Yuezhou Hu

huyz21@mails.tsinghua.edu.cn | +86-17748486233 | huyz2023.github.io

Education

Tsinghua University, BS in Computer Science (Expected in June 2025) Sept 2021 – Present

- GPA: 3.8/4.0
- Department of Computer Science and Technology
- **Coursework:** Calculus A(1,2) (4.0), Physics (1,2) (4.0), Linear Algebra (4.0), Advanced Topics in Linear Algebra (4.0), Introduction to Complex Analysis (4.0)
  Introduction to Artificial Intelligence (4.0), Artificial Neural Networks (4.0), Fundamentals of Computer Graphics (4.0), Cybersecurity Fundamentals (4.0), Software Engineering (4.0)

Georgia Institute of Technology, Visiting Student Jul. 2024 – Aug. 2024

- H. Milton Stewart School of Industrial and Systems Engineering

Chengdu No.7 High School Sep. 2018 – Jul. 2021

- Talented Experimental Class

Research Interests

My research interests include **efficient machine learning**, particularly efficient training and inference. Recently, I am focusing on dynamic sparse training, quantization and speculative decoding. I want to break the monopoly of unicorns in AI and make deep learning affordable and accessible for every researcher.

Publications

- **Accelerating Transformer Pre-training with 2:4 Sparsity** [arXiv] [OpenReview] [PDF] [Project page]
  Yuezhou Hu, Kang Zhao, Weiyu Huang, Jianfei Chen, Jun Zhu
  International Conference on Machine Learning (ICML), 2024
- **S-STE: Continuous Pruning Function for Efficient 2:4 Sparse Pre-training** [arXiv] [OpenReview] [Project page]
  Yuezhou Hu, Jun Zhu, Jianfei Chen
  Neural Information Processing Systems (NeurIPS), 2024
- **Pruning Large Language Models with Semi-Structural Adaptive Sparse Training** [arXiv]
  Weiyu Huang, ***Yuezhou Hu***, Guohao Jian, Jun Zhu, Jianfei Chen
  Submitted to AAAI 2025

Experience

Research Intern, Statistical Artificial Intelligence & Learning Group, Tsinghua University, Advisor: Prof. Jianfei Chen, Prof. Jun Zhu May 2022 – Present

- **Accelerating Transformer Pre-training with 2:4 Sparsity**
  - Propose an end-to-end acceleration method for pre-training transformers based on 2:4 sparsity
  - Propose three accuracy-preserving techniques (two for masked decay and one for dense fine-tune) for 2:4 training
  - Analyze and identify two speed bottlenecks (pruning overhead and gated activation functions' overhead) affecting 2:4 training, which is rarely considered by previous works; propose kernel-level accelerated methods to address each of these bottlenecks
  - To be the first report on end-to-end 2:4 acceleration on pre-training transformers; comparable or even superior in accuracy to those trained with dense training methods
  - A delightful start at my first submission to a ML conference!
- **Continuous Pruning Function for 2:4 Pre-training**

- Study STE-based pre-training methods from the optimization perspective
- Point out that STE-based pre-training defines a discontinuous loss function, which existing optimization theory and algorithms cannot handle
- Reveal several intriguing phenomena highlighting the difficulty of discontinuous optimization, (incorrect descending direction, unpredictable amount of descent, oscillation)
- Introduce a new pruning function which is continuous but can still generate N:M sparse weights at all times to prune weights in 2:4 training
- Surpass previous 2:4 pre-training recipes on a wide range of tasks

Research Intern, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Advisor: Prof. Tuo Zhao — Jul. 2024 – Aug. 2024

- **Selective Knowledge Distillation for Efficient Speculative Decoders**
  - Aim at enhancing predictive precision of the speculative decoder
  - Identify "easy" and "hard" tokens for the decoder to speculate; propose a workflow to differentiate them based on popular knowledge distillation methods
  - Achieve and surpass previous SOTA methods on different tasks

Projects

Image Generation in a Prescribed Aesthetic — Spring 2023

- Course project of Computer Graphics
- Generate images in a given style based on CGGAN, using Python and Jittor

Multi-User Video Sharing Platforms — Spring 2023

- Course project of Software Engineering
- Build and deploy a multi-user video and social platform from scratch, using Python and Django

Technical Skills

Deep learning programming: Python, Pytorch

GPU Programming: OpenAI Triton, C++, CUDA

Others: Docker, Django, Rust

Language Skills: TOEFL 104 (R:28/L:27/S:23/W:26)

Extracurricular Activities

- Volunteer in department's student assistance program (Program Buddy), assisting first and second year undergraduates with programming — Mar. 2024 – Present
- Vice president of Tsinghua University Summer School for Undergraduate Applicants — Aug. 2022
- Volunteer in the Undergraduate Admissions Office of Tsinghua University — Aug. 2022

Honors

- Academic Preeminence Scholarship — Fall 2024
  5000 CNY (about 690 USD)
- Innovation Future Scholarship — Fall 2024
  25000 CNY (about 3450 USD) for a group of three students