Ritchie Zhao

409 Auburn Street #2 – Ithaca, New York 14850 – USA

rzhao01.github.io rz252@cornell.edu rzhao01 +1 607-379-7088

RESEARCH INTERESTS

Hardware specialization for machine learning, efficient neural network architectures, high-level synthesis (HLS) for FPGAs, methodology and automation for hardware design

SKILLS

Python, C, C++, Vivado HLS, TensorFlow, PyTorch, Verilog, LaTek

EDUCATION

Cornell University **Ithaca, NY**
PhD in Electrical and Computer Engineering *Aug 2014 - Sept 2019 (Expected)*
Advisor: Zhiru Zhang

University of Toronto **Toronto, ON**
BS in Electrical and Computer Engineering *Sept 2009 - May 2014*
GPA: 3.92, Graduated with Honors

RESEARCH PROJECTS

DNN Quantization with Outlier Channel Splitting [1] *Aug 2018 - Feb 2019*

- Proposed a technique to improve DNN quantization without retraining, targeting scenarios where training for quantization is not possible (e.g. ML service providers with black-box customer models).
- Results show improved accuracy over state-of-the-art clipping methods.
- Open-source code available at https://github.com/cornell-zhang/dnn-quant-ocs.

Efficient DNNs with Unitary Group Convolutions [2] *Aug 2017 - May 2018*

- Explored the composition of group convolutions with unitary transforms to build efficient DNN architectures; this idea generalizes ShuffleNet and CirCNN from literature.
- Proposed to use the hardware-efficient Hadamard transform. Hadamard networks outperform ShuffleNet with no parameter/multiply overhead and matches CirCNN with fewer multiplies.

Binarized Neural Network Accelerator for FPGA [4,5] *May 2016 – May 2017*

- Designed an accelerator for BNNs on an embedded FPGA platform (Xilinx Zedboard). Achieved 15x speedup over embedded GPU with less power and negligible accuracy loss.
- The same accelerator was ported to Celerity, an academic ASIC taped-out to silicon at TSMC 16nm process.

Synthesizable Halide-to-Verilog *Jan 2016 – Apr 2016*

- Modified the Halide embedded DSL to generate HLS-synthesizable code and to support HLS pragmas.
- Project was passed off to another PhD student.

Decoupled Data Structures for HLS [6] *Jun 2015 – Jan 2016*

- Proposed mapping certain data structures (e.g. heaps, hash tables) to a decoupled architectural template. This enables a modular design flow while exploiting parallelism between method calls and the main program.

Joint Scheduling and Mapping for HLS [7] *Aug 2014 – May 2015*

- Developed a mixed-ILP method to jointly perform HLS pipeline scheduling and downstream LUT-mapping.
- Results show significant area savings in logic-heavy designs via cross-layer optimization.
- Implemented the technique as an LLVM pass using IBM ILOG CLEX as the mILP solver.

INDUSTRY EXPERIENCE

Microsoft Research **Redmond, WA**

Research Intern *May 2018 - Aug 2018*

Research Intern *May 2017 - Aug 2017*

- Researched and prototyped novel techniques to quantize DNNs for efficient deployment on the Brainwave FPGA accelerator system. Gathered experimental data on the quantized accuracy of internal models.
- Assisted in engineering a TensorFlow quantization library used in serving production DNN models.
- Proposed and developed new methods to automate the quantization flow of trained models, making quantization more user-friendly for non-experts in ML.
- Contributed results to the first technical publication of Brainwave [3].

IBM **Toronto, ON**

Extreme Blue Technical Intern *May 2014 - Aug 2014*

- Assisted the development of IBM Dash, a mathematical programming language which generates optimized code for heterogeneous hardware systems.
- Engineered the GPU back-end for Dash using CUDA. Obtained 16x speedup on option pricing with Monte Carlo Black-Scholes compared to hand-optimized C.
- Created a presentation and web demo to pitch the project to IBM executives, culminating in a team pitch at the IBM headquarters in Armonk.

Altera Corporation **Toronto, ON**

Engineering Intern *May 2012 - Aug 2013*

- Assisted timing team to create and maintain timing models for Altera's IV and V FPGA families.
- Developed internal features such as piecewise-linear waveform modeling and per-node clock uncertainty.
- Performed end-to-end testing for the Quartus II Timing Analyzer.
- Created software patches for a key Altera customer while coordinating with field engineers.

SELECTED PUBLICATIONS

1. **R. Zhao**, Y. Hu, J. Dotzel, C. De Sa, and Z. Zhang. "Improving Neural Network Quantization without Retraining using Outlier Channel Splitting". *To appear in Int'l Conf. on Machine Learning (ICML)*, Jun. 2019
2. **R. Zhao**, Y. Hu, J. Dotzel, C. De Sa, and Z. Zhang. "Building Efficient Deep Neural Networks with Unitary Group Convolutions". *To appear in Conf. on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019
3. E. Chung, J. Fowers, K. Ovtcharov, M. Papamichael, A. Caulfield, T. Massengill, M. Liu, D. Lo, S. Alkalay, M. Haselman, M. Abeydeera, L. Adams, H. Angepat, C. Boehn, D. Chiou, O. Firestein, A. Forin, K. S. Gatlin, M. Ghandi, S. Heil, K. Holohan, A. El Husseini, T. Juhasz, K. Kagi, R. K. Kovvuri, S. Lanka, F. van Megen, D. Mukhortov, P. Patel, B. Perez, A. G. Rapsang, S. K. Reinhardt, B. D. Rouhani, A. Sapek, R. Seera, S. Shekar, B. Sridharan, G. Weisz, L. Woods, P. Y. Xiao, D. Zhang, **R. Zhao**, and D. Burger. "Serving DNNs in Real Time at Datacenter Scale with Project Brainwave". *IEEE Micro*, Mar. 2018
4. S. Davidson, S. Xie, C. Torng, K. Al-Hawaj, A. Rovinski, T. Ajayi, L. Vega, C. Zhao, **R. Zhao**, S. Dai, A. Amarnath, B. Veluri, P. Gao, A. Rao, G. Liu, R. K. Gupta, Z. Zhang, R. G. Dreslinski, C. Batten, and M. B. Taylor. "The Celerity Open-Source 511-Core RISC-V Tiered Accelerator Fabric: Fast Architectures and Design Methodologies for Fast Chips". *IEEE Micro*, Mar. 2018
5. **R. Zhao**, W. Song, W. Zhang, T. Xing, J.-H. Lin, M. Srivastava, R. Gupta, and Z. Zhang. "Accelerating Binarized Convolutional Neural Networks with Software-Programmable FPGAs". *Int'l Symp. on Field-Programmable Gate Arrays (FPGA)*, Feb. 2017
6. **R. Zhao**, G. Liu, S. Srinath, C. Batten, and Z. Zhang. "Improving High-Level Synthesis with Decoupled Data Structure Optimization". *Design Automation Conference (DAC)*, Jun. 2016
7. **R. Zhao**, M. Tan, S. Dai, and Z. Zhang. "Area-Efficient Pipelining for FPGA-Targeted High-Level Synthesis". *Design Automation Conference (DAC)*, Jun. 2015