Sai Qian Zhang

zhangs@g.harvard.edu | www.saiqianzhang.com | www.github.com/saizhang0218

Research Interest

My main research interest lies in software/hardware codesign for efficient deep neural network (DNN) implementation. I am also interested in multi-agent reinforcement learning (MARL) and its application.

Education

Harvard University — Cambridge, MA
Doctor of Philosophy in Computer Science — *Aug. 2016 – Sep. 2021 (expected)*

University of Toronto — Toronto, ON
Master of Applied Science in Electrical Engineering — *Aug. 2013 – May 2016*
Master of Science in Statistics — *Aug. 2015 – May 2016*

Professional Experience

Research Engineer Intern — Jul. 2019 - Sep. 2019
Hardware Research Group, Mediatek — *San Jose, CA*

- Designed an efficient routing network on 3D-IC for coordinated parallel use of a plurality of systolic arrays (SAs) in performing deep neural network (DNN) inference.
- This work was published in International Conference on Application-specific Systems, Architectures and Processors (ASAP), 2019.

Research Engineer Intern — May 2017 - Aug. 2017
Wireless Research Group, Intel Labs — *Santa Clara, CA*

- Applied machine learning technique (Conditional random fields) to predict cell quality for aerial drone operation.
- This work was published in IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC) 2018.

Software Developer Intern — May 2011 - Jun. 2012
Signal Integrity Group, Advanced Micro Devices — *Markham, ON*

- Developed software tools to perform the geometric modeling of the vias, voids, traces on PCB package.

Selected Publications

1. **S. Q. Zhang**, B. McDanel, H. T. Kung, X. Dong. Training for Multi-resolution Inference Using Reusable Quantization Terms, in ACM ASPLOS, 2021.
2. **S. Q. Zhang**, J. Lin, Q. Zhang. Succinct and Robust Multi-Agent Communication With Temporal Message Control, in NeurIPS, 2020.
3. **S. Q. Zhang***, B. McDanel*, H. T. Kung*. Term Quantization: Furthering Quantization at Run Time, in ACM/IEEE Supercomputing, 2020 (* equal contribution).
4. **S. Q. Zhang**, J. Lin, Q. Zhang. Efficient Communication in Multi-Agent Reinforcement Learning via Variance Based Control, in NeurIPS, 2019.
5. **S. Q. Zhang***, B. McDanel*, H. T. Kung, X. Dong. Full-stack Optimization for Accelerating CNNs with FPGA Validation, in ACM ICS, 2019 (* equal contribution).

Technical Skills

Languages: Python, C/C++, Verilog, SystemVerilog, Matlab, R

Libraries: Pytorch, Chainer, NumPy

Awards

NSERC Postgraduate Scholarships by Canadian Natural Sciences and Engineering Research Council, 2016-2019
Best paper award at International Conference on Communication (ICC), 2015
ECE Faculty Undergraduate Summer Research Award, 2012, 2013
ADEL S. SEDRA Outstanding Student Award by University of Toronto, 2009

Professional Activities

Conference reviewer: ICML 2021, NeurIPS 2020, AAAI 2020, ICRA 2020, ICC 2019