Sai Qian Zhang
33 Oxford Street, Cambridge, MA, 02138
zhangs@g.harvard.edu
www.saiqianzhang.com

RESEARCH INTEREST

My main research interest lies in software/hardware codesign for efficient deep neural network (DNN) implementation. I develop efficient DNN algorithms which achieve higher energy efficiency and lower latency than prior work while maintaining near state-of-the-art classification performance. Additionally, I am also interested in multi-agent reinforcement learning (MARL) and its application on hardware system.

EDUCATION

Harvard University
Ph.D. in Computer Science — September 2021 (expected)
Advised by Prof. H.T. Kung

University of Toronto
M.A.Sc in Electrical Engineering — May 2016
M.Sc in Statistics — May 2016
B.A.Sc in Electrical Engineering — May 2013

AWARDS

Travel grant, *Neural Information Process System (NeurIPS)* — 2019
NSERC Postgraduate Scholarships by Canadian Natural Sciences and Engineering Research Council — 2016-2019
Best paper award, *International Conference on Communication (ICC)* — 2015
Rogers Scholar — 2014
Dean's Honours List in all 8 academic terms at University of Toronto — 2013
ECE Faculty Undergraduate Summer Research Award — 2013
AMD Appreciation Award — 2012
ECE Alumni Scholarship — 2011
Wallberg Undergraduate Scholarship — 2010
Hosinic Family Scholarship — 2010
ADEL S. SEDRA Outstanding Student Award by University of Toronto — 2009

PUBLICATIONS

Preprints

1. **S. Q. Zhang**, Jieyu Lin, Qi Zhang. *Learning Optimal Client Selection for Efficient Federated Learning Deployment.*

Journal Articles

1. **S. Q. Zhang**, Q. Zhang, A. Tizghadam, B. Park, Hadi Bannazadeh, A. Leon-Garcia, R. Boutaba. *TCAM Space-Efficient Routing in a software-defined network.* Computer Networks 125 (2017): 26-40.
2. **S. Q. Zhang**, Q. Zhang, H. Bannazadeh, A. Leon-Garcia. *Routing Algorithms for Network Function Virtualization Enabled Multicast Topology on SDN.* IEEE transactions on Network and Service Management (2015): 580-594.

Conference Papers

1. **S. Q. Zhang**, B. McDanel, H. T. Kung, X. Dong. *Training for Multi-resolution Inference using Reusable Quantization Terms.* ACM International Conference

on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2021.

2. B. McDanel, H. T. Kung, **S. Q. Zhang**. *Saturation RRAM Leveraging Bit-Level Sparsity Resulting from Term Quantization.* IEEE International Symposium on Circuits and Systems (ISCAS), 2021.
3. **S. Q. Zhang**, J. Lin, Q. Zhang. *Succinct and Robust Multi-Agent Communication With Temporal Message Control.* Neural Information Processing System (NeurIPS), 2020.
4. **S. Q. Zhang**, J. Lin, Q. Zhang. *Adaptive Distributed Convolutional Neural Network Inference at the Network Edge with ADCNN.* International Conference on Parallel Processing (ICPP), 2020.
5. **S. Q. Zhang***, H.T. Kung*, B. McDanel*. *Term Quantization: Furthering Quantization at Run Time.* International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2020 (* equal contribution).
6. Y. Li, X. Dong, **S. Q. Zhang**, H. Bai, Y. Chen, W. Wang. *RTN: Reparameterized Ternary Network.* Thirty-forth AAAI Conference on Artificial Intelligence (AAAI), 2020.
7. J. Lin, K. Dzeparoska, **S. Q. Zhang**, A. Leon-Garcia, N. Papernot . *On the Robustness of Cooperative Multi-Agent Reinforcement Learning.* IEEE Symposium on Security and Privacy (S&P) Deep Learning and Security workshop, 2020.
8. H.T. Kung, B. McDanel, **S. Q. Zhang**. *Packing Sparse Convolutional Neural Networks for Efficient Systolic Array Implementations: Column Combining Under Joint Optimization.* ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2019 (alphabetical author list).
9. **S. Q. Zhang**, J. Lin, Q. Zhang. *Efficient Communication in Multi-Agent Reinforcement Learning via Variance Based Control.* Neural Information Processing System (NeurIPS), 2019.
10. **S. Q. Zhang***, B. McDanel*, H. T. Kung, X. Dong. *Full-stack Optimization for Accelerating CNNs with FPGA Validation.* ACM International Conference on Supercomputing (ICS), 2019 (* equal contribution).
11. H.T. Kung, B. McDanel, **S. Q. Zhang**. X. Dong, C. Chen. *Maestro: A Memory-on-Logic Architecture for Coordinated Parallel Use of Many Systolic Arrays.* IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP), 2019.
12. H.T. Kung, B. McDanel, **S. Q. Zhang**. C. T. Wang, J. Cai, C. Y. Chen, V. Chang, M. F. Chen, J. Sun, D. Yu. *Systolic Building Block for Logic-on-Logic 3D-IC Implementations of Convolutional Neural Networks.* IEEE International Symposium on Circuits and Systems (ISCAS), 2019.
13. H.T. Kung, B. McDanel, **S. Q. Zhang**. "Adaptive Tiling: Applying Fixed-size Systolic Arrays To Sparse Convolutional Neural Networks". International Conference on Pattern Recognition (ICPR), 2018 (alphabetical author list).
14. H.T. Kung, B. McDanel, **S. Q. Zhang**. "Mapping Systolic Arrays Onto 3D Circuit Structures: Accelerating Convolutional Neural Network Inference". IEEE Workshop on Signal Processing Systems (SiPS), 2018 (alphabetical author list).

15. **S. Q. Zhang**, F. Xue, N. Himayat, S. Talwar, H.T. Kung. *A Machine Learning Assisted Cell Selection Method for Drones in Cellular Networks.* IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC) 2018.
16. **S. Q. Zhang**, Q. Zhang, A. Tizghadam, B. Park, Hadi Bannazadeh, A. Leon-Garcia, R. Boutaba. *Sector: TCAM Space Aware Routing on SDN.* International Teletraffic Congress (ITC), 2016.
17. **S. Q. Zhang**, A. Tizghadam, B. Park, H. Bannazadeh, A. Leon-Garcia. *Joint NFV Placement and Routing for Multicast Service on SDN.* IEEE/IFIP Network Operations and Management Symposium (NOMS), 2016.
18. **S. Q. Zhang**, P. Yasrebi, A. Tizghadam, H. Bannazadeh, A. Leon-Garcia. *Fast Network Flow Resumption for Live Virtual Machine Migration on SDN.* IEEE International Conference on Network Protocol (ICNP) workshop on Control, Operation and Application in SDN Protocol, 2015.
19. **S. Q. Zhang**, Q. Zhang, H. Bannazadeh, A. Leon-Garcia. *Network Function Virtualization Enabled Multicast Routing on SDN.* IEEE International Conference on Communication (ICC), 2015 **(best paper award)**.
20. Q. Zhang, **S. Q. Zhang**, A. Leon-Garcia, R. Boutaba. *Aurora: Adaptive Block Replication in Distributed File Systems.* IEEE International Conference on Distributed Computing Systems (ICDCS), 2015.
21. Q. Zhang, **S. Q. Zhang**, H. Bannazadeh, A. Leon-Garcia. *Kaleidoscope: Real-Time Content Delivery in Software Defined Infrastructures.* IEEE/IFIP International Symposium on Integrated Network Management (IM), 2015.

ACADEMIC TALKS

Conference Presentations

ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)	2021
Neural Information Processing System (NeurIPS)	2020
International Conference for High Performance Computing, Networking, Storage and Analysis (SC)	2020
International Conference on Parallel Processing (ICPP)	2020
Neural Information Processing System (NeurIPS)	2019
IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)	2018
IEEE/IFIP Network Operations and Management Symposium (NOMS)	2016
International Teletraffic Congress (ITC)	2016
IEEE International Conference on Communication (ICC)	2015

TEACHING EXPERIENCE

Harvard University

- COMPSCI 242 Computing at Scale (2020 Spring, head TA)
- COMPSCI 143 Computer Networks (2017 Fall, 2018 Spring, 2019 Fall)
- COMPSCI 144/244 Networks Design Projects (2019 Spring)
- ENG-SCI 201 Decision Theory (2017 Spring, 2018 Spring, head TA)

University Of Toronto

- ECE 361 Computer Networks (2014 Fall, 2014 Spring, 2015 Fall)
- ECE 297 Design and Communication (2015 Spring)

TECHNICAL SERVICE

Journal Reviewer

IEEE/ACM Transactions on Networking (ToN) 2021

Conference Reviewer

International Conference on Machine Learning (ICML) 2021
Neural Information Processing System (NeurIPS) 2020, 2021
AAAI Conference on Artificial Intelligence (AAAI) 2020
IEEE International Conference on Robotics and Automation (ICRA) 2020
IEEE International Conference on Communication (ICC) 2017

PROFESSIONAL EXPERIENCE

Research Engineer Intern Mar. 2021 - Now
Brainwave Team, Microsoft Research, Redmond, WA

- Design sparse transformer architecture for efficient hardware implementation on Brainwave FPGA.

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.

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.

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

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

SKILLS SUMMARY

Languages & Software: python, C/C++, Matlab, Verilog, System Verilog
Libraries: PyTorch, TensorFlow, Chainer, NumPy.

EXTRA-CURRICULAR ACTIVITIES

- Singing:
  - Champion of 2013 Chinese Student Singing Competition
  - Champion of 2014 "Singing Loud" Singing Competition
  - Top ten of "The Voice of China Toronto District"
- Jazz Drum
- Piano (Grade 9)
- 1st Running up of 2013 University of Toronto Chinese Student Pool Competition
- Top four in 2007 New Zealand national high school table tennis tournament