Основы научно-исследовательской деятельности. Трек «Исследование»

Вводное задание «Мой исследовательский интерес»

Гладышев В.В.

1 Поиск темы

В настоящее время наиболее динамично развивающейся отраслью информационных технологий является применение машинного и глубокого обучения для формирования и использования больших языковых моделей (large language model - LLM) для решения задач сложных для формализации алгоритма решения. Такие задачи как перевод с одного языка на другой, создание текста на произвольную тему, создание изображений произвольного стиля, написание кода и множество других требуют навыков которые человек накапливает в процессе обучения и часто требуют обобщения различных областей знания. Модели и алгоритмы для решения таких сложно формализуемых задач часто относят к категории искусственного интеллекта.

Сочетание больших вычислительных ресурсов, больших языковых моделей и алгоритмов обучения с подкреплением позволило достичь прорывных результатов ещё десять лет назад относившихся к категории фантастики с отдалённым горизонтом реализации. Достижение прорывных результатов вывело эту область в лидеры по привлечению ресурсов и финансирования. С достижениями в области искусственного интеллекта ведущие геополитические силы Китай и США связывают надежды на достижение технологического превосходства, что порождает высокую конкуренцию и ускорение развития в этой сфере.

Однако, как и в любой практической технологии, в архитектуре и методах применения больших языковых моделей заложены определённые ограничения. Так, кроме потребности в больших вычислительных ресурсах в LLM есть критические ограничения при формировании и удержании контекста – реализации краткосрочной памяти. Сам принцип формирования таких моделей - сложное обучение и «выравнивание», невозможность динамически добавить новые знания,

критически ограничивают объем текущей информации пределами контекстного окна.

Для преодоления этого критического ограничения в настоящее время сосредоточены большие усилия и применяется ряд мер: количественное расширение контекстного окна, применение цепочек рассуждений, мультиагентности, создание новых архитектур [1], дополнение контекста с применением технологий поиска и структурирование с применением графов знаний.

2 Анализ области исследования

2.1 Варианты тем

2.1.1 Направление исследования. Определение объекта исследования

Практическое применение больших языковых моделей для решения текущих задач автоматизации обработки информации, генерации текста или кода в разрезе конкретной проблемы требует поиска и создания алгоритмов управления контекстом.

Кроме наиболее очевидной проблемы – ограниченности краткосрочной контекстной памяти практическое применение больших языковых моделей требует преодоления большого количества препятствий (галлюцинации, потребность в больших вычислительных ресурсах и т.д.).

Основным направлением исследований предлагается выбрать определение средств и реализация решений по преодолению ограничений применения больших языковых моделей.

2.1.2 Какие ключевые вопросы

Предлагается рассмотреть несколько вариантов тем по преодолению ограничений LLM:

1. Исследовать методы организации контекстной информации модели с применением структур данных основанных на графах. Например, применить представление основанное на графах для генерации кода большого проекта со

сложной структурой и определить насколько такой подход улучшает качество генерации.

2. Исследовать влияние применения графов для структурирования данных для генерации запросов к объёмным реляционным базам данных требующих нетривиального промежуточного извлечения и агрегирования данных.

3. Исследовать возможность повышения качества средств автоматизированной подготовки данных в задаче Retrieval Augmented Generation (RAG) для оптимизации формирования представления основанного на графах. Примером фреймворка с реализацией подобного подхода является GraphRAG компании Microsoft (доступна open-source реализация).

2.1.3 Дисциплины учебной программы для изучения этой темы

В учебную программу включены дисциплины необходимые для формирования компетенций для проведения исследований по предполагаемому направлению:

1. Основы машинного обучения;

2. Системы хранения и обработки данных;

3. Глубокое обучение в науках о данных;

4. Анализ естественного языка;

5. Задачи генерации в NLP;

6. Методы обучения с подкреплением.

Таким образом предполагаемое направление исследований тесно связано с программой обучения.

2.2 Академические источники

2.2.1 Список источников по теме

Список источников с установочной информацией по применению структур основанных на графах для формирования контекста LLM:

1. From Local to Global: A Graph RAG Approach to Query-Focused Summarization (апрель 2024 [2]);

2. AriGraph: Learning Knowledge Graph World Models with Episodic Memory for LLM Agents (июль 2024 [3]);
3. Graph Retrieval-Augmented Generation: A Survey (август 2024 [4]);
4. CypherBench: Towards Precise Retrieval over Full-scale Modern Knowledge Graphs in the LLM Era (декабрь 2024 [5]);
5. A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models (январь 2025 [6]).

2.2.2 Наиболее интересные аспекты

Важным фактором является новизна реализации инструментария структурирования контекста больших языковых моделей с применением структур на базе графов. Статья о GraphRAG Microsoft опубликована в апреле 2024 года, остальные ещё боле поздние.

Статьи [1], [2], [4] написаны по результатам реализации программных продуктов и решений. Ключевые элементы данных проектов реализованы в рамках концепции open-source. Анализ данных решений может облегчить и ускорить проведение исследований.

3 Междисциплинарные связи

3.1 Определение набора дисциплин для решения задачи исследования

Исследование применения структур данных основанных на графах для оптимизации формирования контекста больших языковых моделей находится на пересечении дисциплин: глубокого обучения (deep learning - DL), методов обработки естественного языка (natural language processing - NLP), обучения с подкреплением (reinforcement learning - RL) и теории графов.

3.2 Определение вопросов выходящих за рамки учебного курса

При проведении исследований может возникнуть необходимость реализации прототипа. Для реализации прототипа потребуются навыки создания frontend, backend части приложения, использования средств контейнеризации, использования облачных вычислительных ресурсов (DevOps), сбора и анализа данных для датасетов (анализ юридических аспектов).

Список литературы

1. Ali Behrouz, Peilin Zhong, Vahab Mirrokni Titans: Learning to Memorize at Test Time / Ali Behrouz, Peilin Zhong, Vahab Mirrokni [Электронный ресурс] // arxiv.org : [сайт]. — URL: https://arxiv.org/abs/2501.00663 (дата обращения: 02.02.2025);
2. Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Jonathan Larson From Local to Global: A Graph RAG Approach to Query-Focused Summarization / Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Jonathan Larson [Электронный ресурс] // arxiv.org : [сайт]. — URL: https://arxiv.org/abs/2404.16130 (дата обращения: 02.02.2025);
3. Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Mikhail Burtsev, Evgeny Burnaev AriGraph: Learning Knowledge Graph World Models with Episodic Memory for LLM Agents / Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Mikhail Burtsev, Evgeny Burnaev [Электронный ресурс] // arxiv.org : [сайт]. — URL: https://arxiv.org/abs/2407.04363 (дата обращения: 02.02.2025);
4. Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, Siliang Tang Graph Retrieval-Augmented Generation: A Survey / Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, Siliang Tang [Электронный ресурс] // arxiv.org : [сайт]. — URL: https://arxiv.org/abs/2408.08921 (дата обращения: 02.02.2025);
5. Yanlin Feng, Simone Papicchio, Sajjadur Rahman CypherBench: Towards Precise Retrieval over Full-scale Modern Knowledge Graphs in the LLM Era / Yanlin Feng, Simone Papicchio, Sajjadur Rahman [Электронный ресурс] // arxiv.org : [сайт]. — URL: https://arxiv.org/abs/2412.18702 (дата обращения: 02.02.2025);
6. Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Junnan Dong, Hao Chen, Yi Chang, Xiao Huang A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models / Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Junnan Dong, Hao Chen, Yi Chang, Xiao Huang [Электронный ресурс] // arxiv.org : [сайт]. — URL: https://arxiv.org/abs/2501.13958 (дата обращения: 02.02.2025).