Zhaofeng Sun

5556 Forbes Avenue, Pittsburgh, PA | (607)-262-7725 | sun-zf22@mails.tsinghua.edu.cn & zs453@cornell.edu

Research Interests: Efficient Machine Learning; RL for LLM post-training

Education

Computer Science, ***Cornell University (2025 Spring Exchange)*** 2025.1-2025.6

- GPA **4.15**/4.00
- **Major Course works:**

  Stochastic Processes, implemented an efficient sampling scheme for discrete distributions with a large state space using Metropolis algorithm and Monte Carlo techniques

Computer Science and Technology, ***Tsinghua University*** 2022.9-2026.8 (expected)

- GPA **3.97**/4.00**,** Ranking **7**/189
- **Major Course works:**

  Computer Organization and Architecture, developed a 5-stage pipelined CPU with machine mode and user mode

  Artificial Neural Network, studied various architectures, from MLP to modern Transformer models, and completed a project on performance optimization and acceleration for *DiffuSeq*

  Numerical Analysis, studied numerical computation methods suitable for computers to solve various common mathematical problems

  Theoretical Computer Science, studied computational models, complexity theory, randomness and computation, cryptography, logic and computation

  Introduction to Artificial Intelligence, trained a 3D Connect Four model using reinforcement learning, capable of defeating nearly all human players

Publications

Model Preserving Adaptive Rounding ***NeurIPS 2025 (expected)***

Albert Tseng, **Zhaofeng Sun**, Christopher De Sa

Short-ARC: Adaptive Reasoning Control to Prevent LLM Overthinking

ICLR 2026 (expected)

Zhaofeng Sun, Zichong Li, Liming Liu, Haoyu Wang, Tuo Zhao

Research Experience

Chris De Sa's Research Group, Cornell University 2025.1-2025.5

- Researching efficient machine learning, particularly extreme low-bit quantization-aware training, aiming to improve inference efficiency while preserving the accuracy of LLaMA models

  Key Learnings:

  - Understanding classical quantization techniques, such as redistributing quantization difficulty from weights to activations
  - Exploring BitNet's novel architecture and its impact on extreme low-bit quantization
  - Studying methods for mitigating outliers and achieving a more uniform weight distribution to enhance quantization performance (QuIP, QuIP#, QTIP)

  Approaches Explored:

  - Implementing scaling techniques and Gaussian preprocessing to stabilize quantization-aware training (QAT)

- Combining BitNet with low-rank decomposition to improve efficiency without sacrificing too much accuracy
- Investigating new backpropagation strategies and loss functions to reduce precision loss in QAT

Jianfei Chen's Research Group, Tsinghua University 2023.10-2025.1

- Investigated the origins of outliers in model training, analyzing how they emerge during the learning process and what they represent computationally. Our goal is to develop effective strategies to mitigate their impact, ultimately improving the efficiency and performance of model quantization

Key Learnings:

- Reviewed research on interpretability of large models, with a focus on analyzing the root causes of outliers
- Gained foundational understanding of MLSys field

Approaches Explored:

- Attempted to reconstruct the model's Linear and Activation layers by attempts like adding trainable channel-wise bias

Beidi Chen's Research Group, Carnegie Mellon University 2025.5-2025.8

- Systematically evaluating the effectiveness of **Sparse Attention** mechanisms in long reasoning scenarios
- Efficient inference for reasoning models

Haoyu Wang and Tuo Zhao's Lab, SUNY Albany and Georgia Institute of Technology
2025.3-2025.7

- Applying reinforcement learning to dynamically control the reasoning length of a model, enabling adaptive thinking and optimizing the trade-off between response length and accuracy
- Designed a length–accuracy performance benchmark for evaluation

Internship

Noah's Ark Lab, Huawei*, Programmer and Testing* 2024.10-2024.12

- Profiling of large model inference on NPU
- Efficient GEMM implementation on NPU

Beijing Institute of Open-Source Chip*, Programmer* 2024.7-2024.10

- Instruction Cache Performance Optimization and chip verification

Awards and Accomplishments

- Champion of Tsinghua University Supercomputer Competition 2024.11
- Toyota Scholarship for outstanding academic achievements and contributions to CS
2023.11
- Tang Ze'sheng Fellowship for leadership and academic excellence
2023.10

Skills

Coding Skills: Proficient in Python, C, and C++, and experienced with deep learning frameworks
GPU Programming: Proficient in writing custom CUDA and Triton kernels, familiar with profiling tools (e.g. NVIDIA Nsight System)
Mathematics: capable of solving problems in calculus, linear algebra, and probability theory encountered in research