HAOZE HE

Pittsburgh, PA | (646) 673-2552| haozeh@cs.cmu.edu
Homepage: https://hectorhhz.github.io| GitHub: https://github.com/HectorHHZ

Haoze He is currently pursuing Ph.D. degree at Carnegie Mellon University, School of Computer Science, specializing in machine learning and computer software engineering. His research focuses on **large language models (LLMs), large language model systems (LLMSys), parameter-efficient fine-tuning (PEFT) of LLMs, and distributed machine learning systems (distributed MLSys)**. Haoze's work encompasses two main directions: first, he aims to develop efficient systems for LLMs and ML algorithms to enhance performance during training, fine-tuning, and inference; his second research interest is 'LLM Anatomy', focusing on interpreting and improving LLM architectures using system design methods.

EDUCATION

Carnegie Mellon University (CMU) — School of Computer Science
Doctor of Philosophy, Specialized in Machine Learning and Computer Software Engineering — *Aug 2023 – May 2026*
Cumulative GPA: 4.19/4.0

New York University (NYU) — Tandon School of Engineering
Master of Science, Specialized in Computer Engineering — *May 2023*
Cumulative GPA: 3.93/4.0

The Chinese University of Hong Kong (CUHK) — School of Science and Engineering
Bachelor of Science, Specialized in Computer Science and Engineering — *May 2020*

PUBLICATIONS

- **Haoze He**, Juncheng Billy Li, Xuan Jiang, Heather Miller, "*Sparse Matrix in Large Language Model Fine-tuning*", International Conference on Learning Representations (ICLR), Submitted, Oct.2024
  - Proposed a parameter-efficient fine-tuning algorithm to overcome the plateau issue seen in LoRA/DoRA, achieving a 14.6× speedup with similar performance to full fine-tuning, significantly surpassing LoRA/DoRA results while maintaining superior computational efficiency and memory usage.
- **Haoze He**, Jing Wang, Anna Choromanska, "*Adjacent Leader Decentralized Stochastic Gradient Descent*", European Conference on Artificial Intelligence (ECAI), Accepted, June.2024
  - Developed a novel Adjacent Leader decentralized SGD with a gradient step method and a dynamic topology network, significantly accelerating training convergence, enhancing performance for lower-degree workers, and increasing final test accuracy by 5.8% for weaker workers and 2.1% in the output model when training ResNet-50 on CIFAR-10.
- Haoran Zhu, **Haoze He**, Anna Choromanska, Satish Ravindran, Binbin Shi, Phil Chen, Colin Decourt, "*Multi-View Radar Autoencoder for Self-Supervised Automotive Radar Representation Learning*", IEEE Intelligent Vehicles Symposium (IEEE IV), Accepted, Mar.2024
  - Developed MVRAE, the first self-supervised representation learning method for multi-view radar data, addressing severe miscalibration in radar semantic segmentation by refining the fine-tuning loss function, and demonstrated that MVRAE significantly enhances label efficiency and performance in downstream tasks.
- **Haoze He**, Parijat Dube, *"RCD-SGD: Resource-Constrained Distributed SGD in Heterogeneous Environment via Submodular Partitioning"*, International Conference on Computer Vision (ICCV) Workshops, Accepted, Jul.2023
  - Developed the Resource-Constrained SGD algorithm to partition datasets across workers while ensuring similar class-level feature distribution, addressing the straggler problem and reducing computational complexity, achieving up to 32% speedup in wall-clock time and improving final model accuracy by 1.1% compared to SOTA baselines.
- Chaoxun Guo, Zhixing Jiang, **Haoze He,** David Zhang, *"Pulse Signal Acquisition and Analysis for Disease Diagnosis: A review"*, Computers in Biology and Medicine, Accepted, Nov.2021

- Reviewed recent advances in computational pulse diagnosis (CPD), including pulse signal acquisition, preprocessing, feature extraction, and recognition, and introduced a benchmark for evaluating CPD methods.

- **Haoze He,** Parijat Dube, "*Accelerating Parallel Stochastic Gradient Descent via Non-blocking Mini-batches*", The Association for the Advancement of Artificial Intelligence (AAAI), Submitted, Aug.2024
  - Proposed a non-blocking algorithm for distributed machine learning that accelerates convergence, reduces delays, and addresses the straggler problem, extendable to most (de)centralized SOTA algorithms and federated learning, with superior efficiency and convergence compared to MATCHA and D-PSGD, and demonstrated a 2× speedup.

ACDEMIC BLOG POST

- Peter Zhong, **Haoze He**, Omar Khattab, Christopher Potts, Matei Zaharia, Heather Miller, "*A Guide to Large Language Model Abstractions*", Jan.2024

TEACHING

- **Large Language Model System** (CMU-11868, Spring 2025)

SERVICE

- **Reviewer:** International Joint Conference on Neural Networks (IJCNN 2025)
- **Reviewer:** International Conference on Learning Representations (ICLR 2025)
- **Reviewer:** International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022, 2023, 2024)
- **Reviewer:** International Conference on Computer Vision (ICCV 2023) Workshops

PATENTS

- **Haoze He**, Rui Huang, Xiang Zhang, *Biometrically Enhanced Intelligent Smartphone Locking Control System*, China, 201811345387.9, 2019-03-01.
- Zhixia Zheng, **Haoze He**, Saiqin Huang, *Anode Bonding Device*, China, 201720512308.3, 2018-06-12.

SKILLS

- **Programming Languages:** Proficient in Python, Java, C, Cuda , C++, MATLAB, and R
- **Open-Source Software Libraries:** PyTorch, DeepSpeed, Transformer, accelerate, Unsloth, and MPI4PY