Hanrui Wang

38-344, MIT, 50 Vassar Street ⋄ Cambridge, MA, US
hanrui@mit.edu ⋄ https://hanruiwang.me

RESEARCH INTERESTS

Computer architecture and algorithm co-design for efficient sparse computations and machine learning;
Machine learning for systems;
Efficient machine learning and its applications to Natural Language Processing and Computer Vision.

EDUCATION

Massachusetts Institute of Technology, Cambridge, MA, US — *2018.09 - present*
Ph.D. Student in Department of Electrical Engineering and Computer Science — GPA: **5/5**
Advisor: Professor Song Han

Massachusetts Institute of Technology, Cambridge, MA, US — *2018.09 - 2020.05*
M.S. in Electrical Engineering and Computer Science — GPA: **5/5**

Fudan University, Shanghai, China — *2014.09 - 2018.07*
B.Eng. (Honours) in Electrical Engineering — GPA: **3.91/4**(1^{st}/256)
School of Information Science & Technology, School of Microelectronics

University of California, Los Angeles, LA, US — *2017.07 - 2017.09*
CSST Program Exchange Student, CS Department — GPA: **4/4**(1^{st}/94)

PUBLICATIONS

[1] **Hanrui Wang**, Zhanghao Wu, Zhijian Liu, Han Cai, Ligeng Zhu, Chuang Gan, Song Han, "HAT: Hardware-Aware Transformers for Efficient Natural Language Processing," *ACL 2020*.

[2] **Hanrui Wang**, Kuan Wang, Jiacheng Yang, Linxiao Shen, Nan Sun, Hae-Seung Lee, Song Han, "GCN-RL Circuit Designer: Transferable Transistor Sizing with Graph Neural Networks and Reinforcement Learning," *DAC 2020*.

[3] Zhekai Zhang*, **Hanrui Wang***, Song Han, William J. Dally, "SpArch: Efficient Architecture for Sparse Matrix Multiplication," *HPCA 2020*. **(*equal contribution)**

[4] **Hanrui Wang**, Jiacheng Yang, Hae-Seung Lee, Song Han, "Learning to Design Circuits," *NeurIPSW 2018, Oral*.

[5] **Hanrui Wang**, Yize Jin, Liming Wang, Xiaoyang Zeng, Yibo Fan, "A Hardware Friendly Stereo Match Refinement Algorithm Using Disparity Gradient Based Region Growth Method," *IEEE ICSICT 2016, Oral*.

[6] Zhongxia Yan, **Hanrui Wang**, Demi guo, Song Han, "MicroNet for Efficient Language Modeling," *JMLR 2020, Champion of NeurIPS 2019 MicroNet Competition*.

[7] Liming Wang, **Hanrui Wang**, Yize Jin, Xiaoyang Zeng, Yibo Fan, "Hardware Friendly Algorithm of HR Real Time Stereo Matching for Automatic Drive," *IEEE ICSICT 2016*.

[8] Hongzi Mao, Parimarjan Negi, Akshay Narayan, **Hanrui Wang**, *et al*, "Park: An Open Platform for Learning-Augmented Computer Systems, " *NeurIPS 2019, ICMLW 2019 Oral, Best Paper Award*.

[9] Jason Cong*, Zhenman Fang*, Michael Lo*, **Hanrui Wang***, Jingxian Xu*, Shaochong Zhang*, **(*Alphabetical)**, "Understanding Performance Differences of FPGAs and GPUs," *FCCM 2018*.

[10] Yihui He*, Ji Lin*, Zhijian Liu, **Hanrui Wang**, Li-Jia Li, Song Han, "AMC: AutoML for Model Compression and Acceleration on Mobile Devices," *ECCV 2018*.

[11] Tianzhe Wang, Kuan Wang, Han Cai, Ji Lin, Zhijian Liu, **Hanrui Wang**, Yujun Lin, Song Han, "APQ: Joint Search for Network Architecture, Pruning and Quantization Policy," *CVPR 2020*.

[12] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, **Hanrui Wang**, Song Han, "Searching Efficient 3D Architectures with Sparse Point-Voxel Convolution," *ECCV 2020*.

RESEARCH EXPERIENCES

Massachusetts Institute of Technology, Cambridge, MA, US *2018.09 - present*
Advisor: Professor Song Han

Project: Hardware-Efficient Natural Language Processing

- Proposed two new techniques, *heterogeneous layers* and *arbitrary encoder-decoder attention* to enlarge search space and improve Transformer model performance.
- Designed a Transformer SuperNet to include all design choices in the search space and conducted hardware-latency aware Neural Architecture Search to find a low-latency high-performance model. Achieved 3× speedup and 3.7× smaller model size over Transformer baseline on NMT task. [1]
- Attended the NeurIPS 2019 MicroNet efficient NLP challenge and won the champion. [6]
- Designed specialized hardware architecture to support token pruning and head pruning of transformers, achieving 3.0× speedup and 3.2× energy saving over prior state-of-the-art.

Project: Efficient Sparse Matrix Multiplication Accelerator

- Proposed to improve data reuse of both input matrix and output matrix of outer-product based sparse matrix multiplication, achieving 4× speedup and 6× energy saving than previous state-of-the-art.
- Improved output reuse by pipelining two stages of outer-product on-chip, reducing column number of input matrix and using a Huffman-Tree scheduler to schedule the computation sequence. Improved input reuse using a matrix row prefetcher equipped with a cache. [3]

Project: Machine Learning for EDA Systems

- Proposed to leverage reinforcement learning and graph neural networks in analog IC transistor sizing, achieving better Figure of Merits (FoM) on four circuits than Bayesian Optimization, Evolutionary Algorithms, random search and human expert designs.
- Conducted knowledge transfer between circuits in different technology nodes and different circuit topologies with the RL agent to accelerate the design cycle. [2][4][8]

Project: Efficient Speech Recognition and Keyword Spotting System

- Built KWS systems running on laptops and Android cell phones in real time.
- Attended 2018 Kaggle TensorFlow Speech Recognition Challenge and won a bronze medal.

University of California, Los Angeles, LA, US *2017.07 - 2017.09*
Advisor: Professor Jason Cong

Project: Performance & Architecture Comparison between FPGAs & GPUs

- Ported 7 computing kernels in Rodinia to FPGAs and analyzed optimality with Roofline models.
- Compared runtime and FLOPs of computing kernels on FPGAs & GPUs, and analyzed reasons for performance differences based on hardware architecture and algorithm features. [9]

Fudan University, Shanghai, China *2016.09 - 2018.07*
Advisor: Professor C.-J. Richard Shi

Project: Energy Efficient General Purpose Machine Learning Digital Circuits Design

- Proposed a CNN computing dataflow called Channel Stationary to maximize data reuse.

- Proposed a channel-wise CNN feature maps storing method, supporting consecutive processing of different layers with no need for inter-layer off-chip data arrangement.

Fudan University, Shanghai, China *2016.01 - 2016.09*
Advisor: Professor Xiaoyang Zeng and Professor Yibo Fan

Project: Dense Two-Frame Stereo Match

- Proposed a disparity refinement algorithm, utilizing disparity gradients at borders of objects to correct erroneously estimated disparities.
- Proposed a video stereo match algorithm, leveraging data correlation between two successive frames to reduce computational complexity, built GPU acceleration system with CUDA to achieve real-time stereo match. [5][7]

WORKING EXPERIENCES

Nvidia Research, MA, US *2020.06 - 2020.08*
Mentor: Mike O'Connor, Donghyuk Lee, Joel Emer. Manager: Steve Keckler

- Worked on efficient sparse matrix computations in Architecture Research Group.

Xilinx, Beijing, China *2018.07 - 2018.08*
Mentor: Shuang Liang, Junbin Wang. Manager: Shaoxia Fang

- Worked on machine learning hardware accelerators.

TALKS

2020 Google
2020 Nvidia Architecture Research
2020 Sogou Speech Lab
2020 T-Head, Alibaba
2020 Mitsubishi Electric Research Labs, Cambridge, MA
2020 Damo Academy, Alibaba
2020 Nvidia ASIC/VLSI Research
2020 MediaTek
2020 Qualcomm Research Center, San Diego, USA
2020 HPCA 2020 Conference, San Diego, USA
2019 NeurIPS 2019 MicroNet Competition, Vancouver, Canada
2019 MLCAD 2019, Banff, Canada
2019 Samsung Annual Research Review, Cambridge, UK
2019 MIT CICS Research Review, Cambridge, USA
2019 Microsystems Annual Research Conference, Bretton Woods, USA
2019 Workshop on Compiler Techniques for Sparse Tensor Algebra, Cambridge, USA
2018 Xilinx, Beijing, China
2018 NeurIPS 2018 ML for Systems Workshop, Canada
2017 UCLA CSST Conference, Los Angeles, USA
2016 IEEE ICSICT Conference, Hangzhou, China

AWARDS

2020 **DAC Best Video Presentation Award**

2020 **DAC Fellowship**
2019 **Best Paper Award** of ICML 2019 Reinforcement Learning for Real Life Workshop
2019 **Champion** of NeurIPS 2019 MicroNet efficient Language Model Competition
2018 **Bronze Medal** in Kaggle TensorFlow Speech Recognition Challenge
2017 **UCLA Cross-Disciplinary Scholars in Science & Technology Research Fellowship**
2017 **UCLA CSST Best Research Award**
2016 **Chun-Tsung Research Fellowship** (launched by 1957 Nobel laureate in Physics, Tsung-Dao Lee)
2017/16/15 **National Scholarship** (Highest honor for undergraduate academic excellence)

SERVICE

Journal Review	TCAS-II, JMLR, TNNLS, TODAES
Conference Review	NeurIPS, SysML, ICCAD, ISCA, ASPLOS, ICITED

TECHNICAL STRENGTHS

Programming	Python, C/C++, Verilog, MATLAB, CUDA
Tools	PyTorch, TensorFlow, Vivado, Hspice