Yang Zhou

LinkedIn: linkedin.com/in/yang-zhou
Github: github.com/YangZhou08
Pittsburgh, PA, 15206
Google Scholar: link
Homepage: github.io/yangzhou
Email: yangzho6@andrew.cmu.edu

Education

- **University of Texas at Austin** August 2019 - May 2023
  *Bachelor of Electrical and Computer Engineering; GPA: **3.99**/4.00* ***High Honors***
  - **Graduation Track**: Computer Architecture and Embedded System
  - **Courses**: Computer Architecture, ML Hardware Software Co-design, Data Science Lab, Operating Systems, Algorithms, Embedded System Lab
- **Carnegie Mellon University** August 2023 - Present
  Doctorate in Electrical and Computer Engineering; QPA: 3.84/4.00
  - **Advisor**: Professor Beidi Chen

Research Interests

- Efficient Inference of Large Language Models, Large Language Model Reasoning, Distributed Training of Large Neural Networks

Publications

- "Sirius: Contextual Sparsity with Correction for Efficient LLM" **Yang Zhou**, Zhuoming Chen, Zhaozhuo Xu, Victoria Lin, Beidi Chen. *NeurIPS'24*
- "LLM Inference Unveiled: Survey and Roofline Model Insights" Zhihang Yuan*, Yuzhang Shang*, **Yang Zhou***, Zhen Dong, Zhe Zhou, Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, Yan Yan, Beidi Chen, Guangyu Sun, Kurt Keutzer (2024) preprint, cited 29 times
- "DQRM: Deep Quantized Recommendation Models" **Yang Zhou**, Zhen Dong, Ellick Chan, Dhiraj Kalamkar, Diana Marculescu, Kurt Keutzer (2023) preprint
- "Play It Cool: Dynamic Shifting Prevents Thermal Throttling" **Yang Zhou**, Feng Liang, Ting-wu Chin, Diana Marculescu *DyNN @ ICML'22* ***(oral)***

Academic Experience

- **Infini-AI-Lab, CMU** September 2023 - Present
  Graduate Student *Supervisor: Professor Beidi Chen*
  - **LLM model compression for inference optimization** - Identify the problem of existing Contextual Sparsity techniques in complex reasoning tasks, and solve it by proposing a novel method. Published in NeurIPS 2024
  - Ongoing project on Efficient LLM Inference Scaling
- **Pallas Group, UC Berkeley** May 2022 - May 2023
  Undergraduate Research Assistant *Supervisor: Professor Kurt Keutzer and Dr. Zhen Dong*
  - Lead a project that optimizes the state-of-the-art recommendation model (DLRM) in both inference and training
  - Compress the size of the state-of-the-art recommendation models by 8X (4-bit quantization) **without performance sacrifice**
  - Combine sparsification and quantization to compress gradient during training to alleviate communication overhead (>99% sparsification, 8-bit quantization)
  - Project code and preprint are open-sourced (Github Repo: Deep Quantized Recommendation Model DQRM)
- **Energy-Aware Computing Group (EnyAC), UT Austin** March 2021 - May 2023
  Undergraduate Research Assistant *Supervisor: Professor Diana Marculescu*
  - Lead a project that targets solving Thermal Throttling issues on edge/phone CPUs
  - Observe that edge devices suffer from Thermal Throttling when making continuous ML inference
  - Propose to **dynamically shift** between Dynamic Networks to prevent edge devices from Thermal Throttling
  - Published in 2022 ICML DyNN workshop (oral) see talk here

Honors & Awards

Carnegie Institute of Technology Dean's Fellowship — Awarded 2023-2024
2022 Distinguished College Scholar — Top 4%, University of Texas at Austin
2021 Distinguished College Scholar — Top 4%, University of Texas at Austin
2023 College Scholar — Top 10%, University of Texas at Austin
Engineering Honor Student — Top 10%, Since October 2020

Skills

- **Programming Languages**: Python (Proficient), Java (Familiar), C/C++ (Familiar), Kotlin (Familiar)
- **Frameworks**: PyTorch (Proficient), Android Studio (Proficient)
- **System Implementation Backend**: CUDA (Familiar), Triton (Basic)
- **Hardware Languages**: Verilog (Basic), Vitis HLS on C++ (Basic)
- **Languages**: Chinese (Native), English (Proficient)
- **Swimming**: Swam across the Yangzi River in 2016 at the age of 15

Fun Projects

- **Adversarial Attack on NLP Models**: Explored word-level adversarial attack on BERT models (May 2022)
- **Microcontroller-based Alarm Clock**: TM4C123-based alarm clock with internet support (September 2022)