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

Efficient Serving of Large Generative Language Models

Presented by Yanglin Zhang

Abstract - In the rapidly evolving field of artificial intelligence (AI), generative large lan guage models (LLMs) are at the forefront, revolutionizing how we interact with data. However, their substantial computational and memory requirements present significant challenges for efficient operation, particularly in scenarios demanding quick responses and high capacity. This survey addresses the critical need for effective LLM serving strategies from a machine learning systems perspective, linking stateoftheart AI developments with practical system enhancements. It offers an extensive analysis and presents various innovations in LLM system design, including lowbit quantization, parallel computation, memory management, request scheduling, and kernel optimization. The aim is to provide a comprehensive overview of current trends and future directions in efficient LLM serving.