YAOHUI CAI

(+1) 617-335-8109 ⋄ caiyaohui@pku.edu.cn
https://people.eecs.berkeley.edu/~yaohuic/

EDUCATION

Peking University, Beijing, China *Sept. 2016 - Sept. 2020 (expected)*
Yuanpei College
B.S. in Data Science

University of California at Berkeley, Berkeley, CA, USA *May 2019 - Jan. 2020*
Department of Electrical Engineering and Computer Science
Vising student researcher advised by **Kurt Keutzer** and **Michael W. Mahoney**

Massachusetts Institute of Technology, Cambridge, MA, USA *Feb. 2019 - May 2019*
Department of Electrical Engineering and Computer Science
Exchange student
Overall GPA: **5.0/5.0**

PUBLICATION

[1] *ZeroQ: A Novel Zero Shot Quantization Framework.*
Yaohui Cai*, Z. Yao*, Z. Dong*, A. Gholami, M. W. Mahoney, and K. Keutzer.
arXiv preprint. CVPR2020, under review.

[2] *HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks.*
Z. Dong, Z. Yao, **Yaohui Cai***, D. Arfeen*, A. Gholami, M. W. Mahoney, and K. Keutzer.
arXiv preprint. CVPR2020, under review.

[3] *3D Macromolecule Localization in Cryo-Electron Tomography with Deep Reinforcement Learning.*
Yaohui Cai*, X. Zeng*, Y. Zeng, W. Liu, J. Jin, Z. Freyberg, G. Wang, and M. Xu.
Deep Reinforcement Learning Workshop, NeurIPS 2019

[4] *Hessian-Aware Trace-Weighted Quantization.*
Z. Dong, Z. Yao, A. Gholami, **Yaohui Cai**, D. Arfeen, M. W. Mahoney, and K. Keutzer,
Beyond First Order Methods in Machine Learning Workshop (**Oral**), NeurIPS 2019

[5] *Algorithm-hardware Co-design for Deformable Convolution.*
Q. Huang*, D. Wang*, Y. Gao, **Yaohui Cai**, Z. Dong, B. Wu, K. Keutzer, and J. Wawrzynek.
Energy Efficient Machine Learning and Cognitive Computing Workshop (**Oral**), NeurIPS 2019

RESEARCH EXPERIENCE

Efficient Deep Learning with Quantization *May 2019 - Jan. 2020*
Visiting Student Researcher *University of California, Berkeley*

- Advisors: **Kurt Keutzer**, Professor of Department of Electronic Engineering and Computer Science.
  Michael W. Mahoney, Professor of Department of Statistics.
- Proposed a novel method to efficiently compress networks to low-bit without any data or re-training.
- Compress an object detection network, RetinaNet, to ultra-low precision guided by second-order information outperforming state-of-the-art performance.
- Proposed using a randomized Hutchison algorithm to measure the sensitivity of deep neural networks efficiently.
- Hardware-algorithm co-designed an object detection network implemented on FPGA.
- Runner Up at *CVPR2019, Visual Wake Word Challenge* with a specially designed compact model.

Reinforcement Learning on 3D Object Localization *Feb. 2019 - May. 2019*
Remote Research Intern *Carnegie Mellon University*

- Advisor: **Min Xu**, Research Assistant Professor of Computational Biology Department.
- Proposed a reinforcement learning based method to localize the macromolecules in 3D Cryo-ET images automatically and efficiently.

Knowledge Graph Modeling *Jan. 2018 - Dec. 2018*
Research Intern *Peking University*

- Advisor: **Bin Cui**, Professor of School of Electrical Engineering and Computer Science.
- Proposed a novel method in modeling knowledge graph with Graph Neural Network to improve accuracy and efficiency.