Yefan Zhou

yefan.zhou.gr@dartmouth.edu | Homepage | Google Scholar | Linkedin | Hanover, NH, 03755 | 510-809-5378

Education

Dartmouth College — Hanover, NH
Ph.D. candidate in Computer Science — *Sep. 2023 – present*
Advisor: Prof. Yaoqing Yang and Prof. Michael Mahoney
Research Area: LLM evaluation and reasoning, Model diagnostic

University of California, Berkeley — Berkeley, CA
Master in EECS; Major GPA: 4.0/4.0 — *Aug. 2021 – Dec. 2022*
Advisor: Prof. Michael Mahoney
Research Area: Pruning for model efficiency

University of California, Berkeley — Berkeley, CA
Exchange Student; GPA: 4.0/4.0 — *Jan. 2019 – May. 2019*

Southeast University — China
B.Eng in Information Engineering; GPA: 3.7/4 — *Aug. 2016 – Jun. 2020*

Research Summary

- **LLM Reasoning and Algorithm Discovery**: Developed an approach for discovering linear algebra algorithms (e.g., linear system solvers) using LLM-based multi-step inference guided by an evolutionary search pipeline. Incorporated Monte Carlo Tree Search (MCTS) to generate synthetic reasoning steps and distilled the MCTS policy into the LLM model for enhanced performance.
- **LLM Quality Evaluation**[4, 7, 8]: Designed a data-free weight matrix analysis method to evaluate LLM training quality at both layer-wise and model-wise levels. Validated the approach by applying it to an adaptive learning rate optimizer, achieving balanced layer quality and improved fine-tuning performance.
- **LLM Pruning**[1, 5]: Developed a layer-wise compression method that automatically allocates compression factors (e.g., sparsity, quantization precision) to transformer layers. Applied to models like LLaMA, OPT, Mistral, pruning LLaMA-7B to 80% sparsity achieved a 3.06× speedup on CPUs. Enhanced structured/semi-structured pruning and mixed-precision quantization.
- **Transparency and Explainability**[3, 10]: Proposed a post-training diagnostic method to identify pre-trained model failures (e.g., insufficient model size) without access to the training configuration, streamlining hyperparameter tuning.

[·] refers to publications listed below.

Publication

Selected first-author paper:

1. {H. Lu*, **Y. Zhou***}, S. Liu, Z. Wang, M. W Mahoney, Y. Yang "AlphaPruning: Using Heavy-Tailed Self Regularization Theory for Improved Layer-wise Pruning of Large Language Models" (**NeurIPS 2024**)
   LLM pruning · Efficient inference

2. {H. Lu*, X. Liu*, **Y. Zhou***, Q. Li*}, H. Yang, Y. Yan, K. Keutzer, M. W. Mahoney, Y. Yang "Sharpness-diversity tradeoff: improving flat ensembles with SharpBalance" (**NeurIPS 2024**)
   Training data selection · Ensembling · Out-of-distribution

3. {**Y. Zhou***, J. Chen*}, Q. Cao, K. Schürholt, Y. Yang "MD tree: a model-diagnostic tree grown on loss landscape" (**ICML 2024**)
   Model selection · Scaling law · Hyperparameter tuning

4. {**Y. Zhou***, T. Pang*}, K. Liu, C. H Martin, M. W Mahoney, Y. Yang "Temperature Balancing, Layer-wise Weight Analysis, and Neural Network Training" (**NeurIPS 2023 Spotlight**)
   NN optimizer · Efficient training · Layer quality analysis

5. **Y. Zhou**, Y. Yang, A. Chang, M. W Mahoney "A Three-regime model of Network Pruning" (**ICML 2023**)
   NN pruning Model selection Losslandscape analysis

6. **Y. Zhou**, Y. Shen, Y. Yan, C. Feng, Y. Yang "A Dataset-Dispersion Perspective on Reconstruction Versus Recognition in Single-View 3D Reconstruction Networks" *2021 International Conference on 3D Vision* (**3DV 2021**) Image-to-3D, 3D reconstruction

Collaborating or advising paper:

7. {Z. Liu*, Y. Hu*}, T. Pang, **Y. Zhou**, P. Ren, Y. Yang "Model Balancing Helps Low-data Training and Fine-tuning" (**EMNLP 2024 main Oral**)
   LLM fine-tuning Low-resource training

8. P. Qing, C. Gao, **Y. Zhou**, X. Diao, Y. Yang, S. Vosoughi "AlphaExpert: Assigning LoRA Experts Based on Layer Training Quality" (**EMNLP 2024 main**)
   LLM fine-tuning Mixture-of-expert

9. X. Zhu, **Y. Zhou**, Y. Fan, J. Chen, M. Tomizuka "Learn to Grasp with Less Supervision: A Data-Efficient Maximum Likelihood Grasp Sampling Loss" *2022 International Conference on Robotics and Automation* (**ICRA 2022**) 3D understanding Decision making for robotics

10. K. Schürholt, L. Meynent, **Y. Zhou**, Y. Yang, D. Borth "A Model Zoo on Phase Transitions in Neural Networks" (Preprint)

Professional Experience

Research Engineer, International Computer Science Institute — Berkeley, CA
supervised by **Prof. Michael Mahoney** — *Jan. 2023 – Jun. 2023*

- Researched efficient optimization method for deep neural network.
- Researched ensembling methods for improving the OOD robustness of CV models.
- Developed backdoor detection methods to enhance AI model safety.

Graduate Research Assistant, Sky Computing Lab (RISELab), UC Berkeley — Berkeley, CA
advised by **Prof. Michael Mahoney** — *Aug. 2021 – Dec. 2022*

- Researched neural network pruning for CNNs and Transformers.

Services and Award

Reviewers: ICLR 2025-2024, CVPR 2024-2025, NeurIPS 2023, AAAI 2024, ICML 2024, CPAL 2024, IROS 2022, TMLR

Talk

* Invited talk at AI-TIME, "Phase transition, loss landscape and model diagnostics", Jan., 2024.
* Invited talk at UC Berkeley/ICSI TrojAI onsite, "Layer-wise Weight Analysis, and Neural Network Training" Oct. 2023
* Invited talk at UC Berkeley/ICSI TrojAI onsite, "A Three-regime model of Network Pruning" Mar. 2023

Award: ICML 2024 Scholar Award, NeurIPS 2023 Scholar Award

Teaching (Head TAs): CS70: Foundations of Applied Computer Science (Dartmouth College Spring 2024)

Skills

Programming Language: Python, Java, SQL, MATLAB, JAX, CUDA

Developer Tools: PyTorch, Ubuntu, MujoCo, ROS, PyBullet, Slurm, PyRender, Open3D