VLLM

Abstract

- Nếu quản lý không hiệu quả bộ nhớ => lãng phí do phân mảnh và trùng lặp
- limiting the batch size

$\Rightarrow$thuật toán PagedAttention

VLLM

- Quản lý bộ nhớ đệm
- Quản lý share memory

Introduction

- Mô hình tạo ra token dựa trên prompt và chuỗi trước đó. Và lặp đi lặp lại => giới hạn GPU, hạn chế throughput =>
- Pageattention
  - chia nhỏ dữ liệu thành các trang (pages) với kích thước cố định => data được chia nhỏ, dễ quản lý
  - Khi cần truy xuất gọi đến các trang liên quan
  - Chỉ lưu trữ các phần dữ liệu cần thiết

Batching Techniques for LLMs

- Vấn đề
  - Request đến không cùng lúc => đợi (naive)
  - Request shape và output khác nhau

- Giải quyết
  - cellular batching
  - iteration-level scheduling

  Các request mới được lặp lại sau khi một request cũ hoàn thành mà không phải đợi toàn bộ batch hoàn thành

Method

- PagedAttention: Chia nhỏ bộ nhớ thành các bộ nhớ con có kích thước cố định => lưu trữ không liền kề
- KV Cache Manager: sử dụng key-value để ánh xạ bộ nhớ
- Thực thi phân tán: sử dụng nhiều GPU để tính toán