Zhen Dong

PhD Student | Web: https://dong-zhen.com | Email: zhendong@berkeley.edu

RESEARCH INTERESTS

- Model compression for classification/object detection/NLP on embedded platforms.
- Hardware and software co-design for efficient deep learning.
- AutoML and hardware-aware neural architecture search.

EDUCATION

School of Electrical Engineering and Computer Sciences, University of California at Berkeley *Aug 2018 – present*

- Digital Circuits and Computer Architecture (4.00), FPGA RISC-V CPU Lab (4.00).
- Parallel Computing (4.00), Visual Object and Activity Recognition (4.00), Statistical Learning Theory (4.00).

School of Electrical Engineering and Computer Sciences, Peking University, China. *Sept 2014 – Jul 2018*

- **GPA Ranking: 1/327** in School of EECS at Peking University. **3.97 /4.0** (major)
- Major Courses: Micro-Nano Integrated System (4.00), Digital Logic (4.00), Analog Circuits (3.99), Principles of Digital Integrated Circuits (4.00), Advanced Analog Integrated Circuits Design (3.99).
  Computer Systems (3.99), Fundamentals of Solid State Physics (3.98), Semiconductor Device Physics (3.98), Principle of Integrated Circuits Process (3.99). **Ranked 1st in all these major courses.**

AWARDS

- Winner of 2018-2020 Berkeley Fellowship.
- Received waiver for the National College Entrance Exam to enter Peking University, 1st Prize in the National Olympiad in Physics, 1st Prize in the National Physics Competition for college students.
- 2nd place on Visual Wake Word competition at CVPR 2019.
- Reviewer of IEEE Transactions on Electron Devices (TED) and Transactions on Neural Networks and Learning Systems (TNNLS).
- Tang Lixin Scholarship for outstanding students in China. (top 0.5%)
- Tang Lixin 1st Prize Scholarship for graduate students studying abroad. (top 0.05%)
- SenseTime Scholarship, National Scholarship and Fang Zheng Scholarship. (top 1%)
- Pacemaker to Triple-A student and Triple-A student (twice) at Peking University.
- Princeton University Math Competition (PUMac): Top three among all participants in geometry group.
- Top Ten Undergraduate Research Award at PKU EECS.
- Outstanding Graduates at Peking University and Outstanding Graduates in Beijing.

RESEARCH EXPERIENCE

Research Assistant, Electrical Engineering and Computer Sciences, UC Berkeley *Dec 2018 – present*

Advisor: Prof. **Kurt Keutzer**

- **Research on Hessian-AWare Quantization (HAWQ, HAWQ-V2, ZeroQ)**
  1. Propose a second order based method to decide mixed-precision configuration and block-wise fine-tuning order.
  2. Prove theorem to use the trace of Hessian as sensitivity metric and conduct fast Pareto frontier optimization.
  3. Extend HAWQ to segmentation, object detection tasks and achieve state-of-the-art results.
  4. Conduct fast end-to-end quantization without fine-tuning and without using any training/test data.
- **Research on HW-SW Co-design and NAS (HAWQ-V3, CoDeNet)**
  1. Propose efficient deformable operations for object detection on embedded FPGAs.
  2. Design new FPGA-core with ultra-low precision arithmetic.
  3. HW-SW joint architecture search and efficient implementation of mixed-precision NNs on CPU/GPU/FPGAs.
- **Research on Efficient Natural Language Processing (Q-BERT)**
  1. Propose new method to reduce the model size of BERT-base for applications on edge devices.
  2. Use second order information to help reduce communications during distributed training.
  3. Mixed-precision distributed training on the cloud or efficient fine-tuning on the edge.
- **Projects:** Implemented a 3-pipelined RISC-V CPU on Pynq-Z1 FPGA using Verilog.

Research Intern, Facebook AI *June 2020 – Aug 2020*

- Research on efficient natural language processing (NLP) with limited resources.

Brief Summary of Previous Research:

Undergraduate Visiting Researcher (UGVR), Electrical Engineering, Stanford University *Jun 2017 – Sept 2017*
Advisor: Prof. **H.-S. Philip Wong**
Research on utilizing RRAM array for large-scale networks and transfer learning.
Research on building tools based on statistical ML for analyzing energy consumption and delay in 3D RRAM array.

Research Intern, Object Detection Group, SenseTime Corporation *April 2018 – Aug 2018*
Research on 4-bit model compression (both weight and activation) on RetinaNet for SenseTime database.

Research Assistant, Electrical Engineering and Computer Sciences, Peking University *Dec 2016 – Jun 2018*
Advisor: Prof. **Jinfeng Kang**
Research on spike-time-dependent plasticity (STDP) characteristics in Oxide-RRAM for brain-inspired computing.
Research on NVM-based hardware implementation of convolutional neural networks.

SELECTED PUBLICATIONS

[1] **Zhen Dong**, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, Kurt Keutzer. "AMQ: Automatic Mixed-precision Quantization of Neural Networks Based on Hessian Trace," NeurIPS 2020.

[2] **Zhen Dong***, Zhewei Yao*, Amir Gholami*, Michael W. Mahoney, Kurt Keutzer. "HAWQ: Hessian AWare Quantization of Neural Networks with Mixed-Precision," ICCV 2019.

[3] Sheng Shen*, **Zhen Dong***, Jiayu Ye*, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, Kurt Keutzer. "Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT," AAAI 2020.

[4] Yaohui Cai*, Zhewei Yao*, **Zhen Dong*,** Amir Gholami, Michael W. Mahoney, Kurt Keutzer. "ZeroQ: A Novel Zero Shot Quantization Framework," CVPR 2020.

[5] **Zhen Dong**, Zhewei Yao, Yaohui Cai, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, Kurt Keutzer. "Trace-Weighted Quantization of Neural Networks," NeurIPS 2019 Optimization-workshop, Oral Presentation.

[6] **Zhen Dong**, D. Wang, Q. Huang, Y. Gao, Y. Cai, B. Wu, K. Keutzer and J. Wawrzynek "CoDeNet: Algorithm-hardware Co-design for Deformable Convolution," NeurIPS 2019 EMC2 workshop, Oral Presentation.

[7] **Zhen Dong**, Z. Zhou, Z.F. Li, P. Huang, L.F. Liu, X.Y. Liu, J.F. Kang. "RRAM-based Convolutional Neural Networks for High Accuracy Pattern Recognition Tasks," VLSI-SNW, Oral Presentation, pp.145.

[8] **Zhen Dong**, Zheng Zhou, Xinxin Wang, Zefan Li, Peng Huang, Lifeng Liu, Xiaoyan Liu, Jinfeng Kang. "Convolutional Neural Networks for Image Recognition and Online Learning Based on RRAM Devices," IEEE Transactions on Electron Devices 2018, p.793-801.

[9] Jinfeng Kang**, Zhen Dong***, Peng Huang, Renze Han, Lifeng Liu, Xiaoyan Liu. China's patent about 3D RRAM.

[10] Huang P., Li Z., **Zhen Dong**, Han R., Zhou Z., Zhu D., Liu L., Liu X. and Kang J. "Binary Resistive Switching Device Based Electronic Synapse with Spike-Rate-Dependent-Plasticity for Online Learning," ACS Applied Electronic Materials 2019, pp. 845-853.

[11] Xinxin Wang, Peng Huang, **Zhen Dong**, Zheng Zhou, Yuning Jiang, Runze Han, Lifeng Liu, Xiaoyan Liu, Jinfeng Kang. "A Novel RRAM-based Adaptive-Threshold LIF Neuron Circuit for High Recognition Accuracy," published by International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA) 2018, pp. 1-2.

[12] Runze Han, Peng Huang, Yachen Xiang, Chen Liu, **Zhen Dong**, et al. "A Novel Convolutional Computing Paradigm Based on NOR Flash Array with High Computing Speed and Energy Efficiency," published by IEEE Transactions on Circuits and Systems 2019, p.1-12.

[13] Zheng Zhou, Chen Liu, Wensheng Shen, **Zhen Dong**, Zhe Chen, Peng Huang, Lifeng Liu, Xiaoyan Liu, Jinfeng Kang. "The Characteristics of Binary Spike-Time-Dependent Plasticity in HfO2-Based RRAM and Applications for Pattern Recognition," published by Nanoscale Research Letters, 12(1), p.244.

[14] P. Huang, D. B. Zhu, C. Liu, Z. Zhou, **Zhen Dong**, H. Jiang, W. S. Shen, L. F. Liu, X. Y. Liu, and J. F. Kang. "RTN based Oxygen Vacancy Probing Method for Ox-RRAM Reliability Characterization and Its Application in Tail Bits," published by International Electron Devices Meeting (IEDM) 2017, pp. 21-4.

SKILLS

- Programming languages: Python, C/C++, Matlab, Verilog.
- Platforms: PyTorch, Tensorflow(& Keras), TensorflowLite, Caffe.
- Hardware design tools: Cadence, Xilinx Vivado & ISE, HSpice, Modelsim.