An Xu

Websites	[Homepage][Linkedin][Google Scholar]	**Email**	annn.xu@gmail.com
		Mobile Phone	+1 (412) 330 0758

Experiences

Jun 2023 - now **Research Scientist** at **ByteDance/TikTok**, *Bellevue, WA*
Machine Learning System Team:

- Sparse MoE LLM inference latency: reduce the number of activated experts (30%), expert pruning, speculative decoding.
- Mentor research interns.

code LLM Team:

- Lead tensor-parallel LoRA fine-tuning, reducing #GPUs (75%) and wall-clock time (50%).
- Instruction fine-tuning for code completion (+10% on HumanEval benchmark) and FIM (fill-in-the-middle).

Jun 2022 - Aug 2022 **Applied Scientist Intern** at **AWS AI**, *Santa Clara, CA*
Large-batch optimization for LLM pretraining.

Jun 2021 - Aug 2021 **Applied Research Intern** at **NVIDIA Research**, *Remote*
Propose new cross-silo federated learning methods (CVPR 2022, ECCV 2022), which are publicly available in NVIDIA's NVFLARE link1 and link2 respectively.

Jun 2019 - Nov 2019 **Research Intern** at **JD Digits AI Lab**, *Mountain View, CA*
Accelerate deep learning model parallelism by incorporating staleness (pipeline parallelism) (CVPR 2020).

Education

Aug 2018 - Aug 2023 **University of Pittsburgh, *Pittsburgh, PA***

Ph.D in Electrical and Computer Engineering.

Sep 2013 - Jul 2017 **Tsinghua University, *Beijing***
B.Eng in Electrical Engineering. Second degree: B.Eng in Business Administration.

Publications

[Homepage/Publications][Google Scholar]
16 publications with 500+ citations in top machine learning conferences covering Sparse MoE LLM pruning, Large-scale Distributed Training (Data and Model Parallelism) and Federated Learning, etc.

Served as program committee / reviewer for:

- ICLR'22-25; ICML'22-24; NeurIPS'22-24. ICML'22 session chair and outstanding reviewer.
- CVPR'21-23; ICCV'21, 23; ECCV'20, 22, 24
- AAAI'22-25; IJCAI'23
- KDD'20, 23; CIKM'22
- Journals: TNNLS, TMI

Last updated: October 23, 2024