Д. С. Жинко

ОПТИМИЗАЦИЯ ИНФЕРЕНСА БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ МЕТОДОМ КВАНТОВАНИЯ ВЕСОВ

В докладе исследуются методы оптимизации вычислительных ресурсов при работе с большими языковыми моделями (LLM). Рассматривается алгоритм 4-битного квантования (QLoRA) как средство снижения требований к видеопамяти. Приводятся результаты сравнительного анализа точности и скорости работы моделей до и после оптимизации, обосновывается эффективность применения данных методов для локального развертывания нейросетей.

I. Введение

Рост параметров современных LLM (например, LLaMA-3, Mistral) приводит к значительным требованиям к аппаратной части. Для моделей с 70 млрд параметров требуется более 140 ГБ VRAM в формате FP16, что делает их недоступными для большинства потребительских GPU. Методы сжатия моделей, такие как квантование, позволяют снизить порог вхождения без критической потери качества генерации.

II. Методы квантования

Квантование – это процесс отображения непрерывных значений весов (Float32/FP16) в дискретный набор (Int8, Int4).

Наиболее эффективным на данный момент является метод QLoRA (Quantized Low-Rank Adaptation) [**?**]. Он использует 4-битный формат NormalFloat (NF4), который оптимизирован под нормальное распределение весов предобученных моделей.

В таблице III показано изменение объема памяти.

Таблица 1 – Потребление VRAM (модель 7B)

Формат	Веса (ГБ)	KB-кэш (ГБ)
FP16	14.0	1.0
Int8	7.0	1.0
4-bit (NF4)	3.8	0.5

III. Экспериментальная часть

Эксперимент проводился на модели LLaMA-2-7B. Использовался бенчмарк MMLU для оценки точности. Скорость инференса замерялась в токенах в секунду (t/s) на GPU RTX 3060 (12GB). Результаты в таблице IV.

Таблица 2 – Сравнение точности и скорости

Метрика	FP16	4-bit NF4
MMLU Score	45.3%	44.8%
Скорость	18 t/s	42 t/s
Перплексия	5.12	5.24

Снижение точности на MMLU составило менее 1%, при этом скорость генерации увеличилась более чем в 2 раза благодаря снижению нагрузки на шину памяти.

IV. Выводы

4-битное квантование NF4 обеспечивает оптимальный баланс между вычислительной эффективностью и сохранением когнитивных способностей модели. Применение данного метода позволяет запускать модели среднего размера (до 13B параметров) на видеокартах с 8-12 ГБ видеопамяти. Дальнейшие исследования будут направлены на изучение 2-битного квантования и методов дистилляции знаний для сверхмалых моделей.

Список литературы

1. Документация PostgreSQL [Электронный ресурс]. – Режим доступа: https://www.postgresql.org/docs. – Дата доступа: 06.05.2026.

2. Hu E. J., Shen Y., Wallis P. et al. LoRA: Low-Rank Adaptation of Large Language Models [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2106.09685. – Дата доступа: 06.05.2026.

3. Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized LLMs [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2305.14314. – Дата доступа: 06.05.2026.

Жинко Даниил Сергеевич, магистрант кафедры интеллектуальных информационных технологий, БГУИР, danik-zhinko@mail.ru

Научный руководитель: Крапивин Юрий Борисович, доцент кафедры интеллектуальных информационных технологий БГУИР, кандидат технических наук, krapivin@bsuir.by