Zhaofeng Sun

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

Research Interests: Efficient Machine Learning; Reinforcement Learning for LLM Post-training

EDUCATION

Cornell University **Ithaca, NY**
Spring Exchange, Computer Science 01/2025 – 06/2025
- GPA: 4.15/4.00

Tsinghua University **Beijing, China**
Bachelor of Science in Computer Science 09/2022 – 06/2026(expected)
- GPA: 3.97/4.00 | Ranking: 5/187

PUBLICATION

- **Model Preserving Adaptive Rounding**, Albert Tseng, **Zhaofeng Sun**, Christopher De Sa, to be submitted to **NeurIPS 2025**
- **Short-ARC: Adaptive Reasoning Control to Prevent LLM Overthinking, Zhaofeng Sun,** Zichong Li, Liming Liu, Haoyu Wang, Tuo Zhao, to be submitted to **ICLR 2026**

AWARDS AND ACCOMPLISHMENT

- Champion of Tsinghua University Supercomputer Competition 11/2024
- Toyota Scholarship for outstanding academic achievements and contributions to CS (Top 10%) 11/2023
- Tang Ze'sheng Fellowship for leadership and academic excellence (Top 5%) 10/2023
- Shu Tong Scholarship for well-rounded excellence embodying the spirit of "giving back to society." 10/2021

RESEARCH EXPERIENCE

Beidi Chen's Research Group, Carnegie Mellon University 05/2025 – 08/2025
- Conducted a systematic evaluation of different sparse attention methods in different tasks (i.e. retrival tasks: RULER, LongBench, etc., and reasoning tasks: GSM8k, AMC23, AIME24, etc.) and different models (i.e. non-CoT models: Llama3 and CoT models: Qwen3)
- Investigated test-time scaling strategies to improve the performance of reasoning models without additional training; analyzed effects on parallel rollouts with sparse attention in RL sampling
- Identified the need for different landmarks under varying thinking patterns, and proposed a new sparse attention method that adaptively selects key-value tokens based on task demands, enabling better generalization across diverse tasks

Chris De Sa's Research Group, Cornell University 01/2025 – 05/2025
- Developed techniques for low-bit quantization in deep learning, focusing on both post-training quantization (PTQ) and quantization-aware training (QAT) to enhance inference efficiency while maintaining accuracy.
- Implemented outlier mitigation techniques (e.g., QuIP, QuIP#, QTIP) and two-sided Hessian-based analysis to maintain end-to-end model performance.
- Explored layer-wise adaptive bitrates and weight preprocessing methods to achieve more uniform quantization distributions.
- Explored an iterative 2-sided Hessian recovery tuning approach, achieving a ~0.5% improvement in perplexity over the naive 1-sided variant

Jianfei Chen's Research Group, Tsinghua University 10/2023 – 01/2025
- Investigated the origin and computational nature of outliers during neural network training; analyzed when and how outliers emerge and their effects on model behavior.
- Proposed reconstruction methods for Linear and Activation layers, including trainable channel-wise bias insertion to suppress outlier effects and construct outlier-free layers.
- Conducted an in-depth literature review on large model interpretability and outlier phenomena, gaining broad insight into trends and methods in MLSys

PROFESSIONAL EXPERIENCE

Noah's Ark Lab | Research Intern Beijing, China | 10/2024 – 12/2024
- Profiled large model inference on Ascend NPU, analyzing memory bandwidth utilization and compute unit occupancy
- Optimized GEMM kernel on NPU, improving data locality with loop tiling and reducing memory access overhead using shared memory buffering
- Explored quantization-aware acceleration, investigating the impact of mixed-precision arithmetic on inference performance and accuracy

Beijing Institute of Open-Source Chip | Programmer Beijing, China | 07/2024 – 10/2024

- Optimized instruction cache performance, applying prefetching and replacement strategies to reduce latency in ML workloads.
- Conducted hardware-aware model optimization, evaluating cache-aware scheduling techniques for efficient on-chip execution of deep learning models.
- Developed RTL testbenches for chip verification, performing functional and timing validation to ensure hardware reliability in ML accelerators.

SKILLS

Coding Skills: Proficient in Python, C, and C++, and experienced with deep learning frameworks:

- LLM Training Frameworks: PyTorch FSDP, Megatron
- LLM Serving Engines: vLLM, SGLang
- RL Post-training Frameworks: verl

GPU Programming: Proficient in writing custom Triton and CUDA kernels, familiar with profiling tools (e.g. NVIDIA Nsight System)

Mathematics: capable of solving problems in linear algebra and probability theory encountered in research