УДК 004.8

ОБУЧЕНИЕ МУЛЬТИМОДАЛЬНОЙ МОДЕЛИ ПОИСКА ДЛЯ РАБОТЫ С РУССКОЯЗЫЧНЫМИ ДАННЫМИ

Юхневич Е.Д. (ИТМО)
Научный руководитель – Маслюхин С. М. (ООО «ЦРТ»)

Введение. Модели семантического поиска находят применения во множестве различных задач, таких как генерация, дополненная поиском (RAG, англ. Retrieval Augmented Generation) [1]. Большинство существующих моделей семантического поиска работают только с текстовой информацией, что существенно ограничивает их эффективность во многих задачах из-за потери информации, содержащаяся в таблицах, изображениях и инфографике в базе документов.

Из-за этого важным направлением исследования, способным существенно расширить возможности семантического поиска и его применений в RAG-системах, является мультимодальный поиск. Также важным аспектом в изучении работы моделей поиска является их способность поддерживать разные языки и дообучаться для работы с различными языками.

Основная часть. В работе было проведено сравнение различных подходов к решению задачи мультимодального поиска. Выбранная модель – ColPali – отличается наиболее лёгкой архитектурой и высокой скоростью работы без потери качества результатов работы [2]. Данная модель разделяет документы из базы данных на скриншоты страниц документов и выявляет релевантные к запросу пользователя документы путём сравнения близости эмбеддингов текста запроса и изображения скриншота страницы. Модель обучалась на англоязычных данных, однако, как описано в оригинальной статье [2], способна решать задачу работы с другими языками как zero-shot задачу.

Для оценки способности модели дообучаться для работы с русскоязычными данными был собран набор данных из пар изображение-текст: скриншотов страниц документов и вопросов, ответ на которых содержится на страницах. Для создания набора изображений использовались учебные пособия по техническим специальностям и техническая документация. Для генерации вопросов к изображениям использовались модели GigaChat и QwenVL. При сравнении результатов генерации был сделан вывод, что модель GigaChat больше подходит для работы с русскоязычными данными.

Модель ColPali была дообучена на собранном наборе данных с использованием подхода 4-разрядного квантования с LoRA (QLoRA). В процессе дообучения наблюдалось снижение функции ошибки с 0,2952 до 0,1973, что свидетельствует о том, что модель эффективно дообучалась для работы с русскоязычными данными.

Выводы. В работе для исследования была выбрана модель мультимодального поиска ColPali. Для её обучения был собран набор данных, состоящий из скриншотов страниц документов и релевантных к ним вопросов на русском языке. Снижение функции ошибки в процессе дообучения свидетельствует об эффективном обучении модели ColPali для работы с русскоязычными данными.

Список использованных источников:

1. Fan W. et al. A survey on rag meeting llms: Towards retrieval-augmented large language models //Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. – 2024. – С. 6491-6501.
2. Faysse M. et al. Colpali: Efficient document retrieval with vision language models //arXiv preprint arXiv:2407.01449. – 2024.