УДК 004.8

РАЗРАБОТКА АЛГОРИТМА РЕРАНКИНГА МУЛЬТИМОДАЛЬНЫХ ДАННЫХ

Стулов К.В.
Научный руководитель – ассистент Вершинин В.К.
Университет ИТМО

Работа выполнена в рамках темы НИР «Разработка алгоритма реранкинга мультимодальных данных».

Введение

В настоящее время в документах — различных статьях, отчетах и инструкциях, важен не только текст, но и визуальные элементы: графики, схемы, таблицы. Даже современные RAG-системы (Retrieval-Augmented Generation), которые заявлено, что понимают разные типы данных, испытывают трудности на этапе реранкинга результатов поиска. Классические реранкеры анализируют лишь текстовую информацию и не учитывают, содержание визуальных материалах. Таким образом, страница с важным графиком может оказаться в конце рейтинга, при этом в начале будет страница с корректным текстом, но неинформативной иллюстрацией. Вследствие этого языковая модель получает искаженный контекст, и ответ получается неточным.

Как показывают исследования [1, 2], на данный момент в мультимодальном RAG (MRAG) основной акцент делается на первичный поиск (bi-encoder, CLIP) и генерацию ответа с помощью VLM. А вот действительно функциональный мультимодальный реранкинг мало у кого реализован. В связи с этим, существует потребность в таком инструменте, способном учитывать визуальную семантику документов при сортировке результатов поиска, с сохранением жёстких ограничений по латентности (p50/p95) и компактности модели (до 8 ГБ VRAM). Данное исследование как раз и направлено на реализацию такого инструмента для MRAG-систем.

Основная часть

В данной работе реализован и протестирован компактный мультимодальный реранкинг, который предназначен для пересортировки результатов поиска в MRAG-системах. Разработанная архитектура – это продукт, готовый к применению, который включает в себя поиск кандидатов, гибридный мультимодальный реранкинг и генерацию финального ответа.

Для поиска кандидатов используется bi-encoder архитектура, сочетающая обычный поиск по словам (BM25) и векторные мультимодальные эмбеддинги (CLIP). Основное преимущество в пересортировке: модуль обрабатывает пару «запрос и страница документа» вместе, учитывая связь между словами в запросе и визуальным контентом страницы. В отличие от традиционных решений, гибридный метод лучше учитывает, как связаны между собой разные типы данных.

В основе модуля пересортировки находится дообученный кросс-энкодер, способный определять, какой документ лучше подходит к запросу. Чтобы модель была небольшой и работала быстро, был осуществлен перенос знаний из больших VLM-моделей в данную архитектуру, ограничив объём памяти до 8 ГБ. Это позволяет получить хорошее качество ранжирования без ущерба производительности.

Модуль реранкинга реализован как независимый компонент с API на FastAPI и интегрирован в RAG-конвейер. Для удобства был выполнен веб-интерфейс на Streamlit. Финальные ответы генерируются языковой моделью на основе упорядоченного контекста.

Эффективность оценивалась на специальных мультимодальных бенчмарках, содержащими разные типы визуальных материалов. Метрики: Recall@k, nDCG@k, MRR, точность ответа и показатели латентности (p50/p95).

Выводы

В результате был разработан и экспериментально проверен алгоритм мультимодального реранкинга для MRAG-систем. Данная гибридная архитектура, сочетающая bi-encoder для генерации кандидатов и дообученный кросс-энкодер, оптимально обрабатывает текстовые и визуальные данные в документах. За счет дистилляции модель получилась компактной (<8 ГБ) и работает достаточно оперативно для практического использования.

Эта работа полезна тем, что создан готовый к использованию продукт. Модуль реранкинга можно добавить в существующие RAG-системы, чтобы улучшить поиск в базах знаний компаний, библиотеках, патентных ведомствах и аналитических системах, где важна информация не только в тексте, но и в визуальных элементах. Все скрипты, настройки и результаты тестов выложены в открытый доступ, так что любой желающий может повторить эксперименты и дообучить модель на своих данных.

Литература

1. Yang, Y., Zhong, J., Jin, L., Huang, J., Gao, J., Liu, Q., Bai, Y., Zhang, J., Jiang, R., & Wei, K. (2025, February 20). Benchmarking Multimodal RAG through a Chart-based Document Question-Answering Generation Framework. arXiv. https://arxiv.org/abs/2502.
2. Shen, W., Wang, M., Wang, Y., Chen, D., Yang, J., Wan, Y., & Lin, W. (2025, August 5). Are We on the Right Way for Assessing Document Retrieval-Augmented Generation? arXiv. https://arxiv.org/abs/2508.03644.