YIFEI ZHANG

yfeizhang.github.io | github.com/yfeizhang | yzhang2536@wisc.edu | (+1) 608-698-7396

Education

University of Wisconsin-Madison Madison, WI

Bachelor of Science, Computer Science (Distinction in the Major) *Jan. 2021 - Aug. 2023*

- GPA: 4.0/4.0
- Advisor: Prof. Frederic Sala

Research Interests

I am broadly interested in **machine learning on statistical inference** and **deep learning on large-scale efficient training** from theoretical and algorithmic perspectives. I also hold a great passion in **AI for Science** and **Machine Learning System Acceleration**.

Publications

*Equal contribution

[1] InRank: Incremental Low-Rank Learning. **Yifei Zhang***, Jiawei Zhao*, Beidi Chen, Florian Schäfer, Anima Anandkumar. 2023. Preprint. (Short version presented in ICML 2023 ES-FoMO)

[2] Incremental Fourier Neural Operator. Jiawei Zhao, Robert Joseph George, **Yifei Zhang**, Zongyi Li, Anima Anandkumar. 2022. NeurIPS 2022 AI4Science.

Research Experience

Incremental Learning to Accelerate GPT2 Pretraining June 2022 - Oct. 2023

Computing + Mathematical Sciences Department, California Institute of Technology

Advisor: Prof. Anima Anandkumar, Prof. Beidi Chen and Dr. Jiawei Zhao

- Successfully generalized low-rank bias of gradient descent to practical fully connected neural network from theoretical work of matrix/tensor factorization ([ICLR '21], [ICML '21] & [NeurIPS '19]).
- By applying cumulative weight updates, generalized from infinitesimal initialization to any other initializations. Theoretically, helped to prove that cumulative weight updates follow an incremental low-rank trajectory.
- Empirically, demonstrated that the incremental learning holds on a broad range of neural networks (e.g., transformers) and standard training algorithms (e.g., SGD, Adam).
- Evaluated InRank on GPT-2 pretraining from scratch, indicating that InRank achieves comparable prediction performance as the full-rank counterpart while requiring at most 33% of the total ranks.
- Motivated by the exponential increase of singular values and mature DL training acceleration framework DeepSpeed, I revised InRank into an efficient version with only training incremented ranks at initial stage.
- Finally, achieved superior performance up to 0.5 perplexity lower compared with baseline, with at most 38% training time reduction and 42% memory usage reduction on GPT-2 medium and GPT-2 large pretraining.
- Code publicly available with 200+ Github stars [S1] and produced publication [1] [2].

Data-centric Machine Learning Jan. 2022 - Present

Department of Computer Science, University of Wisconsin-Madison

Project 1: Tensor Decomposition on Weakly-Supervised Learning

Advisor: Prof. Frederic Sala

- Motivated by [JMLR '14], reformulated problem of weakly-supervised learning (WSL) into multi-view models. The reformulated model has several desirable properties (e.g., readily accept soft labels).
- Utilizing tensor structure in reformulated model, implemented algorithm of tensor power method to recover accuracies of weak supervision sources and conducted comparison on benchmark WRENCH .
- The work will be drafted as a workshop paper.

Project 2: Optimal Transport on Data Selection

Advisor: Prof. Frederic Sala and Dr. Jieyu Zhang

- Motivated by the sensitivity property to abnormal data of OT [JAMS '21], proposed an augmented algorithm to obtain lower-ranking data valued by OT at first and then reverse the selection by targeting on these low-ranking data.
- By amplifying in-distribution data selection at the second time, the new algorithm consistently surpasses baseline LAVA [ICLR '23] up to 3% precision of selection.
- The work will be drafted as a full conference paper.

Directed Study

Department of Electrical & Computer Engineering, University of Wisconsin-Madison

Project 1: Gaussian Mixture Model on Speech Command Recognition *Dec. 2021 - Jan. 2022*

Advisor: Prof. Matthew Malloy

- Preprocessed data by Mel spectrogram and deployed Quadratic Discriminant Analysis on dataset of speech_commands_v0.02. With regularization for covariance estimation, achieved 68.25% accuracy.
- Based on [arXiv '18], implemented convolutional architecture to do embedding and improved the performance to 84.6% accuracy.

Project 2: Preference Modeling and Crowdsourced Clustering *Sep. 2021 - Dec. 2021*

Advisor: Prof. Ramya Vinayak

- According to [NeurIPS '16], implemented generative graph model to illustrate the performance of triplewise comparisons better than pairwise comparisons, given fixed cost of entropy.
- Implemented salient feature preference model in [ICML '20], conducted sample complexity analysis and through simulation experiments, validated model's capability to capture human irrational behavior.

Volunteer Experience

Popular Science Channel Feb. 2020 - May 2020

- During the pandemic, started up one online channel in YouTube, to interpret the news of public concern based on mathematical knowledge and computer simulation techniques, aiming to ease prevalent anxiety.
- Published five videos (7-8 min for each) in total which received dozens of "like"s and praising comments, and two of videos got more than 10k views.

Open-source Software

[S1] *InRank*. URL: https://github.com/jiaweizzhao/InRank. **200+ GitHub Stars**.

Teaching

Peer Mentor, CS 540 Introduction to Artificial Intelligence, Spring 2023, Fall 2022, Spring 2022

– Positive evaluations from students: "Yifei is very responsive and friendly."

Professional Services

Reviewer: Journal of Machine Learning Research (JMLR)

Relevant Courses

Graduate CS: Matrix Methods in Machine Learning; Probability Theory and Information Theory in Machine Learning; Learn More Weakly-Supervised Learning

Undergraduate CS: Artificial Intelligence, Numerical Analysis, Algorithm, Digital Electronics, Computer Architecture, Operating System, Programming Language and Compiler

Math: Calculus, Ordinary Differential Equation, Linear Algebra, Probability Theory, Statistical Methods, Mathematical Logic, Set Theory, Combinatorics, Modern Algebra

Technical Skills

Programming: Python, Java, C, C++, MATLAB, LaTeX, JavaScript, HTML/CSS, Hydra

Data Analysis: NumPy, scikit-learn, Matplotlib, PyTorch, Weights&Biases, PyTorch Lightning

Honors and Awards

Dean's List (three semesters), University of Wisconsin–Madison

Undergraduate Scholarship for Summer Study ($1,000) (twice), University of Wisconsin-Madison