ZHENG ZHAN

140 The Fenway, Boston, MA 02115

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

EDUCATION

Northeastern University Boston, MA
Ph.d. Candidate in Computer Engineering, GPA: 4.0/4.0 Sep 2019 – May 2023 (expected)

- With a focus on various topics in Machine Learning, especially *Model Compression and Acceleration*.

Syracyse University Syracuse, NY
Master of Science in Computer Engineering, GPA: 3.833/4.0 Sep 2017 – May 2019

Xidian University Xi'an, Shaanxi, China
Bachelor of Engineering in Electronic Science and Technology Sep 2013 – Jun 2017
Excellent Class **(Undergraduate Honor Program)**

EXPERIENCE

Samsung Research America Mountain View, CA
Ph.D. Research Intern May 2022 – Aug 2022

- *Project: Efficient ViT using linear self-attention for large inputs*
  Content: Worked on efficient transformer using linear self-attention for large inputs, especially for improving the efficiency of detecting small objects in high resolution images.
  - We studied the linear self-attention, shifted window of swin transformer and sparse query of QueryDet. We develop a pipeline with the techniques above and with a backbone of Feature Pyramid Network on a 2K high resolution object detection dataset (VisDrone).

Lawrence Livermore National Laboratory Livermore, CA
Ph.D. Research Intern @ DSSI program May 2021 – Aug 2021

- *Project: Multi-Prize Lottery Tickets of Vision Transformer*
  Content: Worked on finding multi-prize lottery tickets for vision transformer to achieve a high performance and robustness using sparse binary network.
  - Developed a framework based on the BInarize-PRune OPtimizer to find the multi-prize lottery tickets of ViT
  - Conducted the experiments on efficient Binary Neural Networks (BNN) training algorithm

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

- We develop the SparCL which explores sparsity for efficient continual learning and achieves both training acceleration and accuracy preservation through the synergy of three aspects: weight sparsity, data efficiency, and gradient sparsity (NeurIPS-22).
- *Project: Compression-Compilation Co-design (CoCoPIE)* Feb 2020 – present
  Content: CoCoPIE , a **startup** developing a platform that optimizes AI models for edge devices, **has raised $6 million in funding**.
  - 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, ECCV-22)
  - Worked on the implementations of vision applications on resource limited platforms
- We propose All-in-One, a highly representative pruning framework (a single neural network contain multiple sparsity ratio) to work with dynamic power management using DVFS. (ICCAD-22)
- Published papers in top-tier conferences (NeurIPS, ICCV, CVPR, ECCV, DAC, ICCAD etc.).

University of Toronto Toronto, ON, Canada
Research Assistant advised by Prof. Baochun Li @ Department of ECE Jul 2018 – Feb 2019

- *Project: Scheduling Machine Learning Jobs with Reinforcement Learning*
  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).
  - 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

Syracuse University Syracuse, NY
Research Assistant advised by Prof. Yanzhi Wang @ College of ECS Sep 2017 – May 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.