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

Efficient Serving System for Large Language Model

Presented by Xi Chen

Abstract – Current frameworks often struggle to meet the demanding requirements of serving large language models (LLMs), particularly when faced with workloads that involve long prompts. To achieve high throughput in LLM serving, it is crucial to process a large number of requests concurrently. In this presentation, we will examine a pioneering study titled "DeepSpeed-FastGen: High-throughput Text Generation for LLMs via MII and DeepSpeed-Inference," published in 2024 by the DeepSpeed team. This study introduces a system that utilizes Dynamic SplitFuse, a novel strategy for prompt and generation composition, to significantly enhance performance: up to 2.3 times higher effective throughput, average latency reduced by half, and tail latency at the token level decreased by up to 3.7 times when compared to the latest systems such as vLLM. The paper also offers a detailed benchmarking methodology, analyzes system performance through latency-throughput curves, and explores scalability via load balancing. During the presentation, we will elaborate on the insights from this study and showcase its design revelations.