Исследование эффективной структуры ортогональных вращений при квантовании активаций больших языковых моделей

Пойманов Дмитрий Романович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: dima.poimanow@yandex.ru

Научный руководитель *— Грабовой Андрей Валериевич*

В данной работе рассматривается задача повышения эффективности квантования активаций в больших языковых моделях. Актуальность данной задачи связана с необходимостью уменьшения вычислительных затрат и требований к памяти при применении современных языковых моделей. Исследуется квантование как весов, так и активаций моделей в 4 бита (режим W4A4). Качество модели существенно зависит от свойств распределения активаций и используемого числового формата. В существующих работах [1, 2] для повышения устойчивости квантования в формате INT4 применяются ортогональные матрицы вращения, позволяющие перераспределить значения активаций для уменьшения выбросов. Одновременно с этим в последние годы были предложены новые форматы числового представления, такие как MXFP4 [3] и NVFP4 [4]. Целью работы является исследование необходимости применения ортогональных преобразований при использовании новых форматов, а также определение эффективной структуры соответствующих матриц вращения.

Идея преобразования моделей с использованием подобных матриц заключается в эквивалентном преобразовании для линейного слоя нейросети:

$$Y_{output} = X_{input} W^T = X_{input}(RR^T)W^T = (X_{input}R)(R^T W^T)$$

В результате вместо исходных активаций и весов квантуются их преобразованные версии. Для проведения квантования использовался метод GPTQ и модель Qwen3-8B. Качество для каждого формата анализировалось в нескольких экспериментальных условиях: без использования вращений, с применением случайных матриц Адамара, а также с их дообучением. Для оценки эффективности различных конфигураций вычислялись метрики: перплексия (PPL), а

также точность на бенчмарке CEVAL.

В результате проведённых экспериментов было показано, что влияние ортогональных преобразований на качество моделей существенно зависит от используемого формата квантования. Для формата MXFP4 было установлено, что наилучшие результаты достигаются при использовании ортогональных преобразований с блочной структурой матрицы вращения, где размер блока составляет 32. Применение такой структуры позволяет улучшить качество модели по сравнению как с отсутствием вращений, так и с использованием случайных матриц. Анализ полученных результатов показал, что наиболее эффективной оказывается структура вращений, учитывающая особенности распределения активаций и механизм масштабирования, используемый в данном формате.

Полученные результаты демонстрируют, что выбор структуры вращений является важным фактором при применении современных форматов низкоразрядного представления активаций и может существенно влиять на итоговое качество моделей.

Литература

1. Ashkboos, S. [et al.]. QuaRot: Outlier-Free 4-Bit Inference in Rotated LLMs // 2024
2. Liu, Z., Changsheng Z. [et al.]. Spinquant: Llm quantization with learned rotations // 2024
3. Rouhani, B., Zhao R. [et al.]. Microscaling data formats for deep learning // 2023
4. Abecassis, F., Agrusa A. [et al.]. Pretraining large language models with nvfp4 // 2025