Yang Guo

1210 W. Dayton St., Madison, WI 53706-1613
yguo@cs.wisc.edu | (+1) 607-379-7444 | https://pages.cs.wisc.edu/~yguo

EDUCATION

University of Wisconsin-Madison **Madison, WI**
Ph.D. in Computer Science, 3.67 / 4.0
Advisor: Robert Nowak, Yingyu Liang *Sep. 2018 - May. 2025*

University of Wisconsin-Madison **Madison, WI**
M.S. in Computer Science, 3.69 / 4.0
Advisor: Yingyu Liang, Somesh Jha
University of Wisconsin Special CS TA Scholarship *Sep. 2018 - May. 2021*

Cornell University **Ithaca, NY**
B.A. in Statistics and Economics, 3.91 / 4.0
Summa Cum Laude • Exceptional Graduating Senior *Sep. 2014 - May 2018*

TECHNICAL SKILLS

- Python, Java, C, Spark, Hadoop, OCaml, SQL, Latex, R, Julia, Matlab, Stata
- Amazon Web Service (AWS), PyTorch, TensorFlow, Scikit-learn, Pandas, Unix

AREA OF INTEREST

My main research interest is on **Understanding the theoretical and empirical aspects of Model Adaptation** with applications in **Adversarial Robustness** and **Foundational Models**, including LLM alignment, and State-Space Models, etc.

PROFESSIONAL EXPERIENCE

Finetuning Large Language Model with Preference Data for Creative Generation **Madison, WI**
Project Assistant *Jan. 2024 – Now*
Supervised by Robert Nowak and Yingyu Liang

- Developed innovative methods for finetuning large language models (LLMs), utilizing a substantial dataset derived from the New Yorker Cartoon Caption Contest (nextml.github.io/caption-contest-data/)
- Benchmarked variants of Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) using human preference data for creative generation tasks (For example, our proposed group comparison method achieves 77.5% ranking accuracy for ranking captions.)
- Proposed systematic modification of reward training and policy training prompts to improve alignment method for creative generation

Understanding In-context Learning Capacity of MAMBA **Madison, WI**
Research Assistant *Jan. 2024 – Now*
Supervised by Yingyu Liang and Robert Nowak

- Empirically investigated the learning ability of MAMBA(/State Space Model) under non-stationary in-context data
- Derived a theoretical learning framework for MAMBA under various types of distribution shifts

Transformer for Time Series Forecasting for Smart Home Domain **Seattle, WA**
Applied Scientist Intern in Amazon *May. 2023 – Aug. 2023*
Hosted by Kathleen Champion

- Benchmarked existing transformer-based time series forecasting methods on smart home domain
- Designed and implemented global univariate transformer-based time series model with custom featurization for home activity state forecast

Multi-agent System Robustness in Online Auction **Seattle, WA**
Research Scientist Intern in Alibaba *Sept. 2022 – Jan. 2023*
Hosted by Bolin Ding

- Designed and implemented the multi-agent reinforcement learning framework in online auction system and studied the economic phenomenon under non-standard auction environment
- Empirically established the equivalence between the theoretical auction equilibriums and converging equilibrium under reinforcement learning environment

Model Robustness for Smart Home Ambience Intelligence **Seattle, WA**
Applied Scientist Intern in Amazon *May 2022 – Aug. 2022*
Hosted by Michael Dillon

- Categorized prototypical noise patterns in smart home data and proposed a novel temporal corruption threat model
- Developed robust training algorithms to improve model robustness via variants of data augmentation and self-supervision techniques, and consistently achieved improved robust accuracy on corrupted samples

Towards Adversarial Robustness via Transductive Robustness **Madison, WI**
Research Assistant *July 2020 – Dec. 2023*
Supervised by Somesh Jha and Yingyu Liang

- Formalized a novel transductive threat models that could generalize common test-time defenses against adversarial attacks
- Proposed the principles for adaptive attack in the transductive model and developed a strong adaptive attack, *Greedy Model Space Attack (GMSA)*
- Proposed a transductive defense method, *Adversarial Training via Representation Matching (ATRM)*, effective against both transfer attacks and strong adaptive attacks
- Performed extensive experiments analysis and achieved significant robustness improvement over the inductive setting (For example, we improved the robustness accuracy from 91.61% to 94.32% in MNIST dataset, and from 41.06% to 53.53% in CIFAR-10 dataset)

Representation Bayesian Risk Decompositions and Multi-Source Domain Adaptation **Madison, WI**
Research Assistant *Sept. 2019 – July 2020*
Supervised by Somesh Jha and Yingyu Liang

- Provided an exact risk decomposition for both single-source and multi-source domain adaptation with Bayesian optimal classifier
- Proposed experimental verifications for our risk decomposition from the Color-MNIST and MNIST-M dataset and constructed adversarial datasets that could fail methods based on other decompositions

Tensorsketch: Low-Rank Tucker Approximation of a Tensor from Streaming Data **Ithaca, NY**
Research Assistant *Sept. 2017 – Dec. 2018*
Supervised by Madeleine Udell

- Designed the first sketching-based streaming model for Tucker decomposition, an important low-rank tensor approximation algorithm

- Conducted rigorous mathematical analysis on the convergence guarantee of our algorithm, which achieves similar bounds as other state-of-the-art tensor approximation algorithms
- Developed an open-source package in python and performed comprehensive simulation study with synthetic datasets and applied it to real-world weather and combustion datasets
- Codebase: https://github.com/tensorsketch/tensorsketch

Tensor Random Projection **Ithaca, NY**
Research Assistant *Sept. 2017 – Dec. 2018*
Supervised by Madeleine Udell

- Developed the Tensor Random Projection based on Khatri-Rao product, and reduced the storage cost for the random map from O(n) to O(log(n))
- Created the variance reduction method for random projection, and enabled the Tensor Random Projection to achieve same the accuracy as conventional methods with much smaller storage cost

Cornell University **Ithaca, NY**
Statistical Consultant *Feb. 2017 – May 2017*
Supervised by Francoise Vermeylen

- Assisted the client to analyze the effect of biochar and sidedress application on growth of maize with the data from Cornell Musgrave research farm with JMP and R
- Computed the optimal combination for maize production and evaluated the finding with multiple test correction

Publications/Preprints

- Jifan Zhang*, Lalit Jain*, **Yang Guo***, Jiayi Chen, Kuan Los Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy Rogers, Kevin Jamieson, Robert Mankoff, Robert Novak. Humor in AI: Massive Scale Crowd-Sourced Preferences and Benchmarks for Cartoon Captioning. In *Neurips Datasets and Benchmarks (Spotlight), 2024*
- **Yang Guo***, Nils Palumbo*, Xi Wu, Jiefeng Chen, Yingyu Liang, Somesh Jha. Two Heads are Actually Better than One: Towards Better Adversarial Robustness via Transduction and Rejection. In *International Conference on Machine Learning* (*ICML), 2024*
- **Yang Guo***, Nils Palumbo*, Xi Wu, Jiefeng Chen, Yingyu Liang, Somesh Jha. Best of Both Worlds: Towards Adversarial Robustness with Transduction and Rejection. In *Neurips ML Safety Workshop, 2022*
- Jiefeng Chen, Xi Wu, **Yang Guo**, Tianqi Li, Qicheng Lao, Yingyu Liang, Somesh Jha. Towards Evaluating the Robustness of Neural Networks Learned by Transduction. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022
- Yiming Sun, **Yang Guo**, Charlene Luo, Joel Tropp, Madeleine Udell. Low-Rank Tucker Approximation of a Tensor From Streaming Data. In *SIAM Journal on Mathematics of Data Science 2,* no. 4 (2020): 1123-1150
- **Yang Guo***, Yiming Sun*, Joel Tropp, Madeleine Udell. Tensor Random Projection for Low Memory Dimension Reduction. In *Neurips 2018 Workshop on Relational Representation Learning (Spotlight)*, 2018
- Xi Wu, **Yang Guo**, Jiefeng Chen, Yingyu Liang, Somesh Jha, Prasad Chalasani. Representation Bayesian Risk Decompositions and Multi-Source Domain Adaptation. *arXiv preprint: 2004.10390, 2020*

** denotes equal contributions or alphabetical order*