УДК 004.912

ОПТИМИЗАЦИЯ ВЫЧИСЛЕНИЙ В НЕЙРОННЫХ СЕТЯХ АРХИТЕКТУРЫ "ТРАНСФОРМЕР" ДЛЯ ЗАДАЧИ ИНФОРМАЦИОННОГО ПОИСКА

Мосунов К.Д. (Университет ИТМО)
Научный руководитель – к. т. н. Шуранов Е.В.
(Университет ИТМО)

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР № 622282 «Разработка русскоязычного персонифицированного диалогового агента с динамической долгосрочной памятью».

Введение. На сегодняшний день в области обработки естественного языка большую популярность обрели подходы, использующие нейросетевые модели архитектуры «Трансформер». Так называемый механизм внимания [1], лежащий в основе данной архитектуры, помогает формировать векторные представления предложений или слов с учётом их контекста, что позволило обучать модели более высокого качества на различных прикладных задачах от построения систем информационного поиска по документам из сети Интернет до разработки персонифицированных диалоговых моделей. Нередко подобные системы включают в себя векторный индекс помимо самой сети-энкодера для быстрого поиска релевантных ответов.

Основная часть. Ввиду высокой вычислительной интенсивности моделей-трансформеров появляется необходимость в ускорении и оптимизации их работы как в случае диалоговых агентов, так и в случае систем информационного поиска и индексирования большого числа документов. В данной работе рассматриваются два подхода к оптимизации подобных систем: оптимизация сети-энкодера [2], преобразующей текстовые данные в их векторные представления, а также оптимизация векторного индекса [3], позволяющего хранить и производить эффективный поиск по векторам.

Выводы. В ходе проведения исследований рассматриваются различные открытые движки оптимизированного исполнения моделей-трансформеров, а также алгоритмы построения векторных индексов с применением сжатия векторов или же без него и проведено сравнение их эффективности.

Список использованных источников:

1. Vaswani A. et al. Attention is all you need // Advances in neural information processing systems. – 2017. – Т. 30.
2. Wang X. et al. LightSeq: A high performance inference library for transformers // arXiv preprint arXiv:2010.13887. – 2020.
3. Malkov Y. A., Yashunin D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs // IEEE transactions on pattern analysis and machine intelligence. – 2018. – Т. 42. – №. 4. – С. 824-836.