Zhuoyan Xu

(+1) 608-960-1191 | zhuoyan.xu@wisc.edu | LinkedIn | GitHub | HomePage

Education

University of Wisconsin-Madison — Madison, WI
Ph.D. in Statistics (Co-advised by: Yingyu Liang, Yin Li, Yiqiao Zhong) — Sep. 2020 - Fall 2025 (expected)
M.S. in Computer Science — Sep. 2021 - May 2023

Wuhan University — China
B.S in Statistics — Aug. 2015 – Jun. 2019

Research Interests

My research interest mainly focus on Foundation Models (including large vision, language and multimodal models). I am interested in investigating and **analyzing behavior of foundation models**, aiming to **enhance their adaptation** to downstream tasks with improved accuracy and greater efficiency.

Recent Publications

- Zhuoyan Xu*, Khoi Duc Nguyen*, Preeti Mukherjee, Somali Chaterji, Saurabh Bagchi, Yingyu Liang, Yin Li. AdaLLaVA: Learning to Inference Adaptively for Multimodal Large Language Models .(* denotes equal contribution).
- Zhuoyan Xu, Haoyang Fang, Boran Han, Bonan Min, Bernie Wang, Shuai Zhang. TabRAG: Efficient Table Retrieval and Understanding with Large Multimodal Models. Work done during internship at AWS.
- Yingyu Liang*, Heshan Liu*, Zhenmei Shi*, Zhao Song*, Zhuoyan Xu*, Junze Yin* Conv-Basis: A New Paradigm for Efficient Attention Inference and Gradient Computation in Transformers arXiv 2024. (* denotes alphabetical order).
- Jiajun Song, Zhuoyan Xu, Yiqiao Zhong. Out-of-distribution generalization via composition: a lens through induction heads in Transformers. arXiv 2024.
- Zhuoyan Xu, Khoi Duc Nguyen, Preeti Mukherjee, Somali Chaterji, Yingyu Liang, Yin Li. AdaInf: Adaptive Inference for Resource-Constrained Foundation Models. ICML 2024 Workshop.
- Zhuoyan Xu*, Zhenmei Shi*, Yingyu Liang. Do Large Language Models Have Compositional Ability? An Investigation into Limitations and Scalability. COLM 2024. (* denotes equal contribution).
- Zhenmei Shi, Jenny Wei, Zhuoyan Xu, Yingyu Liang. Why Larger Language Models Do In-context Learning Differently? ICML 2024.
- Zhuoyan Xu, Zhenmei Shi, Jenny Wei, Fangzhou Mu, Yin Li, Yingyu Liang. Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning. ICLR 2024.

Professional Experience

Research Scientist Intern — May 2024 – Aug. 2024
Amazon AWS AI — Bellevue, WA

- Retrieval-augmented generation (RAG) for Multimodal Large Language Model
- Propose **TabRAG**, a novel framework that addresses table understanding challenges by directly utilizing table images in both retrieval and generation step
- Experimental validation is conducted using a newly constructed table image dataset from 14 public table understanding dataset, demonstrating the robustness and efficiency of our proposed framework

Machine Learning Engineer Intern — May 2022 – Aug. 2022
John Deere — Fargo, ND

- Developed a deep learning framework for electrical machine winding temperature prediction
- Architected an end-to-end data pipeline processing over 1M multivariate time series sequences, optimizing data quality and feature extraction
- Implemented ML models for time series prediction, delivering 50% accuracy improvement over baseline statistical approaches

- Created and deployed Cynet, a custom Python package that streamlined laboratory testing procedures

Research Assistant Sep. 2021 – present
UW - Madison Madison, WI

- Towards understanding and better adaptation of foundation models.

Academic Services

Conference Reviewer: NeurIPS 2024, ICML 2024, ICLR 2025, AISTATS 2025, CVPR 2025

Preprints

- Zhuoyan Xu, Kris Sankaran. Spatial Transcriptomics Dimensionality Reduction using Wavelet Bases. F1000 Research.
- Zhuoyan Xu, Jiaxin Hu, Miaoyan Wang. Generalized tensor regression with covariates on multiple modes. arXiv.

Selected Research Experience

Adaptation of Foundation Models with Multitask Learning Feb. 2022 – Sep. 2023

- Implemented multitask fine-tuning strategy to enhance the performance of foundational models on downstream tasks with scarce labeled data.
- Provide a theoretical framework to substantiate the efficacy of the multitask finetuning methodology.
- Proposed a task selection algorithm based on our theoretical conclusion that effectively identifies related finetuning tasks, thereby boosting the model's effectiveness on specific target tasks.
- Applied the multitask finetuning and task selection algorithm across various experiments in vision, NLP, and multimodal models, resulting in a substantial increase in accuracy.

Adaptive Inference of Vision Foundation Models Nov. 2023 – Now

- Investigate the fast adaptation of pre-trained foundation models balancing accuracy and latency.
- Calculated FLOPS during model inference and contribute to GitHub repo flops-counter.pytorch (Star: 2.6k).
- Train a scheduler that actively deactivates certain components of model during inference, reducing deploy time while keeping accuracy.

Exploring In-Context Learning with Large Language Models Jul. 2023 – Now

- Analyzing general in-context learning tasks through the lens of in-context exemplar selection, utilizing a similarity-based approach.
- Conducted comprehensive studies on the impact of model scale on in-context learning, applying it to classification tasks with models ranging from 70 million to 70 billion parameters.
- Examining the compositional capabilities in in-context learning by designing and testing both simple and complex tasks in linguistics to assess LLM performance.
- Exploring parameter effecient finetuning (PEFT) on LLM on arithrmrtic tasks, focusing on assessing performance on out-of-distribution data.

Technical Skills

Languages: Python, R, Julia, Java, C, SQL, HTML
Developer Tools: Git, Linux, AWS, Azure, Docker, Google Cloud Platform, Slurm, LaTeX

Software

- **Xu, Z.**, Sankaran, K., 2022. R package *waveST*: Spatial Transcriptomics Dimensionality Reduction using Wavelet Bases. Published on GitHub. DOI: 10.5281/zenodo.6823315
- **Xu, Z.**, Hu, J. and Wang, M., 2019. R package *Tensorregress*: Generalized tensor regression with covariates on multiple modes. Published on the Comprehensive R Archive Network
- **Xu, Z.**, Solís-Lemus, C. 2022. Julia Package *HighDimMixedModels.jl*: Fitting mixed-effects models with high dimensional fixed effect variable.