Jiayi Yuan

Website: `https://jy-yuan.github.io/` ⋄ Email: jy101@rice.edu

EDUCATION

Rice University Houston, TX
Ph.D. in Computer Science (Advisor: Dr. Xia "Ben" Hu) *Aug. 2022 - Present*

Tsinghua University Beijing, China
B.Eng. in Computer Science *Aug. 2017 - Jul. 2021*

PUBLICATION

Conference Publications

[C1] ***J. Yuan**, *Z. Liu, H. Jin, S. Zhong, Z. Xu, V. Braverman, B. Chen, X. Hu. "KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache", *in the Forty-first International Conference on Machine Learning (ICML), 2024*

[C2] S. Zhong, D. Le, Z. Liu, Z. Jiang, A. Ye, J. Zhang, **J. Yuan**, K. Zhou, Z. Xu, J. Ma, S. Xu, V. Chaudhary, X. Hu. "GNNs Also Deserve Editing, and They Need It More Than Once", *in the Forty-first International Conference on Machine Learning (ICML), 2024*

[C3] ***J. Yuan**, *R. Tang, Y. Li, Z. Liu, R. Chen, X. Hu. "Setting the Trap: Capturing and Defeating Backdoors in Pretrained Language Models through Honeypots", *in the Thirty-seventh Annual Conference on Neural Information Processing Systems (NeurIPS), 2023*

[C4] **J. Yuan**, R. Tang, X. Jiang, X. Hu. "Large language models for healthcare data augmentation: An example on patient-trial matching", *Best Student Paper, in AMIA Annual Symposium Proceedings (AMIA), 2023*

[C5] C. Chang, **J. Yuan**, S. Ding, Q. Tan, K. Zhang, X. Jiang, X. Hu, N. Zou. "Towards Fair Patient-Trial Matching via Patient-Criterion Level Fairness Constraint", *in AMIA Annual Symposium Proceedings (AMIA), 2023*

[C6] Q. Feng, **J. Yuan**, F.B. Emdad, K. Hanna, X. Hu, Z. He. "Can Attention Be Used to Explain EHR-Based Mortality Prediction Tasks: A Case Study on Hemorrhagic Stroke", *in the 14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB), 2023*

[C7] Z. Yu, Y. Fu, **J. Yuan**, H. You, Y. Lin. "NetBooster: Empowering Tiny Deep Learning By Standing on the Shoulders of Deep Giants", *in Proceedings of the 60th ACM/IEEE Design Automation Conference (DAC), 2023*

[C8] Y. Fu, Y. Yuan, S. Wu, **J. Yuan**, Y. Lin. "Robust Tickets Can Transfer Better: Drawing More Transferable Subnetworks in Transfer Learning", *in Proceedings of the 60th ACM/IEEE Design Automation Conference (DAC), 2023*

[C9] *Y. Fu, *Z. Ye, **J. Yuan**, S. Zhang, S. Li, H. You, Y. Lin. "Gen-NeRF: Efficient and Generalizable Neural Radiance Fields via Algorithm-Hardware Co-Design", *in the 50th IEEE/ACM International Symposium on Computer Architecture (ISCA), 2023*

[C10] ***J. Yuan**, *C. Li, *W. Chen, Y. Lin, A. Sabharwal. "ERSAM: Neural Architecture Search for Energy-Efficient and Real-Time Social Ambiance Measurement", *in 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023*

[C11] Y. Fu, H. Yang, **J. Yuan**, M. Li, C. Wan, R. Krishnamoorthi, V. Chandra, Y. Lin. "DepthShrinker: A New Compression Paradigm Towards Boosting Real-Hardware Efficiency of Compact Neural Networks", *in the Thirty-ninth International Conference on Machine Learning (ICML), 2022*

[C12] *H. You, *Y. Zhao, *Z. Yu, *C. Wan, Y. Fu, **J. Yuan**, S. Wu, S. Zhang, Y. Zhang, C. Li, V. Boominathan, A. Veeraraghavan, Z. Li, Y. Lin. "EyeCoD: Eye Tracking System Acceleration via FlatCam-Based Algorithm and Accelerator Co-Design", *IEEE Micro Top Pick, in the 49th IEEE/ACM International Symposium on Computer Architecture (ISCA), 2022*

Preprints

[P1] ***J. Yuan**, *H. Liu, *S. Zhong, Y. Chuang, S. Li, G. Wang, D. Le, H. Jin, V. Chaudhary, Z. Xu, Z. Liu, X. Hu. "KV Cache Compression, But What Must We Give in Return? A Comprehensive Benchmark of Long Context Capable Approaches"

[P2] *Y. Chuang, *S. Li, ***J. Yuan**, *G. Wang, *K. Lai, L. Yu, S. Ding, C. Chang, Q. Tan, D. Zha, X. Hu. "Understanding Different Design Choices in Training Large Time Series Models"

[P3] G. Wang, Y. Chuang, R. Tang, S. Zhong, J. Yuan, H. Jin, Z. Liu, V. Chaudhary, S. Xu, J. Caverlee, X. Hu. "Secured Weight Release for Large Language Models via Taylor Expansion"

[P4] H. Liu, Z. Liu, R. Tang, **J. Yuan**, S. Zhong, Y. Chuang, L. Li, R. Chen, X. Hu. "LoRA-as-an-Attack! Piercing LLM Safety Under The Share-and-Play Scenario"

[P5] C. Chang, S. Ding, **J. Yuan**, K. Zhang, X. Jiang, X. Hu, N. Zou. "Fair Patient-Trial Matching for Underrepresented Groups"

EXPERIENCE

Rice University Houston, TX
Graduate Research Assistant *Sep. 2022 - Present*

- Working on **Large Language Models (LLMs)**: efficient and trustworthy finetuning and inference. [C1] [P4]
- Designed a defender algorithm against natural language backdoor attacks. [C3]
- Worked on several projects regarding health informatics. [C4] [C5] [C6] [P5]

Amazon.com, Inc. Seattle, WA
Applied Scientist Intern *May 2024 - Aug. 2024*

- Worked on LLM agents for Amazon stores.
- Built a RAG-based modalized pipeline.

Rice University Houston, TX
Research Assistant *Aug. 2021 - Aug. 2022*

- Proposed re-parameterization-based efficient training and inference algorithms. [C7] [C11]
- Proposed a NAS pipeline for real-time social ambiance measurement. [C10]
- Worked on several projects regarding machine learning algorithms and systems co-design. [C9] [C12]
- Took part in efficient computer vision challenges: LPCVC-UAV, DAC-SDC.

HONORS AND AWARDS

SDM'24 Doctoral Forum Travel Award, ***by SIAM*** *Mar. 2024*
NeurIPS 2023 Scholar Award, ***by NeurIPS*** *Nov. 2023*
AMIA Best Student Paper, ***by AMIA*** *Nov. 2023*
AMIA 2023 KDDM Student Innovation Award, ***by AMIA*** *Oct. 2023*
IEEE Micro Top Picks, ***by ACM*** *Jul. 2023*