точными, разнообразными и верными, чем лучшие модели до, которые использовали только систему языковой памяти.

Принцип работы систем с RAG. Активные RAG-системы динамически извлекают информацию из внешних источников, таких как базы данных, интернет-ресурсы или конкретные наборы документов различных форматов, в режиме реального времени. Этот процесс — не просто пассивный поиск информации, а активный, контекстный поиск,

основанный на запросе пользователя или контексте разговора. Затем система использует полученные данные для формирования ответов. Ключевыми компонентами активного RAG являются:

- Динамический поиск информации: активные системы RAG постоянно ищут и обновляют внешние источники знаний, обеспечивая актуальность и релевантность используемой информации.
- Контекстно-ориентированная обработка: система понимает и анализирует контекст запроса, что позволяет ей получать информацию, точно соответствующую потребностям пользователя.
- Интеграция с LLM: полученная информация легко интегрируется в процесс генерации ответа большой языковой моделью, обеспечивая не только точность ответов, но и естественность формулировок.

RAG, как правило, работает в два этапа:

- Фаза поиска: алгоритмы ищут и извлекают релевантную информацию на основе запроса пользователя, включая специфические для пользователя сведения и обновленный фактический контекст на основе «подмешенной» дополнительной информации.
- Фаза генерации контента: после получения информации генеративная языковая модель использует этот контекст для создания ответов. Ответы зависят от точности найденных данных и могут ссылаться на источники информации.

Лучшие практики внедрения RAG. При использовании RAG в платформе машинного обучения важно помнить о нескольких ключевых передовых практиках:

- Качество и релевантность данных: эффективность RAG в значительной степени зависит от качества получаемых данных. Обеспечение актуальности и точности данных в базе знаний имеет решающее значение.
- Тонкая настройка для контекстного понимания: важно точно настроить генеративные модели, чтобы они понимали и эффективно использовали контекст, предоставляемый полученными данными.
- Баланс между поиском и генерацией: достижение баланса между полученной информацией и творческим вкладом генеративной модели является ключом к сохранению оригинальности и ценности результата.
- Этические соображения и уменьшение предвзятости: учитывая зависимость от внешних источников данных, важно учитывать этические аспекты и активно работать над снижением предвзятости получаемых данных.

Примеры использования. Некоторые распространенные примеры использования, в которых RAG особенно эффективен и заслуживает внимания [4]:

- Чат-боты для поддержки клиентов: в сфере обслуживания клиентов RAG позволяет чат-ботам давать более точные и контекстуально подходящие ответы. Получая доступ к актуальной информации о продукте или данным о клиенте, такие чат-боты могут оказывать более качественную помощь, повышая уровень удовлетворенности клиентов.

- Бизнес-аналитика и анализ: бизнес может использовать RAG для создания отчетов об анализе рынка или аналитических материалов.
- Информационные системы здравоохранения: в здравоохранении RAG может улучшить системы, предоставляющие медицинскую информацию или консультации. Получая доступ к последним медицинским исследованиям и рекомендациям, такие системы могут предлагать более точные и безопасные медицинские рекомендации.
- Юридические исследования: специалисты в области права могут использовать RAG для быстрого поиска соответствующих нормативно-правовых актов, уставов и иных юридических документов, что упрощает процесс исследования и обеспечивает более полный правовой анализ.
- Создание контента: при создании контента, например при написании статей или отчетов, RAG может повысить качество и релевантность результатов.
- Образовательные инструменты: RAG можно использовать в образовательных платформах, чтобы предоставить учащимся подробные объяснения и контекстуально значимые примеры, опираясь на широкий спектр учебных материалов.

Перспективы технологии RAG. Уже сейчас, на момент написания данной статьи, в глобальной сети достаточно руководств по созданию инструментов с применением поисковой расширенной генерацией. Этому способствует и бесплатный доступ к ассистентам на основе генеративного ИИ, и фреймворки для разработчиков, типа LangChain, предоставляющих высокий уровень абстракции при написании кода. Также появляются больше исследований, направленных на оценку точности доступных генеративных языковых моделей. Так в недавнем исследовании Дмитрий Гуреев – руководительно цифровой трансформации бизнеса в одной из медицинских компаний, совместно с командой экспертов-юристов протестировал 8 моделей от компаний OpenAi, Yandex и Sber. Наилучший результат показывает GPT4 с токенайзером ada-02 с результатом в 71% правильных ответов. Результаты YandexGPT2 и GigaChat оказались значительно ниже. В исследовании также подчеркивается важность разработки локализованных под конкретную задачу токенайзеров и качественной подготовки текста перед использованием языковых моделей. [5] Возможные качественные скачки в области применения расширенной поисковой генерации ограничиваются несколькими факторами:

- Финансовый: стоимость запроса к той же GPT4 в разы выше по отношению к GPT3.5.
- Объемы информации: локальные базы знаний компаний в большинстве случаев превосходят в количественном выражении максимальный размер запроса, передаваемого языковой модели.
- Безопасность: безусловно каждый бизнес так или иначе заботиться об информационной безопасности, поэтому задача применения RAG с LLM становиться сложнее из-за локального развертывания проектируемых систем.

Постепенно с решением двух первых вопросов со стороны компаний, продвигающих свои большие языковые модели, будет реагировать и потребительский бизнес, предлагая новые решения для своих цифровых продуктов.

В заключение следует отметить, что RAG — это значительный шаг вперед в области ИИ, предлагающий одновременно точность и творческий подход, обладающий огромным потенциалом. Понимая и внедряя лучшие практики, а также изучая различные варианты его использования, мы сможем использовать весь потенциал RAG в различных областях. По мере того, как мы будем продолжать изучать и совершенствовать эту технологию, возможности ее применения будут казаться безграничными, обещая будущее, в котором ИИ станет более полезным, точным и проницательным.

Список литературы:

1. Chui, M., Roberts, R., Yee, L., Hazan, E., Singla, A., Smaje, K., Sukharevsky, A., Zemmel, R. / The economic potential of generative AI: The next productivity frontier. URL: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier (дата обращения 20.12.2023)
2. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H. / Retrieval-Augmented Generation for Large Language Models: A Survey. URL: https://doi.org/10.48550/arXiv.2312.10997 (дата обращения 20.12.2023)
3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., Kiela, D. / Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. URL: https://doi.org/10.48550/arXiv.2005.11401 (дата обращения 20.12.2023)
4. Kimothi, A. / Creating Impact: A Spotlight on 6 Practical Retrieval Augmented Generation Use Cases. URL: https://www.linkedin.com/pulse/creating-impact-spotlight-6-practical-retrieval-use-cases-kimothi-ya02c (дата обращения 20.12.2023)
5. Гуреев, Д. / Большой тест GPT4, GPT3.5, YandexGPT, GigaChat, Saiga в RAG-задаче. Часть 1. — URL: https://habr.com/ru/articles/782484 (дата обращения 20.12.2023)

References:

1. Chui, M., Roberts, R., Yee, L., Hazan, E., Singla, A., Smaje, K., Sukharevsky, A., Zemmel, R. "The economic potential of generative AI: The next productivity frontier.". URL: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier (Accessed: 12/20/2023)
2. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H. "Retrieval-Augmented Generation for Large Language Models: A Survey." . URL: https://doi.org/10.48550/arXiv.2312.10997 (Accessed: 12/20/2023)
3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., Kiela, D. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." . URL: https://doi.org/10.48550/arXiv.2005.11401 (Accessed: 12/20/2023)
4. Kimothi, A. "Creating Impact: A Spotlight on 6 Practical Retrieval Augmented Generation Use Cases.". URL: https://www.linkedin.com/pulse/creating-impact-spotlight-6-practical-retrieval-use-cases-kimothi-ya02c (Accessed: 12/20/2023)
5. Gureev, D. "Bolshoy Test GPT4, GPT3.5, YandexGPT, GigaChat, Saiga V RAG-zadache. Chast 1.". URL: https://habr.com/ru/articles/782484 (Accessed: 12/20/2023)