Zhaofeng Sun

327 Eddy St. Ithaca, NY | (607)-262-7725 | zs453@cornell.edu | Date of Birth: 4th June 2005

Research Interests: efficient machine learning algorithms, especially in model quantization, distillation, and pruning; RL fine-tuning on LLMs

Education

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

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

Beijing No.8 High School 2019.9-2022.6

- Top 0.2% in College Entrance Examination

Publications

Model Preserving Adaptive Rounding, NeurIPS 2025 (expected)

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

Short-ARC: Adaptive Reasoning Control to Prevent LLM Overthinking, ICLR 2026 (expected)

Zhaofeng Sun, Zichong Li, Haoyu Wang, Tuo Zhao

Research Experience

Chris De Sa's Research Group, Cornell University 2025.1-2025.6

- 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. However, these methods did not yield significant improvements.

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

- Building a benchmark to systematically evaluate the effectiveness of **Sparse Attention** mechanisms

Tuo Zhao's Research Group, Georgia Institute of Technology 2025.3-2025.12 (expected)

- Implementing an adaptive reasoning post-training method using GRPO to control the length of Chain-of-Thought reasoning

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

PyTorch: Proficient in using deep learning frameworks

Coding Skills: Proficient in Python, C and C++

Solid Foundation of Mathematics: Solid mathematical background, with demonstrated ability to apply calculus, linear algebra, and probability theory to solve research-level problems