Yueming Yuan

+1-447-902-6347 | yy28@illinois.edu | https://yueming-yuan.github.io/
Yueming Yuan | yueming-yuan
Urbana, IL - 61801, United States

Research Interests

I'm interested in 1) developing LLM agents for real-world tasks; 2) improving multi-modal understanding ability and efficiency on real-world scenarios.
Previously, I worked on machine learning systems and ML efficiency.

Education

- **University of Illinois Urbana-Champaign** *Aug. 2024 - Present*
  Ph.D. in Computer Science Urbana, IL
- **Zhejiang University** *Aug. 2020 - May 2024*
  B.E. in Electrical and Computer Engineering (GPA: 3.97/4.0) Zhejiang, China
- **University of Illinois Urbana-Champaign** *Aug. 2022 - May 2023*
  Exchange Student Urbana, IL

Publications

[1] **X-MoE: Enabling Scalable Training for Emerging Mixture-of-Experts Architectures on HPC Platforms.** Yueming Yuan, Ahan Gupta, Jianping Li, Sajal Dash, Feiyi Wang, Minjia Zhang. Under review at *SC 2025*; review ratings: 4/4/5 (max rating: 5).

[2] **MiLo: Efficient Quantized MoE Inference with Mixture of Low-Rank Compensators.** Beichen Huang*, Yueming Yuan*, Zelei Shao*, Minjia Zhang. (*equal contribution). Accepted to *MLSys 2025*. [pdf].

[3] **SPLAT: A framework for optimized GPU code-generation for SParse reguLar ATtention.** Ahan Gupta, Yueming Yuan, Devansh Jain, Yuhao Ge, David Aponte, Yanqi Zhou, Charith Mendis. Accepted to *OOPSLA 2025*. [pdf].

Preprints

[1] **FLuRKA: Fast and accurate unified Low-Rank & Kernel Attention**. Ahan Gupta, Hao Guo, Yueming Yuan, Yanqi Zhou, Charith Mendis. [pdf].

Selected Research Experience

- **Research Assistant** *Mar. 2024 - Present*
  Advisor: Minjia Zhang Urbana, IL
  - Large-scale training system for DeepSeek-style MoE: Scale and accelerate **DeepSeek-style Mixture-of-Expert (MoE) training** with up to **1024 GPUs** on the Frontier supercomputer through hybrid parallelism and communication optimizations.
  - MoE compression: Proposed extreme-bit **model weight compression** algorithm for MoE models, leveraging **low-rank compensators** to correct quantization error. Implemented W3A16 GEMM CUDA backend.

- **Undergraduate Research Assistant** *Jan. 2023 - Jan. 2024*
  Advisor: Yanqi Zhou, Charith Mendis Urbana, IL
  - Code generation for regular sparse attention: Developed framework to automatically generate efficient CUDA kernels for **regular sparse attentions**, e.g., blocked sparse attention.
  - Efficient attention mechanism: **approximated linear attention** for extremely long-sequence Transformer training.

Selected Projects

- **Efficient Small Language Model (SLM) Training via Pruning and Distillation** *Jan. 2025 - May 2025*
  Advisor: Fan Lai Urbana, IL
  - Curated high-quality **reasoning** dataset, providing 1% overall improvement in commonsense reasoning tasks.
  - Trained the SLM on a synthesized **function-calling** dataset for Android and deployed the model to mobile devices.

Talks

Efficient Quantized MoE Inference with Mixture of Low-Rank Compensators May. 2025
MLSys Conference Santa Clara, CA

Selected Awards

Academic Scholarship (¥10000), Zhejiang University Oct. 2023
Travel Grant ($1000), MLSys 2025 May. 2025

First Prize in National Olympiad in Informatics in Provinces (NOIP) 2017
Second Prize in Provincial Chinese Chemistry Olympiad (CChO) 2018

Skills

- **Programming Languages:** Python, C/C++
- **Tools & Frameworks:** CUDA, NCCL, PyTorch, JAX
- **Languages:** English, Chinese