ANDREW BAI

andrewbai@cs.ucla.edu

I am currently a fourth year PhD student at UCLA Computer Science Department advised by Prof. Cho-Jui Hsieh. My research interests involve understanding the memorization and forgetting of machine learning models and mechanisms that control them, with more recent focus on large language models. My most recent research topic of focus is characterizing the instruction-tuning techniques that are best for downstream preference alignment (RLHF) performance in LLMs. I also engage in concurrent collaboration projects spanning multiple topics, including LLM agents, forgetting in instruction-tuning, multi-modal model (MLLM) interpretability, diffusion model data memorization, and prompt optimization.

EDUCATION

University of California, Los Angeles (Los Angeles, USA) Sep 2021 – Exp. Jun 2027
Ph.D. in Computer Science (advised by Prof Cho-Jui Hsieh).

National Taiwan University (Taipei, Taiwan) Sep 2016 – Jan 2021
B.S. in Computer Science and Information Engineering. (GPA: 4.2/4.3)
Minor in Mechanical Engineering.

SELECTED PUBLICATIONS

- A. Bai, C.-K. Yeh, C.-J. Hsieh, and A. Taly. **An Efficient Rehearsal Scheme for Catastrophic Forgetting Mitigation during Multi-stage Fine-tuning.** Under submission review.
- Y. Wang*, A. Bai*, N. Peng, and C.-J. Hsieh. **On the Loss of Context-awareness in General Instruction Fine-tuning.** Under submission review.
- T. Xie*, H. Li*, A. Bai, C.-J. Hsieh. **Data Attribution for Diffusion Models: Timestep-induced Bias in Influence Estimation** In *Transactions on Machine Learning Research (TMLR)*, Jun 2024.
- A. Bai, C.-K. Yeh, P. K. Ravikumar, N. Lin, and C.-J. Hsieh. **Concept Gradient: Concept-based Interpretation Without Linear Assumption.** In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, May 2023.
- A. Bai, H.-T. Lin, C. Raffel, and W. Kan. **On training sample memorization: Lessons from benchmarking generative modeling with a large-scale competition.** In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Aug 2021.

WORK EXPERIENCE

Google Bard (Remote) Jun 2024 – Oct 2024
Student Researcher

- Investigated early-stopping metrics for supervised fine-tuning (SFT) on instruction data to maximize downstream preference alignment (specifically DPO) performance.
- Benchmarked "alignment tax" – the deterioration of instruction following capabilities after preference alignment (RLHF) on LLMs, and identified it as an artifact caused by longer generated responses.

Google Cloud (Remote) Apr 2023 – Aug 2023
Student Researcher

- Designed a computationally-free rehearsal scheme to mitigate catastrophic forgetting for multi-stage training by increasing the likelihood of sampling "useful" samples online. Achieved equivalent performance to baselines with up to 50% less computation.
- Investigated targeted active learning setting on improving specific data slice while maintaining performance on the overall data distribution.

Amazon (Palo Alto, California) Jun 2022 – Sep 2022
Applied Scientist Intern

· Implemented and optimized factorization machine training and inferencing in C++, increasing the training speed by 43x compared to `libffm` (see open-source code PECOS for details).
· Investigated the impact of replacing inner product search with cross-attention methods (e.g. factorization machine) in two-tower deep neural network retrieval models.

RESEARCH EXPERIENCE

Dept. of Computer Science, UCLA (Los Angeles, CA) Sep 2021 – Present
Graduate Student Researcher (advised by Prof. Cho-Jui Hsieh)

· Identified the loss of context-awareness after instruction-tuning LLMs, traced the problem source to the bias in instruction data, and proposed straighforward solutions utilizing attention-steering and conditional supservied fine-tuning (CSFT).
· Designed concept-based interpretability methods for general differentiable models (e.g. neural networks) by propagating gradients through shared input feature representation.

Dept. of Computer Science and Engineering, NTU (Taipei, Taiwan) Jun 2018 – Jan 2021
Research Assistant (advised by Prof. Hsuan-Tien Lin)

· Collaborated with Kaggle (now a subsidiary of Google) on generative modeling metric design and held the first-ever public large-scale generative modeling competition with 900+ participating teams.
· Designed the first algorithm to reduce training sample memorization during Generative Advesarial Networks (GANs) training with rejection sampling.
· Designed the first deep neural network model to predict tropical cyclone rapid intensification using satellite image data and establish strong baseline for our proposed benchmark.

TEACHING AND MENTORSHIP

Undergraduate Research Mentorship Jan 2023 – Present

· Mentored 20+ undergraduate students on 9-month machine learning research projects (voluntarily).
· Volunteered instructing *Algorithms* for high school students from the Los Angeles area (LACC 2024).
· Assisted teaching *Introduction to Algorithms and Complexity* and *Introduction to Programming* for 2 academic quarters each.

GRANTS AND FELLOWSHIP

Kaggle, Alphabet Inc. Jul 2019 – Aug 2019
Generative Adversarial Network Research Grant

· Funding for holding the Kaggle Generative Dog Images competition

Taiwan Ministry of Science and Technology (MOST) Jul 2019 – Feb 2020
MOST Research Grant for University Students

· Funding for tropical cyclone rapid intensification prediction resesarch