ЕФЕКТИВНА РЕАЛІЗАЦІЯ МАТРИЧНОГО МНОЖЕННЯ ДЛЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ НА CUDA

Сальніков Д.В., Васильченков О.Г.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Великі мовні моделі стали ключовим інструментом у галузі штучного інтелекту, проте їх ефективне використання потребує значних обчислювальних ресурсів та пам'яті, що напряму впливає на ціну використання моделі. Сучасні LLM, такі як GPT-4, мають обсяги вагових коефіцієнтів понад 300 ГБ, що перевищує об'єми пам'яті більшості пристроїв користувачів. Одним із методів зменшення витрат є квантування вагових коефіцієнтів, зокрема до декількох біт на ваговий коефіцієнт [1 – 2].

Сучасні великі мовні моделі побудовані на базі архітектури трансформерів з використанням механізму уваги, що включає велику кількість операцій множення матриць [3]. Поточні ядра множення матриць що використовуються включають етап деквантування. Такий підхід призводить до необхідності використання додаткової пам'яті для проміжних результатів, та додаткових трансферів пам'яті квантованих коефіцієнтів.

Авторами було розроблено ядро множення матриць із квантованими вхідними даними, що є ефективнішим з точки зору часу обчислень та апаратних витрат і досягає продуктивності класичних ядер CUDA (які використовують єдиний тип даних операндів). Запропоноване ядро зменшує об'єм необхідної пам'яті, оскільки відпадає необхідність зберігати розпаковані коефіцієнти на розрахунковому пристрої. Реалізоване ядро використовує кешування вхідних даних та виконує розпаковку коефіцієнтів множення без збереження в проміжкові масиви, що дозволяє ефективно використовувати пропускну здатність пам'яті GPU.

Подальша робота буде спрямована на розширення типів даних що підтримує ядро та покращення алгоритму, зокрема шляхом векторизації його операцій.

Література:

1. The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits Shuming Ma Hongyu Wang Lingxiao Ma Lei Wang Wenhui Wang Shaohan Huang Li Dong Ruiping Wang Jilong Xue Furu Wei

2. Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa.QuIP: 2-bit quantization of large language models with guarantees.CoRR, abs/2307.13304, 2023

3. Ashish Vaswani, Attention Is All You Need, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, arXiv, 2023.