

ZHENG ZHAN

140 The Fenway, Boston, MA 02115

☎ (857)891-6751 ✉ zhan.zhe@northeastern.edu 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 Deep Learning and its applications.	
Syracuse 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

Lawrence Livermore National Laboratory	Livermore, CA
Ph.D. Research Intern @ DSSI program	May 2021 – Aug 2021
• <i>Project: Multi-Prize Lottery Tickets of Vision Transformer</i>	May 2021 – present
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 vision transformer	
– 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
• <i>Project: Compression-Compilation Co-design (CoCoPIE)</i>	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)	
– Worked on the implementations of vision applications on resource limited platforms	
• 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.).	

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

- **Zheng Zhan**[†], Yifan Gong[†], Pu Zhao[†] et al, "Achieving on-Mobile Real-Time Super-Resolution with Neural Architecture and Pruning Search", in *Proceedings of the IEEE International Conference on Computer Vision (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. In *Conference on Neural Information Processing Systems (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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. (**Oral paper, top 5%**)
- Tianyun Zhang, Xiaolong Ma, **Zheng Zhan** et al, "A Unified DNN Pruning Weight Framework Using Reweighted Method", *Design Automation Conference (DAC)*, 2021.
- Wei Niu[†], Pu Zhao[†], **Zheng Zhan** et al, "Towards Real-Time DNN Inference on Mobile Platforms with Model Pruning and Compiler Optimization" in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2020. (Demonstration Track in Proceedings) (Acceptance Rate: 26.7%)
- Yifan Gong, **Zheng Zhan**, Zhengang Li et al, "A Privacy-Preserving-Oriented DNN Pruning and Mobile Acceleration Framework", *ACM Great Lakes Symposium on VLSI (GLSVLSI)*, 2020.
- Yifan Gong, Baochun Li, Ben Liang, **Zheng Zhan**, "Chic: Experience-driven Scheduling in Machine Learning Clusters", in *Proceedings of the International Symposium on Quality of Service (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", in *Proceedings of the AAAI Conference on Artificial Intelligence (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.

SERVICE

Program Committee & Reviewer of AAAI 2022