УДК 004.912

АДАПТЕРНАЯ КАЛИБРОВКА КОДИРУЮЩИХ LLM ДЛЯ ЗАДАЧ ИНФОРМАЦИОННОГО ПОИСКА

Посохов П.А. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Коротеева О.В. (ИТМО)

Введение. Одной из ключевых проблем в нейронном информационном поиске с использованием векторных представлений является усечение ранжированного списка кандидатов из-за недостаточной разделимости распределений значений функций подобия релевантных и нерелевантных данных. Для решения этой проблемы применяются калибровочные модели. Активное развитие больших языковых моделей (LLM) значительно повлияло на обработку естественного языка, однако их применение в информационном поиске остается малоизученным. В частности, отсутствует анализ эффективности калибровочных моделей для усечения ранжированных списков в LLM, что делает данное исследование актуальным.

Основная часть. В данной работе проводится анализ имеющихся моделей и методов калибровки значений функции подобия кандидатов, а именно Cosine Adapter [1] и Surprise [2] для кодирующих LLM. А также представлен новый подход к калибровке кодирующих моделей (LLM) для задач информационного поиска — TMP Adapter. Он основан на улучшенной архитектуре Cosine Adapter и включает Threshold Margin Penalty (TMP). TMP представляет собой модификацию маржинальной функции ошибки [3], используемую как дополнительную функцию потерь для калибровки релевантности. Основная проблема существующих методов заключается в низкой емкости моделей, нестабильности обучения и несогласованности выбранных пороговых значений, особенно в условиях сдвига распределения между обучающими и тестовыми данными. Для решения этих проблем предложен метод TMP Adapter, который улучшает согласованность пороговых значений и разделение между положительными и отрицательными парами, что критически важно для эффективного усечения ранжированных списков.

Для проведения исследований были отобраны три ключевых набора данных из BEIR-бенчмарка: FiQA (финансовые вопросы и ответы), NFCorpus (медицинские документы) и Robust04 (новостные статьи). Эти наборы данных представляют разнообразные задачи информационного поиска, что позволяет оценить качество калибровки кодирующих моделей в различных доменах. В качестве оценки калибровки использовались метрики качества усечения ранжированного списка – F1-score.

Выводы. Экспериментальные результаты показали, общую эффективность существующих методов калибровки LLM кодирующих моделей улучшив изначальные значения метрик F1-score на 2.01%. Вместе с этим предложенный TMP Adapter позволяет улучшить исходные метрики LLM модели на 4.25% что доказывает эффективность данной модели в сравнении с существующими методами калибровки.

Список использованных источников:

1. Rossi N. et al. Relevance filtering for embedding-based retrieval //Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. – 2024. – С. 4828-4835.
2. Bahri D. et al. Surprise: Result List Truncation via Extreme Value Theory //Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2023. – С. 2404-2408.
3. Zhang Q. et al. Threshold-consistent margin loss for open-world deep metric learning. – 2024.