DIXI YAO

(+86)1772-147-1282 | dixi.yao@mail.utoronto.ca | github.com/dixiyao | https://dixiyao.github.io

EDUCATION

University of Toronto — Toronto, Canada

M.A.Sc, Department of Electrical and Computer Emgineering — *Sept. 2022 –*

- Advisor: Prof. Baochun Li.

Shanghai Jiao Tong University — Shanghai, China

BEng, Department of Computer Science — *Aug. 2018 – Jun. 2022*

- AI Honors Program. Bachelor Thesis: **Research on privacy preserving methods via Transnformer**
- Advisor: Prof. Liyao Xiang, Prof. Xinbing Wang. Theme: Intelligent Edge
- Cumulative GPA: 3.87/4.3, Major GPA: 3.92/4.3 (ranking top 8%)
- Standard Testing: TOEFL: 105 (R28, L28, S23, W26), GRE: V157, Q170, W3.5
- Core Courses Performance:
  - Programming: Thinking and Approaching Programming [C++] (A+) / Data Structure [C++] (A+) / Problem Solving and Practice [C++] (A+)
  - Mathematics: Probability and Statistics (A+) / Discrete Mathematics (A+) /Algorithm and Complexity (A+) / Mathematical Foundations (A+)
  - Artificial Intelligence: Digital Graphics Processing (A+) / Machine Learning (A) / Data Mining Techniques (A) /Science and Technology Innovation (Bioinformatics) (A+)
  - Computer Networks: Computer Networks (A) /Mobile Internet (A+)

Max Planck Institute for Informatics — Saarbrücken,Germany

Research Intern Fellowship — *Jul. 2021 – Dec. 2021*

- Advisor: Prof. Yiting Xia. Theme: Distributed Deep Learning Benchmark

PUBLICATIONS

Privacy-Preserving Split Learning via Patch Shuffling over Transformers

Proc. IEEE International Conference on Data Mining (ICDM), Orlando, USA — *Nov 28 - Dec 1, 2022*

- **Dixi Yao**, Liyao Xiang, Hengyuan Xu, Hangyu Ye, Yingqi Chen

Context-Aware Compilation of DNN Training Pipelines across Edge and Cloud

The Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT) — *Vol. 5, No. 4, 2021*

- **Dixi Yao**, Liyao Xiang, Zifan Wang, Jiayu Xu, Chao Li, Xinbing Wang

Federated Model Search via Reinforcement Learning

Proc. IEEE International Conference on Distributed Computing Systems (ICDCS), USA — *July 7-10, 2021*

- **Dixi Yao***, Lingdong Wang*, Jiayu Xu, Liyao Xiang, Shuo Shao, Yingqi Chen, Yanjun Tong

Context-Aware Deep Model Compression for Edge Cloud Computing

Proc. IEEE International Conference on Distributed Computing Systems (ICDCS), Singapore — *July 8 - 10, 2020*

- Lingdong Wang, Liyao Xiang, Jiayu Xu, Jiaju Chen, Xing Zhao, **Dixi Yao**, Xinbing Wang, Baochun Li

TEACHING EXPERIENCES

- Teaching Assistant Experience: **MA500 Application of fuzzy Math** (Mar. 2020 – Jun. 2020).
- Course Lab Design: Edge-Cloud Computing for SJTU **EE447 Mobile Internet**.
- Two open lectures for Data Structure (over **100** students participated) and multiple tutor experiences.
- Programming Language: Python (torch, TF), C++.
- Software and Hardware skills: AWS, Verilog, EMU8086, VMware/Vbox, Arduino, Raspberry Pi, Nvidia Jetson TX2.
- Video recording and publishment for CS214 Algorithm and Complexity

RESEARCH EXPERIENCES

Distributed Deep Learning Benchmarking Research Assistant

Advisor: Prof. Yiting Xia, Max Planck Institute for Informatics *Jul. 2021 – Dec. 2021*

- Explored several design drawbacks of current distributed deep learning frameworks under certain GPU cluster configurations.
- Benchmarked the latent causes of training latency fluctuation and training failure when training CNN and transformers with Pytorch native Allreduce DDP under several different GPU and network configurations.
- Investigated how Ray and Hoplite framework's scheduler react to previous latency fluctuation and failures and tested Ray's performance for synchronized distributed deep learning.

Context-Aware Compilation of DNN Training Pipelines Research Assistant

Advisor: Prof. Liyao Xiang, SJTU *Oct. 2020 – Jun. 2021*

- Sped up distributed training across edge and cloud when mobile users need personalized AI applications with low latency;algorithm maintains users' privacy.
- Developed a grouped parallelization algorithm to break the backpropagation lock, thus allowing for different layers of a neural network which can be trained at the same time. Proposed error feedback compression to compensate for accuracy loss brought about by lossy compression methods.
- Established a context-aware training system which can adapt to various network and device conditions. The system can reduce latency by up to 90% with less than 1.5% accuracy loss, less than 10% extra memory cost and less than 10% extra energy consumption.
- First author of publication which will be featured in Issue 4 of IMWUT Vol. 5. paper

Federated Neural Architecture Search Research Assistant

Advisor: Prof. Liyao Xiang, SJTU *Apr. 2020 – Sept. 2020*

- Designed a novel neural architecture search algorithm based on reinforcement learning, searching for a better model architecture fit for non-i.i.d. data distribution over each participant in federated learning where previous models would have demonstrated sub-optimal performance.
- Solved the delay compensation problem when multiple workers are uploading gradients while some of them are out of date, through updating model weights and architecture hyper parameter asynchronously, efficiency was greatly enhanced.
- Implemented our work in real distributed computation environment, outperforming most federated neural architecture searches in accuracy (up to 5% higher) and efficiency (up to 95% less parameters).
- First author, published in ICDCS 21.paper

Pruning Neural Architecture Search Research Assistant

Advisor: Prof. Liyao Xiang, SJTU *Jan. 2020 – Apr. 2020*

- Discovered abundant model search space in neural architecture search works.
- Extended the lottery-ticket theory by pruning abundant model weights to pruning architecture parameters of a supernet adaptively, to reduce the search space and speed up NAS.
- Derived bi-level annealing by dynamically justifying the cutting rate of supernet architecture parameters, in order to avoid unbalanced pruning on different nerual networks' operations.

Context-Aware Deep Model Compression Research Assistant

Advisor: Prof. Liyao Xiang, SJTU *Jun. 2019 – Jan. 2020*

- Addressed the issue when an edge equipment cannot run the DNN with its own computing power and ensured the protection of user privacy.
- Proposed a method based on reinforcement learning to design an engine which is capable of making a decision of where to partite and how to compress a model on the fly depending on real-time network condition.
- Implemented the algorithm and reduced 30% to 50% inference latency while keeping the accuracy loss at about 1%.
- Published in ICDCS20. paper

Review Experience: ICCC22

More research and course work projects can refer to personal website.