Chinese University of Hong Kong in Shenzhen
School of Data Science
Shenzhen, Guangdong, PROC

Efficient Memory Management for Large Language Model Serving

Presented by Xiaozhuang Song

Abstract – Achieving high throughput in serving large language models (LLMs) necessitates processing numerous requests simultaneously. Yet, the fluctuating and substantial memory demands of the key-value (KV) cache for each request pose significant challenges for existing large language model systems. In this presentation, we will delve into a recent study "Efficient Memory Management for Large Language Model Serving with PagedAttention", which is published at SOSP2023, and it tries to address the inefficiencies and challenges associated with dynamically managing the significant memory requirements of LLMs. This study introduces PagedAttention, an attention mechanism inspired by the virtual memory and paging concepts of operating systems. This mechanism significantly substantially reducing memory fragmentation and duplication by managing KV cache memory in non-contiguous blocks. Moreover, the study introduces vLLM, a serving system developed on the PagedAttention framework, which boosts throughput by 2-4 times without compromising latency and outperforms established systems like FasterTransformer and Orca. These improvements are pronounced in dealing with extended sequences, larger model sizes, and intricate decoding algorithms. This presentation aims to highlight the latest developments in LLM serving efficiency and their potential to catalyze further innovation and research in this evolving field.