Bellevue, WA
xiaoxiawu10@gmail.com
xwushirley.github.io
xiaoxiawu1
Google Scholar

Xiaoxia (Shirley) Wu

Professional Summary

Dr. Wu is a highly accomplished senior researcher and engineer specializing in the optimization of large language models (LLMs), with expertise in quantization (FP8/mxFP4/nvFP4/INT4, ZeroQuant), training & inference efficiency and cross-stack optimization.

Experience

Jul. 2024 – Present **Senior Staff Scientist**, *Together AI*, Remote
- Lead the development of core quantization methodologies (FP8/FP4/INT4) for LLM servers including vLLM, SGLang, and TRT-LLM.
- Build and direct a specialized team, delivering robust evaluation frameworks .
- Drive innovation in bridging advanced research with scalable deployment, focusing on quantization, inference optimization, and full-stack performance.

Nov. 2021 – Jun. 2024 **Senior Researcher**, *DeepSpeed Team, Microsoft*, Redmond, WA
- Led quantization projects: *DeepSpeed-FP6*, *ZeroQuant-FP*, *ZeroQuant-HERO*, and *ZeroQuant*.
- Core contributor to *DeepSpeed-Chat* and *DeepSpeed-VisualChat*.
- Published papers including *Extreme Compression Made Simple & Efficient* in NeurIPS 2022 oral.

Sep. 2020 – Dec. 2020 **Research Intern**, *Birch.AI (Mentor: Yinhan Liu)*, Seattle, WA
Developed speech-to-text generation and summarization using encoder-decoder Transformers.

May – Aug. 2020 **Research Intern**, *Google (Mentor: Behnam Neyshabur)*, Mountain View, CA
Researched curriculum learning for CIFAR10 and ImageNet; results published as ICLR 2021 oral.

Aug. 2017 – Oct. 2018 **Research Intern**, *Meta AI Research (Mentor: Léon Bottou)*, New York, NY
- Implemented and benchmarked state-of-the-art optimization algorithms in PyTorch.
- Analyzed optimization of layer and weight normalization; work presented as ICML 2019 oral.

Education

Aug. 2014 – Dec. 2020 **Ph.D. in Machine Learning**, *The University of Texas at Austin*
- Advisors: Rachel Ward (supervisor) & Léon Bottou (co–supervisor)
- Thesis (Frank Gerth III Dissertation Award):
  Gradient-based Optimization and implicit regularisation over non-convex landscapes

Publications and Preprints (selected)

Model Compression

[1] H. Xia, Z. Zheng, **X. Wu**, *et al. FP6-LLM: Efficiently Serving Large Language Models through FP6-Centric Algorithm–System Co-Design. arXiv:2401.14112* (2024).

[2] **X. Wu**, H. Xia, S. Youn, Z. Zheng, *et al. ZeroQuant(4+2): Redefining LLM quantization with an FP6 Strategy. arXiv:2312.08583* (2023).

[3] **X. Wu***, Z. Yao* *ZeroQuant-FP: W4A8 Post-Training Quantization Using Floating-Point Formats*. NeurIPS ENLSP (2023).

[4] G. Wang, H. Qin, S.A. Jacobs, **X. Wu**, *et al. ZeRO++: Extremely Efficient Collective Communication for Large Model Training*. ICLR (2024).

[5] Z. Yao, **X. Wu**, C. Li, S. Youn, Y. He. *ZeroQuant-V2: Exploring Post-training Quantization in LLMs*. AAAI (2024).

[6] **X. Wu**, C. Li, R.Y. Aminabadi, Z. Yao, Y. He. *Understanding INT4 Quantization for Transformer Models*. ICML (2023).

[7] **X. Wu**, Z. Yao, *et al. XTC: Extreme Compression for Pre-trained Transformers Made Simple and Efficient*. NeurIPS (2022), oral.

[8] Z. Yao, R.Y. Aminabadi, M. Zhang, **X. Wu**, *et al. ZeroQuant: Efficient Post-Training Quantization for Transformers*. NeurIPS (2022).

Optimization and Theory

[10] **X. Wu**, E. Dyer, B. Neyshabur. *When Do Curricula Work?* ICLR (2021), oral.

[11] **X. Wu**, Y. Xie, *et al. Adaloss: A Computationally-Efficient Adaptive Gradient Method*. AAAI (2022).

[13] **X. Wu***, E. Dobriban*, T. Ren*, *et al. Implicit Regularisation and Convergence for Weight Normalisation*. NeurIPS(2020).

[14] R. Ward*, **X. Wu***, L. Bottou. *Adagrad Stepsizes: Sharp Convergence over Nonconvex Landscapes*. ICML (2019), oral.

NLP and Multi-modal

[19] Z. Yao, **X. Wu**, C. Li, M. Zhang, H. Qin, *et al. DeepSpeed-VisualChat: Multi-Round Multi-Image Interleave Chat*. *arXiv:2309.14327* (2023).

[20] Z. Yao, R.Y. Aminabadi, O. Ruwase, S. Rajbhandari, **X. Wu**, *et al. DeepSpeed-Chat: Affordable RLHF Training of ChatGPT-like Models*. *arXiv:2308.01320* (2023).

[21] C. Li, Z. Yao, **X. Wu**, M. Zhang, Y. He. *DeepSpeed Data Efficiency: Improving Deep Learning Model Quality*. AAAI (2024).

Teaching Experience

- Probability I (Spring 2019)
- Scientific Computation in Numerical Analysis (Spring 2018)
- Calculus: Sequences & Series; Multivariate (Spring 2016, Fall 2016)
- Differential & Integral Calculus (Fall 2014, Spring 2015, Fall 2015)

Honours and Awards

Professional Development Award, UTAustin (2018,2019)

ICML & NeurIPS Travel Awards (2019)

Graduate School Fellowship, UTAustin (2018)

Frank Gerth III Dissertation Award (2020)

Scotland Saltire Scholarship, University of Edinburgh (2012)

Professional Service

- Journal Reviewer: *Journal of Machine Learning Research*
- Conference Reviewer: AISTATS 2020, MSML 2020, NeurIPS 2020, WiML 2019
- Mentor: Directed Reading Program, UTAustin
- Volunteer Teacher: Sanger Center, UTAustin