Topic: Key-Value Cache Compression and Reuse of LLMs

Presenter: Xinye Zhao

Research Questions: When an LLM input contains multiple text chunks, how to quickly combine their precomputed KV caches in order to achieve the same generation quality as the expensive full prefill (i.e., without reusing KV cache)?

Challenge: Fully reusing KV caches from individual chunks ignores cross-attention between them, leading to information loss.

Claims: The absence of cross-attention in full KV reuse causes significant discrepancies in the forward attention matrix. Selectively recomputing part of the tokens layer-wise could fuse those KV-Cache chunks.

Evidence: The model design is based on two insights: **1)** Tokens with high KV deviation remain consistent across layers, recomputing those tokens will dramatically reduce attention score deviation; **2)** Tokens with the highest KV deviations on one layer are likely to have the highest KV deviations on the next layer.

Statistical Analysis: Computing Spearman correlations between adjacent layers to support Insight 1. Benchmarking CacheBlend across multiple tasks and datasets using ROUGE-L and F1 scores to evaluate performance. Selecting full KVCache recomputation and full KVCache reuse as baselines.