

ZHENG ZHAN

140 The Fenway, Boston, MA 02115

☎ (857)891-6751 ✉ zhan.zhe@northeastern.edu in Zheng Zhan

EDUCATION

Northeastern University Ph.d. Candidate in Computer Engineering, GPA: 4.0/4.0 • With a focus on Deep Learning and its applications.	Boston, MA Sep 2019 – May 2023 (expected)
Syracuse University Master of Science in Computer Engineering, GPA: 3.833/4.0	Syracuse, NY Sep 2017 – May 2019
Xidian University Bachelor of Engineering in Electronic Science and Technology Excellent Class (Undergraduate Honor Program)	Xi'an, Shaanxi, China Sep 2013 – Jun 2017

EXPERIENCE

Samsung Research America <i>Ph.D. Research Intern</i>	Mountain View, CA May 2022 – Aug 2022
Lawrence Livermore National Laboratory <i>Ph.D. Research Intern @ DSSI program</i> • <i>Project: Multi-Prize Lottery Tickets of Vision Transformer</i> Content: Worked on finding multi-prize lottery tickets for vision transformer to achieve a high performance and robustness using sparse binary network. <ul style="list-style-type: none">– Developed a framework based on the Binarize-PRune Optimizer to find the multi-prize lottery tickets of vision transformer– Conducted the experiments on efficient Binary Neural Networks (BNN) training algorithm	Livermore, CA May 2021 – Aug 2021 May 2021 – Aug 2021
Northeastern University <i>Research Assistant advised by Prof. Yanzhi Wang @ College of Engineering</i> • <i>Project: Compression-Compilation Co-design (CoCoPIE)</i> Content: CoCoPIE, a startup developing a platform that optimizes AI models for edge devices, <u>has raised \$6 million in funding.</u> <ul style="list-style-type: none">– Led the Core project of achieving Real-Time Super-Resolution on Mobile platform, we are the first to achieve real-time SR inference (with only tens of milliseconds per frame) for implementing 720p resolution with competitive image quality (in terms of PSNR and SSIM) on mobile platforms (ICCV-21)– Worked on the implementations of vision applications on resource limited platforms <ul style="list-style-type: none">• Proposed a unified DNN weight pruning framework with dynamically updated regularization terms bounded by the designated constraint to achieve extremely high compression rates.• Contributed to the development of Radiofrequency Machine Learning System on resource limited platform, such as mobile phone, FPGA, TX2.• Published papers in top-tier conferences (ICCV, CVPR, DAC, etc.).	Boston, MA Sep 2019 – present Feb 2020 – present
University of Toronto <i>Research Assistant advised by Prof. Baochun Li @ Department of ECE</i> • <i>Project: Scheduling Machine Learning Jobs with Reinforcement Learning</i> Content: Adopted reinforcement learning to find the scheduling decision for distributed machine learning training jobs by designing the state, action, and rewards appropriately (IWQoS-19). <ul style="list-style-type: none">– Helped to model the scheduling problem for reinforcement learning agent with carefully designed state space, action space, and reward– Simulated the results and compared it with the state-of-art method	Toronto, ON, Canada Jul 2018 – Feb 2019

- *Project: Stochastic Computing and Universal Approximation Theory*
Content: Proved the equivalence of Stochastic Computing-based Neural Networks (SCNN) and BNN by using Universal Approximation theory (AAAI-19)
 - Calculated error bound of SCNN by using Universal Approximation theory and Chebyshev's Inequality, and got the error bound of BNN by using the equivalence
 - Calculated the energy complexity of SCNN, and used the equivalence to get the energy complexity of BNN

PUBLICATIONS

Conference Papers, * means equal contribution.

- Zifeng Wang*, **Zheng Zhan***, Yifan Gong et al, "SparCL: Sparse Continual Learning on the Edge". NeurIPS 2022.
- Yifan Gong*, **Zheng Zhan***, Pu Zhao et al, "All-in-One: A Highly Representative DNN Pruning Framework for Edge Devices with Dynamic Power Management". ICCAD 2022.
- Yushu Wu*, Yifan Gong*, Pu Zhao, Yanyu Li, **Zheng Zhan** et al, "Compiler-Aware Neural Architecture Search for On-Mobile Real-time Super-Resolution". ECCV 2022.
- **Zheng Zhan***, Yifan Gong*, Pu Zhao* et al, "Achieving on-Mobile Real-Time Super-Resolution with Neural Architecture and Pruning Search". ICCV 2021.
- Geng Yuan, Xiaolong Ma, Wei Niu, Zhengang Li, Zhenglun Kong, Ning Liu, Yifan Gong, **Zheng Zhan** et al, "MEST: Accurate and Fast Memory-Economic Sparse Training Framework on the Edge". NeurIPS 2021. (**Spotlight paper, top 3%**)
- Zhengang Li*, Geng Yuan*, Wei Niu*, Pu Zhao, Yanyu Li, Yuxuan Cai, Xuan Shen, **Zheng Zhan** et al, "NPAS: A Compiler-Aware Framework of Unified Network Pruning and Architecture Search for Beyond Real-Time Mobile Acceleration". CVPR 2021. (**Oral paper, top 5%**)
- Tianyun Zhang, Xiaolong Ma, **Zheng Zhan** et al, "A Unified DNN Pruning Weight Framework Using Reweighted Method". DAC 2021.
- Yifan Gong, **Zheng Zhan**, Zhengang Li et al, "A Privacy-Preserving-Oriented DNN Pruning and Mobile Acceleration Framework". GLSVLSI 2020.
- Yifan Gong, Baochun Li, Ben Liang, **Zheng Zhan**, "Chic: Experience-driven Scheduling in Machine Learning Clusters". IWQoS 2019.
- Yanzhi Wang, **Zheng Zhan**, Liang Zhao et al, "Universal Approximation Property and Equivalence of Stochastic Computing-based Neural Networks and Binary Neural Networks". AAAI 2019.

Journal Papers

- Yifan Gong*, Geng Yuan*, **Zheng Zhan** 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.
- Tong Jian, Yifan Gong, **Zheng Zhan** et al, "Radio Frequency Fingerprinting on the Edge", *IEEE Transactions on Mobile Computing*, 2021.

SKILLS

- Research: Machine Learning, Model Compression, Computer Vision, AI in Communications.
- Software: PyTorch, TensorFlow.
- Programming Languages: Python, C/C++ (wrote a Remote Test Harness), C# (wrote a Remote Build Server), MATLAB.