HAO ZHANG

402 West Dayton Street ◇ Madison, United States

(+1) 669-338-7703 ◇ sufeidechabei@gmail.com

Homepage ◇ Google Scholar ◇ GitHub

EDUCATION

UW-Madison — Madison, US
Department of Computer Science. — 2022-Present
PhD student supervised by Prof.Suman Banerjee.

Chongqing University Of Posts And Telecommunications — Chongqing, China
Department of Electronic Engineering — 2016-2021
Bachelor Degree

RESEARCH INTERESTS

. LLM Deployment.

. Machine Learning.

RESEARCH WORK

1. Yilong Li, Jingyu Liu, **Hao Zhang**, M Badri Narayanan, Utkarsh Sharma, Shuai Zhang, Pan Hu, Yijing Zeng, Jayaram Raghuram, Suman Banerjee. *"PalmBench: A Comprehensive Benchmark of Compressed Large Language Models on Mobile Platforms."* ICLR 2025

2. Anran Wang, Maruchi Kim, **Hao Zhang**, Shyam Gollakota. *"Hybrid Neural Networks for On-device Directional Hearing."* AAAI 2022

3. Mufei Li, **Hao Zhang**, Xingjian Shi, Minjie Wang, Zheng Zhang. *"A Statistical Characterization Of Attentions In Graph Neural Networks."* ICLR 2019 (Representation Learning on Graphs and Manifolds Workshop)

4. Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, Ziyue Huang, Qipeng Guo, **Hao Zhang**, Haibin Lin, Junbo Zhao, Jinyang Li, Alexander Smola, Zheng Zhang. *"Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs."* ICLR 2019 (Representation Learning on Graphs and Manifolds Workshop)

5. Quanshi Zhang, Yingnian Wu, **Hao Zhang**, Songchun Zhu. *"Mining deep And-Or object structures via cost-sensitive question-answer-based active annotations."* Computer Vision and Image Understanding

6. Mufei Li, **Hao Zhang**, Xingjian Shi, Minjie Wang, Yixing Guan, Zheng Zhang. *"Characterize and Transfer Attention in Graph Neural Networks."*

7. **Hao Zhang**, Giacomo Giuliari, Adrian Perrig. *"Data Exfiltration Detection: A Learning Game."* Semester Project

RESEARCH EXPERIENCE

UW-Madison September 2022 - Present
Research assistant advised by Prof. Suman Banerjee

· Work on deploying LLM on power-constrained devices.
· Quantization and benchmark platforms for evaluating LLM on mobile devices.

Network Security Group at ETH Zurich August 2021 - November 2021
Research assistant advised by Prof. Adrian Perrig

· Work on using deep learning algorithms for simulating the defense and attacking process for IoT devices and using learning algorithms to detect exfiltration for IoT devices.
· Use deep neural networks for flow size prediction.
· Collect TCP packet dataset and use it to train a flow size predictor, then deploy the flow size predictor in multi-path sockets to select the best-suited path, improving networking system performance.

UW Networks and Mobile Systems Lab July 2020 - August 2021
Remote visiting student advised by Prof.Shyam Gollakota

· Work on developing and deploying AR application on wearable devices like deploying face detection algorithm to VR headset and interact with eve tracking function in VR device.
· Explore multi-channel real-time speech separation and help to replicate state-of-the-art real-time speech separation baselines.
· Simulate synthetic audio dataset and help to validate Beamforming for speech separation with angle information.

AWS AI Lab Shanghai March 2019 - June 2019
Visiting guest

· Help to build a deep learning system for graph-based models.
· Work on the graph-based model benchmarks.
· Propose algorithms that can sparsify the molecule dataset according to the edge attention. .

Learning and Vision Lab at National University of Singapore June 2019 - August 2019
Research assistant

· Work on NeurIPS 2019: MineRL Competition and our team get 4th place.
· Work on distributed reinforcement learning.
· Combine PPO with Vtrace (an off-policy correction method) to improve the performance of PPO in MineCraft and Impala reproduction.

OPENSOURCE PROJECTS

GluonCV: GluonCV provides implementations of state-of-the-art (SOTA) deep learning algorithms in computer vision. It aims to help engineers, researchers, and students quickly prototype products, validate new ideas, and learn computer vision. (**5k stars** in GitHub) .

DGL (Deep Graph Library): DGL is a Python package dedicated to deep learning on graphs, built atop existing tensor DL frameworks (e.g. Pytorch, MXNet) and simplifying the implementation of graph-based neural networks.(**13k stars** in Github).

TECHNICAL STRENGTHS

Computer Languages	Python, C/C++, MATLAB
Software & Tools	Git, Latex, Mxnet, Pytorch, Ray, Tensorflow, Caffe