Ritchie Zhao

Redmond, WA, USA

rzhao01.github.io rzhao01@gmail.com rzhao01 +1 503-858-5933

HIGHLIGHTS

Proven innovator in data formats for deep learning. Engineering lead for LLM experiment infrastructure and model quantization library at Microsoft. Key contributor to the Microscaling (MX) data format and MX library (github.com/microsoft/microxcaling).

AREAS OF EXPERTISE

Hardware acceleration for deep learning, efficient LLM training and inference, CUDA kernels for deep learning, high-level synthesis for FPGAs

EDUCATION

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

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

INDUSTRY EXPERIENCE

Microsoft **Redmond, WA**
Senior Data Science Manager *May 2022 - Present*
Senior Hardware Engineer *Sept 2019 - Apr 2022*

- Researched and implemented novel data formats and quantization algorithms for AI hardware accelerators. Ran hundreds of experiments on quantized LLM training and inference.
- Engineering lead and codebase owner for our team's Pytorch training and quantization library. Implemented custom CUDA kernels to speed up quantization operations.
- Key contributor to the Microscaling (MX) Specification, a data format ratified by 7 major companies including NVIDIA, Meta, and Intel. Met with representatives from other companies to align on the format. Led the design of the FP32 to MX conversion protocol. Lead developer for the open-source MX library.
- Key engineer for Microsoft's OpenAI engagement. Ran experiments on highly confidential production models to explore opportunities for acceleration.
- Promoted to manager in 2022, leading a team of two engineers.
- Work had major impact on the Azure Maia 100 accelerator and OpenAI inference workloads. Research led to three publications [1,2,3] and five successful patents.

Microsoft Research **Redmond, WA**
Research Intern *May 2018 - Aug 2018*
Research Intern *May 2017 - Aug 2017*

- Researched and prototyped DNN quantization techniques, targeting deployment on Microsoft's FPGA-based Brainwave accelerator. Ran deep learning experiments to evaluate the quality of quantized models. Gathered experimental data for the first technical publication of Brainwave [7].
- Assisted in engineering a TensorFlow quantization library used in serving production DNN models.
- Proposed a novel differentiable neural architecture search (NAS) method to find the optimal bit-width for each layer in a quantized DNN, leading to a patent.

IBM **Toronto, ON**

Extreme Blue Technical Intern *May 2014 - Aug 2014*

- Assisted the research and development of a Matlab-like programming language which statically compiles to optimized code for CPU or GPU. Worked on the compiler backend.
- Performed engineering on the GPU backend which generated CUDA. Obtained 16x speedup on option pricing with Monte Carlo Black-Scholes compared to hand-optimized C.

Altera Corporation (now Intel) **Toronto, ON**

Engineering Intern *May 2012 - Aug 2013*

- Performed end-to-end testing for the Quartus II Timing Analyzer. Maintained timing models for Altera's IV and V FPGA families.
- Created software patches for a key Altera customer while coordinating with field engineers.

SELECTED PUBLICATIONS

1. B. Darvish Rouhani, **R. Zhao**, A. More, M. Hall, A. Khodamoradi, S. Deng, D. Choudhary, M. Cornea, E. Dellinger, K. Denolf, S. Dusan, V. Elango, M. Golub, A. Heinecke, P. James-Roxby, D. Jani, G. Kolhe, M. Langhammer, A. Li, L. Melnick, M. Mesmakhosroshahi, A. Rodriguez, M. Schulte, R. Shafipour, L. Shao, M. Siu, P. Dubey, P. Micikevicius, M. Naumov, C. Verrilli, R. Wittig, D. Burger, E. Chung "Microscaling Data Formats for Deep Learning". *arXiv e-print*, Oct. 2023
2. B. Darvish Rouhani, **R. Zhao**, V. Elango, R. Shafipour, M. Hall, M. Mesmakhosroshahi, A. More, L. Melnick, M. Golub, G. Varatkar, L. Shao, G. Kolhe, D. Melts, J. Klar, R. L'Heureux, M. Perry, D. Burger, E. Chung "With Shared Microexponents, A Little Shifting Goes a Long Way". *Int'l Conf. on Computer Architecture (ISCA) Industry Track*, Jun. 2023
3. B. Darvish Rouhani, D. Lo, **R. Zhao**, M. Liu, J. Fowers, K. Ovtcharov, A. Vinogradsky, S. Massengill, L. Yang, R. Bittner, A. Forin, H. Zhu, T. Na, P. Patel, S. Che, L. C. Koppaka, X. Sogn, S. Som, K. Das, S. Reinhardt, S. Lanka, E. Chung, and D. Burger. "Pushing the Limits of Narrow Precision Inference at Cloud Scale with Microsoft Floating Point". *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2020
4. Y. Zhang, **R. Zhao**, W. Hua, N. Xu, G. E. Suh, Z. Zhang "Precision Gating: Improving Neural Network Efficiency with Dynamic Dual-Precision Activations". *Int'l Conf. on Learning Representations (ICLR)*, Apr. 2020
5. **R. Zhao**, Y. Hu, J. Dotzel, C. De Sa, and Z. Zhang. "Improving Neural Network Quantization without Retraining using Outlier Channel Splitting". *Int'l Conf. on Machine Learning (ICML)*, May. 2019
6. **R. Zhao**, Y. Hu, J. Dotzel, C. De Sa, and Z. Zhang. "Building Efficient Deep Neural Networks with Unitary Group Convolutions". *Conf. on Computer Vision and Pattern Recognition (CVPR)*, May. 2019
7. 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. Darvish 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
8. **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
9. **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
10. **R. Zhao**, M. Tan, S. Dai, and Z. Zhang. "Area-Efficient Pipelining for FPGA-Targeted High-Level Synthesis". *Design Automation Conference (DAC)*, Jun. 2015

GRANTED PATENTS

1. K. Ovtcharov, E. Chung, V. Akhlaghi **R. Zhao** "Quantization-Aware Neural Architecture Search". *US Patent 11790212*, 10/17/2023
2. D. Burger, E. Chung, B. Darvish Rouhani, D. Lo, **R. Zhao** "Flow for Quantized Neural Networks". *US Patent 11645493*, 05/09/2023
3. K. Ovtcharov, E. Chung, V. Akhlaghi, **R. Zhao** "Differential Bit Width Neural Architecture Search". *US Patent 11604960*, 03/14/2023
4. E. Chung, D. Lo, J. Zhang, **R. Zhao** "Residual Quantization for Neural Networks". *US Patent 11586883*, 02/21/2023
5. E. Chung, D. Lo, **R. Zhao** "Outlier Quantization for Training and Inference". *US Patent 11574239*, 02/07/2023

RESEARCH PROJECTS

DNN Quantization with Outlier Channel Splitting [5] *Aug 2018 - Feb 2019*

- Proposed a technique to improve DNN quantization without retraining, targeting post-training quantization for inference.
- Results show improved accuracy over state-of-the-art post-traininng clipping methods. Open-source code available.

Efficient DNNs with Unitary Group Convolutions [6] *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 [8] *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 [9] *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 [10] *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.