An Xu

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

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

- SparseMoE 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).

Internships

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

Large-scale Distributed Training:

- **A. Xu**, Y. Bai. *Distributed Adaptive Learning with Divisible Communication.* ECML PKDD 2023 research track. [paper].
- **A. Xu**, H. Huang. *Detached Error Feedback for Distributed SGD with Random Sparsification.* ICML 2022 (**spotlight**). [paper] [arxiv] [codes]
- **A. Xu**, Z. Huo, H. Huang. *Step-Ahead Error Feedback for Distributed Training with Compressed Gradient.* AAAI 2021. [paper] [arxiv] [codes]
- Y. Huang, X. Yan, G. Jiang, T. Jin, J. Cheng, **A. Xu**, Z. Liu, S. Tu. *Tangram: bridging immutable and mutable abstractions for distributed data analytics.* USENIX ATC 2019. [paper]

Model Parallelism:

- **A. Xu**, Y. Bai. *Cross Model Parallelism for Faster Bidirectional Training of Large Convolutional Neural Networks.* ECML PKDD 2023 research track. [paper].

- **A. Xu**, Z. Huo, H. Huang. *On the Acceleration of Deep Learning Model Parallelism with Staleness.* CVPR 2020. [paper] [arxiv]

Federated Learning:

- **A. Xu**, W. Li, P. Guo, D. Yang, H. Roth, A. Hatamizadeh, C. Zhao, D. Xu, H. Huang, Z. Xu. *Closing the Generalization Gap of Cross-silo Federated Medical Image Segmentation.* CVPR 2022. [paper] [arxiv] [codes]
- **A. Xu**, H. Huang. *Coordinating Momenta for Cross-silo Federated Learning.* AAAI 2022 (**oral**). [paper] [arxiv]
- P. Guo, D. Yang, A. Hatamizadeh, **A. Xu**, Z. Xu, W. Li, C. Zhao, D. Xu, S. Harmon, E. Turkbey, B. Turkbey, B. Wood, F. Patella, E. Stellato, G. Carrafiello, V. M Patel, H. Roth. *Auto-FedRL: Federated Hyperparameter Optimization for Multi-institutional Medical Image Segmentation.* ECCV 2022. [paper] [arxiv] [codes]
- H. Gao, **A. Xu**, H. Huang. *On the Convergence of Communication-Efficient Local SGD for Federated Learning.* AAAI 2021. [paper]
- B. Gu, **A. Xu**, Z. Huo, C. Deng, H. Huang. *Privacy-Preserving Asynchronous Vertical Federated Learning Algorithms for Multi-Party Collaborative Learning.* TNNLS 2021. [paper] [arxiv] [codes]

Misc:

- Y. Liu, **A. Xu**, and Z. Chen. *Map-based Deep Imitation Learning for Obstacle Avoidance.* IROS 2018. [paper]
- J. Li, X. Yan, J. Zhang, **A. Xu**, J. Cheng, J. Liu, K. Ng, and T. Cheng. *A General and Efficient Querying Method for Learning to Hash.* SIGMOD 2018. [paper]
- Y. Li, Z. Jiang, **A. Xu**, S. Zhou, and Z. Niu. *Elastic Local Breakout Strategy and Implementation for Delay-Sensitive Packets with Local Significance.* In Proceedings of the 9th International Conference on Wireless Communications and Signal Processing, 2017. [paper] [arxiv]

Academic Services

Conference program committee / reviewer:

- ICLR'22-25; ICML'22-24; NeurIPS'22-24
- CVPR'21-23; ICCV'21, 23; ECCV'20, 22, 24
- AAAI'22-25; IJCAI'23
- KDD'20, 23; CIKM'22

ICML'22 session chair (session 7, track 9, MISC/Deep Learning)
ICML'22 outstanding reviewer (top 10%).
Journal reviewer: TNNLS, TMI

Last updated: October 18, 2024