Yufa Zhou

yufazhou@seas.upenn.edu
Tel.: +1 445 260 7203
Homepage

Education

University of Pennsylvania, Philadelphia, United States — Aug. 2023 — May. 2025 (expected)
M.S.E. in Scientific Computing — Current GPA: 3.96/4.00

- Selected Courses: CIS 5450 Big Data Analytics (A), CIS 5810 Computer Vision & Computational Photography (A), ESE 5460 Principles of Deep Learning (A), ESE 6050 Modern Convex Optimization (A), ESE 6500 Learning in Robotics (A-), STAT 5330 Stochastic Processes (A)

Wuhan University, Wuhan, China — Sep. 2019 — Jul. 2023
B.E. in Engineering Mechanics — GPA: 3.68/4.00

- Thesis: *Solving the Plane Thermal Stresses Problem using Deep Operator Network*
- Honors: Third Prize Scholarship (Top 30%), 2020 and 2021; Guangzhou Alumni Association Scholarship Grant, 2020; Jucheng Self-improvement Scholarship, 2022; Mitacs Globalink Research Internship Award, 2022
- Selected Courses: Advanced Mathematics (91), Linear Algebra (97), Probability and Mathematical Statistics (98), Computational Method (99), Machine Learning and Pattern Recognition (93), Complex Function & Integral (98)

Research Interests

- Large Language Models (LLMs): LLM Understanding, LLM Accleration, Safe Generative AI

Research Experiences

Mentors: **Zhenmei Shi**, **Yingyu Liang**, and **Zhao Song** — Remote, Mar. 2024 — Nov. 2024

- Multi-Layer Transformers Gradient Can be Approximated in Almost Linear Time (OPT @ NeurIPS 2024) [Paper]
  We prove that gradients in multi-layer transformer models can be computed in almost linear time $n^{1+o(1)}$ using a novel fast approximation method with polynomially small error, overcoming the quadratic complexity bottleneck of self-attention and enabling more efficient training and deployment of long-context language models with general loss functions and common sub-modules like residual connections, causal masks, and multi-head attention.

- Differential Privacy of Cross-Attention with Provable Guarantee (SafeGenAi @ NeurIPS 2024) [Paper]
  We present the first differential privacy (DP) data structure for cross-attention modules—securing sensitive information in key and value matrices across AI applications like retrieval-augmented generation and guided stable diffusion—with theoretical guarantees on privacy and efficiency, robustness to adaptive attacks, and potential to inspire future privacy designs in large generative models.

- Beyond Linear Approximations: A Novel Pruning Approach for Attention Matrix (ICLR 2025) [Paper]
  We introduce a novel LLM weight pruning method that directly optimizes for approximating the non-linear attention matrix—with theoretical convergence guarantees—effectively reducing computational costs while maintaining model performance.

- Fine-grained Attention I/O Complexity: Comprehensive Analysis for Backward Passes [Paper]
  We establish tight I/O complexity bounds for attention mechanisms in large language models across small and large cache sizes—confirming FlashAttention's optimality in large caches, improving algorithms for small caches, extending analysis to sparse attention, and offering insights for efficient LLM training and inference.

Projects

- GPT2-tinytiny Nov. 2023 — Dec. 2023
  - Develop and fine-tune a GPT2 model as part of the ESE 5460: Principles of Deep Learning course at the University of Pennsylvania with the code and detailed report available on GitHub.
  - Implement GPT2 from scratch in PyTorch, pre-train on the WikiText dataset, and perform supervised fine-tuning (SFT) on the LIMA dataset using Low Rank Adaptation (LoRA). Evaluate and analyze the model's completions for given prompts.
- Dog Behavior Classification with Time Series Movement Dataset Nov. 2023 — Dec. 2023
  - Conduct a data science project for CIS 5450: Big Data Analytics at the University of Pennsylvania and participate in a Kaggle competition with the code and slides available on GitHub.
  - Preprocess accelerometer and gyroscope data collected from dogs' necks and backs. Carry out Exploratory Data Analysis (EDA) and feature engineering for data cleaning and transformation. Leverage Plotly for visualization, employ resampling for uniform data distribution, and implement multiple models (random forest, GRU, ResNet) for classification. Achieve a 93.6% accuracy, outperforming the original paper's SVM-based accuracy of 91.4%.

Selected Publications

Accepted in the peer-reviewed journals and conferences:

1. Yingyu Liang, Zhenmei Shi, Zhao Song, **Yufa Zhou**. (alphabetical order) "Tensor attention training: Provably efficient learning of higher-order transformers." In *NeurIPS 2024 Workshop: Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2024. [Paper]
2. Yingyu Liang, Zhenmei Shi, Zhao Song, **Yufa Zhou**. (alphabetical order) "Differential Privacy of Cross-Attention with Provable Guarantee." In *NeurIPS 2024 Workshop: Safe Generative AI*, 2024. [Paper]
3. Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, **Yufa Zhou**. (alphabetical order) "Multi-Layer Transformers Gradient Can be Approximated in Almost Linear Time." In *NeurIPS 2024 Workshop: Optimization for Machine Learning*, 2024. [Paper]
4. Xuan Shen, Zhao Song, **Yufa Zhou**, Bo Chen, Yanyu Li, Yifan Gong, Kai Zhang, Hao Tan, Jason Kuen, Henghui Ding, Zhihao Shu, Wei Niu, Pu Zhao, Yanzhi Wang, Jiuxiang Gu. "LazyDiT: Lazy Learning for the Acceleration of Diffusion Transformers." In *AAAI*, 2025. [Paper]
5. Xuan Shen, Zhao Song, **Yufa Zhou**, Bo Chen, Jing Liu, Ruiyi Zhang, Ryan A. Rossi, Hao Tan, Tong Yu, Xiang Chen, Yufan Zhou, Tong Sun, Pu Zhao, Yanzhi Wang, Jiuxiang Gu. "Numerical Pruning for Efficient Autoregressive Models." In *AAAI*, 2025. [Paper]
6. Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, **Yufa Zhou**. (alphabetical order) "Looped relu mlps may be all you need as practical programmable computers." In *AISTATS*, 2025. [Paper]
7. Yingyu Liang, Jiangxuan Long, Zhenmei Shi, Zhao Song, **Yufa Zhou**. (alphabetical order) "Beyond Linear Approximations: A Novel Pruning Approach for Attention Matrix." In *ICLR*, 2025. [Paper]

Preprints:

1. Xiaoyu Li, Yingyu Liang, Zhenmei Shi, Zhao Song, **Yufa Zhou**. (alphabetical order) "Fine-grained Attention I/O Complexity: Comprehensive Analysis for Backward Passes." arXiv preprint arXiv:2410.09397, 2024. [Paper]

Academic Services

Conference Reviewer: ICLR 2025.

Skills

Techniques: Natural Language Processing, Large Language Models, Generative AI
Programming: Python (PyTorch, Jax), MATLAB, Linux, SQL, Data Analysis, LaTeX