Huizi Mao

PERSONAL INFO

Email: ralphmao95 at gmail.com
Website: http://stanford.edu/~huizi/

EDUCATION

June 2021 (expected) — **Ph.D Candidate**, Electrical Engineering, Stanford University
Advisor: Bill Dally
Research Interests: Deep Learning; Computer Architecture

July 2016 — **B.E.** (with honors), Electronic Engineering, Tsinghua University
Thesis: Task-related Compression Methods of Deep Neural Networks

July 2016 — **B.S.** (secondary), Mathematics, Tsinghua University
Thesis: Scheduling Algorithms for Distributed Principal Component Analysis

SELECTED PUBLICATIONS

[1] **Huizi Mao**, Xiaodong Yang, William J Dally. "A Delay Metric for Video Object Detection: What Average Precision Fails to Tell", International Conference on Computer Vision (ICCV), 2019

[2] **Huizi Mao**, Taeyoung Kong, William J Dally. "CaTDet: Cascaded Tracked Detector for Efficient Object Detection from Video", The Conference on Systems and Machine Learning (SysML), 2019

[3] **Huizi Mao**, Song Han, Jeff Pool, Wenshuo Li, Xingyu Liu, Yu Wang, William J Dally. "Exploring the Regularity of Sparse Structure in Convolutional Neural Networks", CVPR workshop, 2017

[4] Song Han, Xingyu Liu, **Huizi Mao**, Jing Pu, Ardavan Pedram, Mark Horowitz, William J. Dally, "EIE: Efficient Inference Engine on Compressed Deep Neural Network", International Symposium on Computer Architecture (ISCA), 2016

[5] **Huizi Mao**, Song Yao, Tianqi Tang, Boxun Li, Jun Yao, Yu Wang, "Towards Real-Time Object Detection on Embedded Systems", in IEEE Transactions on Emerging Topics in Computing, vol.PP, no.99, pp.1-1

[6] Song Han, **Huizi Mao**, William J. Dally, "Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding", International Conference on Learning Representations (ICLR), 2016 (Best Paper Award)

EXPERIENCE

June 2020-Sept 2020 — **Intern**, Google Research
Mentor: Bo Chen
Object detection and keypoint tracking with temporal CenterNet

June 2019- Sept 2019 **Intern**, NVIDIA Autonomous Vehicle
Mentor: Yue Wu
Explored hybrid CNN/RNN models for temporal motion prediction from monocular camera for autonomous vehicles. Built the sequence training framework.

June 2018 - Sept 2018 **Intern**, NVIDIA GPU Architecture
Mentor: Paulius Micikevicius
Explored special types of sparsity, including impact of sparse training on RNN accuracy and potential speed-ups on Turing GPU. Obtained more than 2x speedup for low-batch inference compared with cuBLAS.

June 2017 - Sept 2017 **Intern**, Facebook Applied Machine Learning
Mentor: Peter Vajda
Developed an automated model search framework upon FBLearner Flow that enables management of hundreds of gpus in parallel. Designed a better neural network model for classification and detection. Optimized the model on mobile platform.

June 2016- Feb 2014 **Undergraduate Researcher**, NICS Lab, Tsinghua University
Advisor: Yu Wang
Co-developed a speech recognition framework with multi-GPU acceleration. Collaborated with senior students to design an FPGA-based neural network accelerator for image recognition.

Sept 2015- July 2015 **Visiting Researcher**, CVA Group, Stanford University
Advisor: Bill Dally
Developed a neural network compressing pipeline with ultra-high compression rates. Designed and simulated customized circuits for sparse neural networks.

SCHOLARSHIPS AND AWARDS

2018 NVIDIA Graduate Fellowship
2016 Outstanding Undergraduate Thesis, Tsinghua University
2015 Winner, Low-Power Image Recognition Challenge (LPIRC), San Francisco
2014 National Scholarship of China
2013 Tsinghua-Evergrande Scholarship