

CMSC828G: Special Topics in Systems for Machine Learning

Long Context Optimizations

Instructor: Yafan Huang



DEPARTMENT OF
COMPUTER SCIENCE

A System Perspective for Long-context Optimization

- Attention Approximation: Reduce the amount of history (or attention) needed.
 - KV Cache Compression: Reduce the memory footprint of retained KV states.
 - Long Context Optimization: Process extremely long streams or sequences under limited resources.
-
- They overlap and complement each other in addressing long-context challenges.

Two Different Meanings of Long Context



- Long-running streaming: Process an indefinitely growing token stream with limited active memory.
 - A chatbot processes millions of tokens over time but only needs recent context to continue.
- Long-range attention: Process a very long sequence while preserving interactions across distant tokens.
 - A model analyzes a million-token codebase where relevant information may appear anywhere, including the beginning.

Student Presentations

- [ICLR'24] Efficient Streaming Language Models with Attention Sinks
- [ICLR'24] Ring Attention with Blockwise Transformers for Near-Infinite Context