YIFAN (EVELYN) GONG

(she / her / hers) ⋄ 140 The Fenway, Boston, MA 02115

(857)891-0509 gong.yifa@northeastern.edu Personal Website Yifan Gong

EDUCATION

Northeastern University Boston, MA

Ph.D. Candidate in Computer Engineering, advised by Prof. Yanzhi Wang Sep 2019 – current

- With a focus on **energy-efficient deep learning and artificial intelligence systems, accelerations of deep neural networks including large-scale models for AIGC, trustworthy machine learning**

University of Toronto Toronto, ON, Canada

Master of Applied Science (Thesis-based with Fellowship) Sep 2017 – Sep 2019

- With a focus on deep reinforcement learning and its applications

Xidian University Xi'an, Shaanxi, China

Bachelor of Engineering (**Valedictorian**, **with highest honor**), GPA: 3.83/4.0 (rank $\mathbf{1}^{st}$) Sep 2013 – Jun 2017

Education Experimental Class **(Undergraduate honor program)**

PUBLICATIONS

Submitted, † means equal contribution.

[I5] **Yifan Gong**, Zheng Zhan, Yushu Wu et al, "An Automatic Framework for Adaptive Deep Neural Network Deployment with DVFS on the Edge", *under review* for TCAD.

[I4] **Yifan Gong**, Yushu Wu, Pu Zhao, et al, "Lotus: learning-based online thermal and latency variation management for two-stage detectors on edge devices", *under review* for DAC 2024.

[I3] **Yifan Gong**†, Chao Wu†, Liangkai Liu†, et al, "AyE-Edge: Automated Deployment Space Search Empowering Accuracy yet Efficient Real-Time Object Detection on the Edge", *under review* for DAC 2024.

[I2] **Yifan Gong**, Zheng Zhan, Qing Jin, et al, "E^{2}GAN: Efficient Training of Efficient GANs for Image-to-Image Translation", *under review* for CVPR 2024.

[I1] Yuguang Yao†, Jiancheng Liu†, **Yifan Gong**†, et al, "Can Adversarial Examples Be Parsed to Reveal Victim Model Information?", *under review* for CVPR 2024.

Conference Proceedings, † means equal contribution.

[C13] **Yifan Gong**†, Yushu Wu†, Zheng Zhan et al, "MOC: Multi-Objective Mobile CPU-GPU Co-optimization for Power-efficient DNN Inference", in **ICCAD** 2023. (**Acceptance rate: 22.9%**)

[C12] **Yifan Gong**, Pu Zhao, Zheng Zhan, et al, "Condense: A Framework for Device and Frequency Adaptive Neural Network Models on the Edge", in **DAC** 2023. (**Acceptance rate: 23%**)

[C11] **Yifan Gong**, Zheng Zhan, Pu Zhao, et al, "All-in-One: A Highly Representative DNN Pruning Framework for Edge Devices with Dynamic Power Management", in **ICCAD** 2022. (**Acceptance rate: 22.5%**)

[C10] **Yifan Gong**, Yuguang Yao, Yize Li, Yimeng Zhang, Xiaoming Liu, Xue Lin, Sijia Liu, "Reverse Engineering of Imperceptible Adversarial Image Perturbations", in **ICLR** 2022. (**Acceptance rate: 32.2%**)

[C9] **Yifan Gong**†, Yushu Wu†, Pu Zhao, et al, "Compiler-Aware Neural Architecture Search for On-Mobile Real-time Super-Resolution", in **ECCV** 2022. (**Acceptance rate: 28%**)

[C8] **Yifan Gong**†, Zheng Zhan†, Pu Zhao, et al, "Achieving on-Mobile Real-Time Super-Resolution with Neural Architecture and Pruning Search", in **ICCV** 2021. (**Acceptance rate: 25.9%**)

[C7] **Yifan Gong**, Zheng Zhan, Zhengang Li, et al, "A Privacy-Preserving-Oriented DNN Pruning and Mobile Acceleration Framework", in **GLSVLSI** (invited) 2020.

[C6] **Yifan Gong**, Baochun Li, Ben Liang, Zheng Zhan, "Chic: Experience-driven Scheduling in Machine Learning Clusters", in **IWQoS** 2019.

[C5] Zifeng Wang, Zheng Zhan, **Yifan Gong**, et al, "DualHSIC: HSIC-Bottleneck and Alignment for Continual Learning", in **ICML** 2023.

[C4] Sizhe Chen, Geng Yuan, Xinwen Cheng, **Yifan Gong**, et al, "Self-Ensemble Protection: Training Checkpoints Are Good Data Protectors" in **ICLR** 2023.

[C3] Zifeng Wang, Zheng Zhan, **Yifan Gong**, et al, "Sparcl: Sparse continual learning on the edge"in **NeurIPS** 2022.

[C2] Geng Yuan, Xiaolong Ma, Wei Niu, Zhengang Li, Zhenglun Kong, Ning Liu, **Yifan Gong**, et al, "Mest: Accurate and fast memory-economic sparse training framework on the edge"in **NeurIPS** 2021.

[C1] Peiyan Dong, Siyue Wang, Wei Niu, Chengming Zhang, Sheng Lin, Zhengang Li, **Yifan Gong**, et al, "RTMobile: Beyond Real-Time Mobile Acceleration of RNNs for Speech Recognition", in **DAC** 2020.

Journal Papers

[J2] **Yifan Gong**, Geng Yuan, et al, "Automatic Mapping of the Best-Suited DNN Pruning Schemes for Real-Time Mobile Acceleration", ACM Transactions on Design Automation of Electronic Systems (**TODAES**), 2021.

[J1] Tong Jian, **Yifan Gong**, et al, "Radio Frequency Fingerprinting on the Edge", IEEE Transactions on Mobile Computing, 2021.

RESEARCH EXPERIENCE

Snap Inc. Santa Monica, CA
Ph.D. Research Intern @ Creative Vision Group May 2023 – Aug 2023

- *Project: Model Generation with Knowledge from Diffusion Models*
  Content: Worked on efficient distillation of GANs from diffusion models (**E^{2}GAN**).
  - Proposed a novel knowledge transfer framework to train efficient GANs with knowledge from diffusion models
  - Built model weight generation pipeline

IBM Research Cambridge, MA
Ph.D. Research Intern @ MIT-IBM Watson AI Lab May 2021 – Aug 2021

- *Project: Improving Vision Transformers by Attention Graph*
  Content: Worked on improving the performance of vision transformers by incorporating the interpretability of an image with structural information.

Northeastern University Boston, MA
Research Assistant advised by Prof. Yanzhi Wang @ College of Engineering Sep 2019 – present

- *Project: Effective Compression-DVFS Co-design* Feb 2022 – present
  Content: Worked on reducing runtime variation of DNNs on edge devices under dynamic power management with DVFS (**DAC-23**, **ICCAD-22**)
  - Developed a framework to get multiple subnets in one DNN to reduce latency variation for different hardware frequency levels with DVFS (**ICCAD-22**)
  - Proposed a two-level algorithm for obtaining subnets with arbitrary ratios in a single model with theoretical proof for a more automatic framework that works for arbitrary devices (**DAC-23**)
- *Project: Intelligent Diagnosis for Machine and Human-Centric Adversaries* Jan 2021 – Mar 2023
  Content: Explored a new adversarial learning paradigm-Reverse Engineering of Deceptions (**ICLR-22**).
  - Formulated the Reverse Engineering of Deceptions (RED) problem to estimate adversarial perturbations and provided the feasibility of inferring the adversary intention
  - Identified a series of RED principles and built a comprehensive evaluation pipeline
  - **Recognized and valued by the community, we had the privilege of hosting the CVPR'23 tutorial on Reverse Engineering of Deceptions (RED) based on my two works on RED against machine-centric attacks**

- *Project: Compression-Compilation Co-design (CoCoPIE)* Feb 2020 – present
  Content: Optimizing AI models for the implementation on edge devices (**ICCV-21**, **ECCV-22**).
  - Worked on achieving Real-Time Super-Resolution on Mobile platform, we are **the first** to achieve real-time SR inference for implementing 720p resolution with competitive image quality on mobile platforms

University of Toronto Toronto, ON, Canada
Research Assistant advised by Prof. Baochun Li @ Department of ECE Sep 2017 – Sep 2019

- *Project: Scheduling Machine Learning Jobs with Reinforcement Learning*
  Content: Proposed a scheduler to find the scheduling decision for distributed machine learning workloads to minimize the average completion time based on reinforcement learning (**IWQoS-19**).
  - Modeled the scheduling problem for reinforcement learning agent and simulated the results to compare with SOTA methods

TEACHING EXPERIENCES

Teaching Assistant of Advances in Deep Learning @ Northeastern University

- Taught a 2-hour course about how to use SOTA deep learning frameworks such as Pytorch and Tensorflow
- Prepared course materials and final project topics
- Held office hours to address the questions of students in the class
- Assigned and graded homework, quiz, and final projects

Teaching Assistant of Operating System and Computer Fundamentals @ University of Toronto

- Gave lab demonstrations to students
- Marked and graded tests and exams

INVITED TALKS

Tutorials

[T4] "Reverse Engineering of Deceptions: Foundations and Applications", @ **CVPR'23**.

Invited Seminars

[T3] "Automatic Mapping of the Best-Suited DNN Pruning Schemes for Real-Time Mobile Acceleration", in ROAD4NN @ **DAC'21**.

[T2] "A Privacy-Preserving-Oriented DNN Pruning and Mobile Acceleration Framework", @ **GLSVLSI'20**.

[T1] "Towards Best Possible Deep Learning Acceleration on the Edge - A Compression-Compilation Co-Design Framework", in MGHPCC @ **SC'20**.

REVIEW SERVICES AND SKILLS

Review services: ICLR'23, NeurIPS'23, ICCV'23, CVPR'23, ISCAS'23, AICAS'23, AdvML'22, TCAD'22

Research interests: Hardware and Software Co-Design for Artificial Intelligence Acceleration, Energy-Efficient Deep Learning and Artificial Intelligence Systems, Accelerations of Emerging Large-Scale Models for AIGC such as Diffusion Models and Large Language Model, Model Compression, Efficient and Robust Deep Learning

SELECTED SCHOLARSHIP, HONORS AND AWARDS

ICCAD Travel Award	09/2023
College of Engineering **Outstanding TA Awards** of Northeastern University	04/2023
College of Engineering **Dean's Fellowship** of Northeastern University	2019-2020
ECE Student Fellowship of University of Toronto	2017-2019
Valedictorian of Xidian University	06/2017
Excellent Graduate of Xidian University (**10 of 5180**)	06/2017
National Scholarship (1%)	10/2015, 10/2016
Role Model Outstanding Student	11/2014, 11/2015
Provincial 1st Prize in CUMCM	11/2015