Introduction

INTRODUCTION TO LLM
INFERENCE SERVING SYSTEMS

CHUHONG YUAN

Outline

- Course Overview
- Syllabus & Logistics
- LLM Inference Basics
- Homework
- Q & A

Course Overview

Instructor

- Chuhong Yuan
- 5th Year Ph.D. Student
- Research focuses on LLM Inference Systems, particularly P-D Disaggregation and Quantization

Course Information

- 6PM to 7PM ET every Tuesday, except on 10/7 (Fall Break)
- Zoom, the link is on Canvas, and it is the same every time
- Assignments are published and submitted through Canvas
- Discussions can be posted on Ed Discussion
- Prerequisite courses
  - Operating System
  - Machine Learning

What Is This Seminar For?

- Topic: LLM **inference systems**
  - Not for LLM
  - Not for LLM training systems
- Content:
  - Serving infrastructure
  - Single-instance LLM serving optimization
  - Multi-instance LLM serving optimization

What To Learn From This Seminar?

- Basics of LLM inference systems
- Hot topics of research in this area
- System research methodology
- Critical thinking

Syllabus & Logistics

Syllabus

8/19/2025	Introduction
8/26/2025	Serving Infrastructure I
9/2/2025	Serving Infrastructure II
9/9/2025	Serving Infrastructure III
9/16/2025	Single-Instance Serving I
9/23/2025	Single-Instance Serving II
9/30/2025	P-D Disaggregation I
10/7/2025	Fall Break

10/14/2025	P-D Disaggregation II
10/21/2025	P-D Disaggregation III
10/28/2025	Quantization
11/4/2025	Sparse Attention
11/11/2025	Multi-Instance Serving I
11/18/2025	Multi-Instance Serving II
11/25/2025	Multi-Instance Serving III
12/2/2025	Multi-Instance Serving IV

Seminar Format

- 30 - 40 min paper reading
- 10 - 20 min homework review
- 10 min Q & A

Homework Format

- Type (links will be published on Canvas)
  - Code reading
  - Paper review
- Format
  - Summary
  - Questions
  - Critical thinking
  - Paper review

Submission & Grading

- Submission
  - No more than 1 page PDF
  - Font: Times New Roman, 12
  - Line spacing: 1.5
  - Only necessary pictures
- Grading
  - The seminar is pass/fail, as long as your submission is completed, it is fine

Homework Integrity

- Please obey Georgia Tech's Academic Honor Code, no plagiarism
- Use of LLM
  - Sending homework questions or papers to LLM is highly unrecommended
  - Refining homework writing with LLM is fine
  - Asking related technical questions to LLM is fine

LLM Inference Serving Basics

Transformer

- Tokenization: raw text to numbers
- Embedding layer: mapping tokens to vectors -> $X \in R^{L \times d}$
- Linear projection: $Q = XW_Q, K = XW_K, V = XW_V, W_Q, W_K, W_V \in R^{d \times d_h}$
- Attention: $Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_h}}\right)V, output \ \in R^{L \times d_h}$
- MLP (Multi-Layer Perceptron): refine representations
- Output projection: $h \in R^{1 \times d}, W_{vocab} \in R^{d \times V}, P(next\ token) = softmax(hW_{vocab})$
- https://poloclub.github.io/transformer-explainer/

LLM Inference Phases

- KV Cache
  - $Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_h}}\right)V$
  - When a new token is added, we only need to compute its $Q \in R^{1\times d}$, K and V are fixed, so they can be cached
- Prefill
  - Process all input tokens and generate KV Cache
  - Parallelizable, compute-bound
- Decoding
  - Use KV Cache to generate the later tokens one by one
  - Unparallelizable, memory-bound

Challenges in LLM Inference Serving

- Costs
  - Computational cost – limit by GPU capacity
  - Memory cost – KV Cache, which fluctuates and is unpredictable
- Latency
  - Serving-level objectives (SLOs): TTFT (Time To First Token), TPOT (Time Per Output Token), E2E (End To End)
  - Average, Median (P50), Tail (P90, P95, P99)
- Throughput
  - Tokens per second, different in the two phases
- Tradeoffs
  - Latency vs. Throughput: batching strategies, SLO attainment
  - Cost vs. Gain: scheduling strategies

Homework

Start Learning vLLM!

- vLLM
  - https://docs.vllm.ai/en/latest/design/arch_overview.html
  - https://github.com/vllm-project/vllm
- Task
  - Find the codes that correspond to the Figure showing the class hierarchy of vLLM
  - State how the codes work together

Q & A