FSU | FLORIDA STATE UNIVERSITY

STUDENT SEMINARS @ FSU Computer Science

CRELAY: Composable REpresentative LAYer Attention - Runtime Head Consolidation for LLM Inference

Oteo Mamo

Presentation Abstract

The key-value (KV) cache has become a primary bottleneck for large language model (LLM) inference, limiting memory efficiency and throughput. While Grouped Query Attention (GQA) reduces KV heads in multi-head attention (MHA) through architectural modifications and fine-tuning, it enforces static, factor-based configurations that preclude dynamic, data-driven adaptation and optimization. This work presents CRELAY (Composable Representative Layer Attention), a runtime KV-head clustering framework that reduces KV cache size beyond GQA without modifying model weights. CRELAY analyzes attention patterns from representative prompts to identify redundant KV heads per layer using complementary clustering criteria, then routes all query heads through a compact set of representative KV heads. Importantly, CRELAY preserves all query heads, enabling composability with orthogonal techniques such as token eviction and query-side sparsity. Implemented in vLLM, CRELAY supports non-static, per-layer KV configurations unconstrained by head divisibility. Across LLaMA and Gemma models, CRELAY achieves a 20–25% KV-head reduction with less than 5% accuracy degradation. CRELAY establishes a practical baseline for composable, multi-axis sparsity in LLM inference.

Speaker's Profile

I am a second-year PhD student at Florida State University. I conduct my research in the Computer Architecture and SysTems Research Laboratory (CASTL) under the guidance of Dr. Weikuan Yu. My research interests lie at the intersection of large language models and systems optimization, with a particular focus on improving the memory efficiency of key-value (KV) caches during LLM inference. As models grow larger and context lengths increase, managing memory effectively becomes a critical challenge, and my work aims to design techniques that reduce memory overhead while maintaining performance. Currently, my primary research explores sparsity techniques for long-context summarization tasks. In addition to this work, I have contributed to several related projects. These include research on visual-spatial reasoning in multimodal models, I/O-disaggregated memory storage systems, checkpoint-restart mechanisms, and optimizations for LLM inference pipelines.

Seminar Time: 2026-03-24 Tuesday, 11:45 AM - 12:45 PM

Location (In-Person Only): Room 353 LOVE Building

Contact FSU Student Seminar Organizers

Dr. Yushun Dong

Email: yd24f@fsu.edu