Zhen Dong

https://dong-zhen.com | Google Scholar | Dissertation Poster

RESEARCH INTERESTS

Efficient AI: Efficient inference and training for generative models (Vision & NLP)
LLM compression, AI systems, serving and acceleration
Function-calling LLM agents and multi-agent systems
Hardware-software co-design and AI for science
Efficient evaluation and alignment of foundation models

EDUCATION

Postdoc in CS at UC Berkeley, Advisor: Prof. Kurt Keutzer *Oct. 2022 – Oct. 2023*

Ph.D. in CS at UC Berkeley, Advisor: Prof. Kurt Keutzer *Aug. 2018 – Oct. 2022*

PhD & Postdoc graduated from Berkeley AI Research (BAIR). Created startup Nexusflow.AI with Berkeley advisors and working as a founding member after graduation, with continued research collaboration with Berkeley to date.

B.S. in EECS at Peking University (Rank 1st/327) *Sept. 2014 – July 2018*

AWARDS

Winner of 2018-2020 Berkeley Fellowship
Winner of PhD Forum (2nd Place among all candidates) at DAC 2024
Doctoral Consortium at CVPR 2024
Best Paper Nomination at Practical DL Workshop at AAAI 2023
1st Place on EMCC 2020 Competition on Classification Track and Object Detection Track
2nd Place on Visual Wake Word Competition at CVPR 2019
AWS Research Credits Award 2021 and Google Cloud Research Credits Award 2022
1st Place Research Funding Proposal at Berkeley Deep Drive (BDD) 2019
Winner of SenseTime Scholarship in 2018
1st Prize in the National Olympiad in Physics (China), 1st Prize in National Physics Competition for College Students
Winner of Tang Lixin Scholarship (Highest honor for undergraduate academic and research excellence in China)
Winner of Tang Lixin 1st Prize Scholarship, Winner of Fang Zheng Scholarship
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

Founding Member, **Nexusflow.AI** *Jun. 2023 - present*

Research on Large Language Models (LLMs) with Function Calls: NexusRaven, NaxusRaven-V2, Athene-V2

- Propose novel methods to construct structured data and conduct instructional finetuning for NexusRaven models.
- Publicize benchmarks and a leaderboard that can measure the function calling abilities of LLMs.
- NexusRaven-V2 is around 7% better than GPT-4 on function calling, with only 13B parameters.
- NexusRaven-V2-13B gained 461 likes and **184k total downloads** on Huggingface. It ranked Top-5 on Huggingface Trending when released.

Ph.D./Postdoc, **Berkeley AI Research** (BAIR), UC Berkeley *Dec. 2018 – Jun. 2023*

Advisor: Prof. Kurt Keutzer

Research on Hessian-aware Quantization, Pruning & Distillation: HAWQ (ICCV'19), HAWQ-V2 (NeurIPS'20), ZeroQ (CVPR'20), HAP (WACV'22), Quantization Review (BLPCV'22), QD-BEV (ICCV'23), NoisyQuant (CVPR'23)

- Propose a Hessian-based method to decide mixed-precision configuration and block-wise fine-tuning order.
- Prove theorem to use the trace of Hessian as sensitivity metric and conduct fast Pareto frontier optimization.
- Generalize to segmentation, 2D/3D object detection tasks and achieve state-of-the-art results.
- Conduct fast end-to-end quantization without fine-tuning and without using any training/test data.
- This line of work has obtained **2762 citations** to date.

Research on HW-SW Co-design:

HAWQ-V3 (ICML'21), CoDeNet (FPGA'21), HAO (FCCM'21), CSQ (DAC'23), EPIM (DAC'24)

Achieve hardware-aware quantization and utilize 4-bit Tensor Cores for inference acceleration.

Implement 4-bit kernels and mixed-precision support on TVM, achieve 7.4x compression ratio and 5.4x speedup compared to fp32.

Propose efficient deformable operations on embedded FPGAs and design new FPGA-core with ultra-low precision arithmetic.

HW-SW joint architecture search and efficient implementation of mixed-precision NNs on CPU/GPU/FPGAs/PIM (Processing-in-Memory).

This line of work has obtained **420 citations** to date.

Research on Efficient LLMs and Diffusion Models:

Q-BERT (AAAI'20), Q-Diffusion (ICCV'23), SqueezeLLM (ICML'24), PB-LLM (ICLR'24)

Propose sensitivity-based non-uniform quantization and dense-and-sparse decomposition for efficient handling of outliers.

Pioneer the usage of Hessian information to guide LLM quantization in both post-training quantization (PTQ) and quantization-aware training.

Implement 3/4-bit CUDA kernels and achieve 4.6x compression ratio compared to fp16 and 2.4x speedup when deployed on an A6000 GPU.

Propose timestep-aware calibration and split shortcut quantization to achieve 4-bit diffusion models at the first time.

This line of work has obtained **880 citations** to date.

Research on Multi-agent Systems: MAgIC (EMNLP'24)

Pioneer the integration of probabilistic graphical modeling (PGM) to enhance the cognitive abilities of LLMs and obtain better interpretability.

Present a framework to evaluate LLM-powered multi-agent systems by employing social deduction games alongside game theory scenarios.

Research on Image & Video Generative Models:

PromptCoT (CVPR'24), ViewControl (IJCAI'24), D-Edit (AAAI'25), Meissonic, Magic-Me, VEditBench, KSort Arena

Propose new methods to achieve better control ability of generative diffusion models.

Develop novel efficient Arena algorithms for human-in-the-loop evaluation and alignment, and collect benchmarks for less costly auto-eval.

Present Meissonic-1B that elevates masked image modeling (MIM) text-to-image models to a level comparable to SDXL.

Research on AI for Science:

FastML (Frontiers in Big Data'22), High-momentum Particle Trigger Decisions (TRETS'24)

Review AI inference acceleration techniques and how they help dark matter search, morphology characterization, synthesis dynamics, etc.

Implement efficient AI on ASICs and FPGAs to reduce time cost and enable particle trigger decisions at CERN Large Hadron Collider (LHC).

Research Intern, **Bytedance AI** *Jan. 2023 – Apr. 2023*

Research Intern, **Nvidia AI** *Jun. 2021 – Sept. 2021*

Research Intern, **Facebook AI** *Jun. 2020 – Aug. 2020*

Research Intern, **SenseTime AI** *Apr. 2018 – Aug. 2018*

Undergraduate Visiting Researcher (UGVR), Electrical Engineering, **Stanford University** *Jun. 2017 – Sept. 2017*

Advisor: Prof. H.-S. Philip Wong

Research Assistant, Electrical Engineering and Computer Sciences, **Peking University** *Dec. 2016 – Jun. 2018*

Advisor: Prof. Jinfeng Kang

INDUSTRIAL COLLABORATIONS

Intel, Amazon, Alibaba, Nvidia, Panasonic, Bytedance, Google, Meta, Apple, AMD, Nexusflow, Samsung, Tesla

OPENSOURCE

In total received: Github 4k Stars, Huggingface 1.2k Likes, over 200k Huggingface Model Downloads

HAWQ, ZeroQ, CoDeNet, BitPack, AwesomeQuantizationPapers, LOVEU-TGVE, HAP, Q-Diffusion, SqueezeLLM, NexusRaven [huggingface], NexusRaven-V2 [huggingface][demo][leaderboard], Magic-Me [website][demo], Ksort Arena [huggngface], Meissonic [huggingface], D-Edit [huggingface], AtheneV2-Agent [huggingface], [AtheneV2-Chat-72B][huggingface], NexusBench

FIRST- & CORRESPONDING-AUTHOR PUBLICATIONS

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

[2] **Zhen Dong**, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, Kurt Keutzer. "HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks," **[NeurIPS 2020]**.

[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," Spotlight, **[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] Zhewei Yao*, **Zhen Dong***, Zhangcheng Zheng*, Amir Gholami*, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael W. Mahoney, Kurt Keutzer. "HAWQV3: Dyadic Neural Network Quantization," **[ICML 2021]**.

[6] **Zhen Dong***, Yizhao Gao*, Qijing Huang, John Wawrzynek, Hayden K.H. So, Kurt Keutzer. "HAO: Hardware-aware neural Architecture Optimization for Efficient Inference," Oral, **[FCCM 2021]**.

[7] **Zhen Dong***, D. Wang*, Q. Huang*, Y. Gao, Y. Cai, B. Wu, T. Li, K. Keutzer and J. Wawrzynek. "CoDeNet: Algorithm-hardware Co-design for Deformable Convolution," **[FPGA 2021]** Oral Presentation.

[8] Amir Gholami*, Sehoon Kim*, **Zhen Dong*,** Zhewei Yao*, Michael W. Mahoney, Kurt Keutzer. "A Survey of Quantization Methods for Efficient Neural Network Inference", **[BLPCV 2021]** (Book of Low-Power Computer Vision).

[9] Chenyu Wang*, **Zhen Dong*✉,** Daquan Zhou*✉, Zhenhua Zhu, Yu Wang, Jiashi Feng, Kurt Keutzer. "EPIM: Efficient Processing-In-Memory Accelerators based on Epitome," **[DAC 2024]**.

[10] Yifan Zhang*, **Zhen Dong***, Huanrui Yang, Ming Lu, Cheng-Ching Tseng, Yandong Guo, Kurt Keutzer, Li Du, Shanghang Zhang. "QD-BEV: Quantization-aware View-guided Distillation for Multi-view 3D Object Detection," **[ICCV 2023]**.

[11] Yuzhang, Shang, Zhihang Yuan Qiang Wu, **Zhen Dong✉**. "PB-LLM: Partially Binarized Large Language Models," **[ICLR 2024]**.

[12] Lin Xu, Zhiyuan Hu, Daquan Zhou✉, Hongyu Ren, **Zhen Dong✉,** Kurt Keutzer, See-Kiong Ng, Jiashi Feng. "MAgIC: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration," **[EMNLP 2024]**.

[13] Shixing Yu*, Zhewei Yao*, Amir Gholami*, **Zhen Dong***, Michael W. Mahoney, and Kurt Keutzer. "Hessian-Aware Pruning and Optimal Neural Implant," Oral, **[WACV 2022]**.

[14] **Zhen Dong,** Yaohui Cai, Amir Gholami, Tianjun Zhang, Kurt Keutzer. "Ultra-low Bit Quantization for Visual Wake Word Challenge", 2nd Place at VWW Competition, **[CVPR 2019]**.

[15] **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 **[TED 2018]**.

[16] **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 2018]**, Oral Presentation.

[17] Zhikai Li, Xuewen Liu, Dongrong Fu, Jianquan Li, Qingyi Gu, Kurt Keutzer, **Zhen Dong✉**. "K-Sort Arena: Efficient and reliable benchmarking for generative models via K-wise human preferences," arXiv 2024.

[18] Ze Ma, Daquan Zhou✉, Chun-Hsiao Yeh, Xue-She Wang, Xiuyu Li, Huanrui Yang, **Zhen Dong✉,** Kurt Keutzer, Jiashi Feng. "Magic-Me: Identity-Specific Video Customized Diffusion," arXiv 2024.

[19] Rui Ma, Qiang Zhou, Yizhu Jin, Daquan Zhou, Bangjun Xiao, Xiuyu Li, Yi Qu, Aishani Singh, Kurt Keutzer, Jingtong Hu, Xiaodong Xie, **Zhen Dong✉,** Shanghang Zhang✉, Shiji Zhou. "A Dataset and Benchmark for Copyright Protection from Text-to-Image Diffusion Models," arXiv 2024.

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

SELECTED PUBLICATIONS (chronological order)

[21] Jinbin Bai, **Zhen Dong,** Aosong Feng, Xiao Zhang, Tian Ye, Kaicheng Zhou, Mike Zheng Shou. "Integrating View Conditions for Image Synthesis," **[IJCAI 2024]**.

[22] Rongyu Zhang, Yulin Luo, Huanrui Yang, **Zhen Dong**, … & Shanghang Zhang. "Efficient Deweahter Mixture-of-Experts with Uncertainty-Aware Feature-wise Linear Modulation," **[AAAI 2024]**.

[23] Junyi Yao, Yijiang Liu, **Zhen Dong,** Mingfei Guo, Jiashi Feng, Kurt Keutzer, Li Du, Daquan Zhou, Shanghang Zhang. "PromptCoT: Align prompt distribution via adapted chain of thought," **[CVPR 2024]**.

[24] Sehoon Kim, Coleman Hooper, Amir Gholami, **Zhen Dong,** Xiuyu Li, Sheng Shen, Michael W. Mahoney, Kurt Keutzer. "SqueezeLLM: Dense-and-Sparse Quantization," **[ICML 2024]**.

[25] Anthony Chen, Huanrui Yang, Yulu Gan, Denis A Gudovskiy, **Zhen Dong,** Haofan Wang, Tomoyuki Okuno, Yohei Nakata, Shanghang Zhang, Kurt Keutzer. "Split-Ensemble: Efficient OOD-aware ensemble via task and model splitting," **[ICML 2024]**.

[26] Javier Campos, **Zhen Dong,** Javier Duarte, Amir Gholami, Michael Mahoney, Jovan Mitrevski, Nhan Tran. "End-to-end codesign of Hessian-aware quantized neural networks for FPGAs and ASICs," ACM Transactions on Reconfigurable Technology and Systems **[TRETS 2024]**.

[27] Venkat Srinivasan, **Zhen Dong,** Banghua Zhu, Brian Yu, Damon Mosk-Aoyama, Kurt Keutzer, Jiantao Jiao, Jian Zhang. "NexusRaven: A Commercially-Permissive Language Model for Function Calling," **[FMDM@NeurIPS 2024]**.

[28] Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, **Zhen Dong,** Lei Zhu, Shuicheng Yan. "Meissonic: Revitalizing

Masked Generative Transformers for Efficient High-Resolution Text-to-Image Synthesis," arXiv 2024.

[29] Zhihang Yuan, Yuzhang Shang, Yang Zhou, **Zhen Dong,** Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, Yan Yan, Beidi Chen, Guangyu Sun, Kurt Keutzer. "LLM Inference Unveiled: Survey and Roofline Model Insights," arXiv 2024.

[30] Javier Campos, **Zhen Dong,** Javier Duarte, Amir Gholami, Michael Mahoney, Jovan Mitrevski and Nhan Tran. "End-to-end codesign of Hessian-aware quantized neural networks for FPGAs and ASICs," OSCAR Workshop at **[ISCA 2023]**.

[31] Lirui Xiao, Huanrui Yang, **Zhen Dong,** Kurt Keutzer, Li Du, Shanghang Zhang. "CSQ: Growing Mixed-Precision Quantization Scheme with Bi-level Continuous Sparsification," **[DAC 2023]**.

[32] Yijiang Liu, Huanrui Yang, **Zhen Dong,** Kurt Keutzer, Li Du, Shanghang Zhang. "NoisyQuant: Noisy Bias-Enhanced Post-Training Activation Quantization for Vision Transformers," **[CVPR 2023]**.

[33] Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, **Zhen Dong,** Daniel Kang, Shanghang Zhang, Kurt Keutzer. "Q-Diffusion: Quantizing Diffusion Models," **[ICCV 2023]**.

[34] Mingfei Guo, **Zhen Dong,** Kurt Keutzer. "SANA: Sensitivity-Aware Neural Architecture Adaptation for Uniform Quantization," Applied Sciences, 2023.

[35] Tian Li, Xiang Chen, **Zhen Dong,** Weijiang Yu, Yijun Yan, Shanghang Zhang, Kurt Keutzer. "Domain-Adaptive Text Classification with Structured Knowledge from Unlabeled Data", Long Oral, **[IJCAI 2022]**.

[36] Allison McCarn Deiana, Nhan Tran, … **Zhen Dong,** … Olivia Weng. "Applications and Techniques for Fast Machine Learning in Science", **[Frontiers in Big Data 2022]**.

[37] Tian Li, Xiang Chen, Shanghang Zhang, **Zhen Dong**, Kurt Keutzer. "Cross-Domain Sentiment Classification with Contrastive Learning and Mutual Information Maximization," **[ICASSP 2021]**.

[38] Peng Huang, Zefan Li, **Zhen Dong**, Runze Han, Zheng Zhou, Dongbin Zhu, Lifeng Liu, Xiaoyan Liu, and Jinfeng Kang. "Binary Resistive Switching Device Based Electronic Synapse with Spike-Rate-Dependent-Plasticity for Online Learning," ACS **[Applied Electronic Materials 2019]**.

[39] 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 **[TCAS 2019]**.

[40] 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]**.

[41] 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," Nanoscale Research Letters **[NRL 2018]**.

[42] 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]**.

ACADEMIC SERVICE

Program Committee Member: NeurIPS, ICML, CVPR, ICCV, EMNLP, ICLR, AAAI, ECCV, IJCAI, WACV, KDD, MLSys, TinyML, ECV, BLPCV

Reviewer for TPAMI (Transactions on Pattern Analysis and Machine Intelligence), TMLR (Transactions of Machine Learning Research), JMLR (Journal of Machine Learning Research), TNNLS (IEEE Transactions on Neural Networks and Learning Systems), IEEE Micro, TED (IEEE Transactions on Electron Devices), PR (Pattern Recognition), TCSVT (IEEE Transactions on Circuits and Systems for Video Technology), OJCAS (IEEE Open Journal of Circuits and Systems), JCST (Journal of Computer Science and Technology) and Fundamental Research (Elsevier)

TALKS & ORGANIZED WORKSHOPS & MEDIA

[1] Meissonic gets recommended by AI era (新智元) and 36Kr, Link to Post.

[2] KSort Arena gets recommended by Qingke Lab, Link to Post.

[3] I presented "Efficient Deep Learning via Quantization and Co-Design" at CVPR 2024 Doctoral Consortium and DAC 2024 PhD Forum.

[4] I co-organized the LOVEU (LOng-form VidEo Understanding) workshop at CVPR 2024.

[5] Q-Diffusion is featured in the newest TensorRT post.

[6] I co-organized the 3rd Workshop on Practical Deep Learning: Towards Efficient and Reliable LLMs at IEEE Conference on Artificial Intelligence (IEEE CAI) 2024.

[7] NexusRaven-V2-13B is presented at NeurIPS 2023 EXPO.

[8] NexusRaven and NexusRaven-V2 are recommended by: Deci AI Top 10 Under-13B LLMs, Huggingface's Post, Together AI's Post, etc.

[9] Invited Talk "Efficient Inference and Training of Large Neural Network Models" at Intel oneAPI DevSummit for AI and HPC 2023.

[10] Invited Talk "Hardware-Aware Efficient Deep Learning" at Peking University Institute of Artificial Intelligence (PKU-IAI), on June 11, 2023.

[11] I co-organized the LOVEU (LOng-form VidEo Understanding) workshop at CVPR 2023, Link to Zhihu.

[12] Invited to host the Practical DL Workshop at AAAI 2023 in Washington DC.

[13] Invited Talk "Efficient Deep Learning via Quantization and HW-SW Co-Design" at Hardware and Algorithms for Learning On-a-chip Workshop (HALO) at ICCAD 2022.

[14] My dissertation on "Hardware-aware Efficient Deep Learning" was defended on June 29, 2022.

[15] "Efficient Neural Networks through Systematic Quantization and Co-Design", virtually at Matchlab (Imperial College London), [slides].

[16] CoDeNet and HAO are presented at ML@B Seminar (Machine Learning at Berkeley).

[17] "Hessian-Aware Pruning and Optimal Neural Implant", WACV 2022, Hawaii, US, [slides].

[18] Berkeley AI Research (BAIR)/ Berkeley Deep Drive (BDD) Workshop 2021, Berkeley, US.

[19] The book that I contributed to, "Low-Power Computer Vision: Improve the Efficiency of Artificial Intelligence", is online for ordering.

[20] "HAO: Hardware-aware neural Architecture Optimization for Efficient Inference", FCCM 2021 (online).

[21] "HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks", NeurIPS 2020.

[22] HAWQ-V2 gets recommended by JiangMen (将门) AI media (in Chinese), Link to ZhiHu.

[23] "Systematic Neural Network Quantization", NVIDIA GTC 2021.

[24] "Efficient Neural Networks through Systematic Quantization", BAIR/CPAR/BDD Seminar 2020, [slides].

[25] "HAWQ-V3: Dyadic Neural Network Quantization" is presented at TVM Conference 2020.

[26] "ZeroQ: A novel Zero-Shot Quantization Framework", Real-Time Intelligent Secure Explainable Systems (RISELab) Retreat 2020, Lake Tahoe (online), US, [slides].

[27] Berkeley AI Research (BAIR)/ Berkeley Deep Drive (BDD) Workshop 2020, Santa Rosa, US.

[28] "Q-BERT: Hessian Based Quantization of BERT", AAAI 2020, New York, US, [slides].

[29] Q-BERT gets recommended by Synced (机器之心) AI media (in Chinese), Link to WeChat.

[30] Q-BERT gets recommended by AI.Science (Aggregate Intellect), Link to YouTube.

[31] "Hessian-Aware trace-Weighted Quantization", Beyond First-Order Methods in ML Workshop at NeurIPS 2019, Vancouver, Canada.

[32] Real-Time Intelligent Secure Explainable Systems (RISELab) Retreat 2019, Monterey, US.

[33] Berkeley AI Research (BAIR)/ Berkeley Deep Drive (BDD) Workshop 2019, Berkeley, US.

[34] Visual Wake Word Challenge, LPIRC Workshop at CVPR 2019, Long Beach, US, [slides], [link].

[35] "RRAM Based Convolutional Neural Networks for High Accuracy Pattern Recognition and Online Learning Tasks", VLSI-SNW 2017, Kyoto, Japan, [slides].

TEACHING EXPERIENCE

Online Course of CS267 Parallel Computing on Moodle XSEDE: Course Coordinator

Applications of Parallel Computers, Berkeley CS 267: Head Graduate Student Instructor

Optimization Analytics, Berkeley INDENG 240: Graduate Student Instructor

Mathematical Programming, Berkeley INDENG 262A: Graduate Student Instructor

BAIR Mentoring Program for Underrepresented Undergraduates

GRANT & FUNDING WRITING

I have significantly contributed to the writing of the following grants and fundings:

[1] Berkeley Deep Drive (BDD) 2019 Funding Proposal: Large Scale Second-Oder Stochastic Training of Neural Networks Through K-FAC

[2] Google Cloud Research Grant 2019: Hessian-aware Mixed-Precision Quantization with Distillation

[3] Wave Computing Research Grant 2019: Model Compression of RoBERTa on NLP Tasks

[4] Berkeley Deep Drive (BDD) 2020 Funding Proposal: Efficient Neural Networks Through Systematic Quantization

[5] AWS Research Grant 2020: Hardware-aware Quantization with End-to-end Inference Acceleration

[6] Google Cloud Research Grant 2020: Hardware Software Co-Design for NLP and Recommendation Systems

[7] Facebook Research Grant 2020: A Study of Communication Avoiding Algorithms for Training Large Scale Recommendation Systems

[8] Berkeley Deep Drive (BDD) 2021 Funding Proposal: Real-time and Accurate Object Detection through Quantization of Transformer- and MLP-based Computer Vision Models

[9] Alibaba Berkeley Commons Grant 2021: TASC: Topology-Aware Structured Communications for Efficient Deep Neural Network Training

[10] Panasonic Research Grant 2022: Sensitivity-aware DETR Quantization

[11] Alibaba Berkeley Commons Grant 2022: DQRM: Deep Quantized Recommendation Models

[12] Berkeley Deep Drive (BDD) 2022 Funding Proposal: Efficient Transformer Inference and Training for Fast Unsupervised Learning Through Attention-Aware Pruning

[13] Intel Research Grant 2023: Efficient Distributed Training of Large-Scale Neural Networks

[14] Panasonic Research Grant 2023: Controllable AI with LLM/VLM

[15] Berkeley Deep Drive (BDD) 2023 Funding Proposal: Quantization on Vision Models for Real-time and Accurate Inference in ADAS/AV

[16] Intel Research Grant 2024: Efficient Deep Learning on Intel Processors and Networks

NOTABLE MENTORING EXPERIENCE

I have had the privilege of mentoring many talented students over the years, listed below in chronological order.

1. Yaohui Cai (Undergrad at PKU, now PhD at Cornell)
   Yaohui was a visiting research intern at Berkeley. Then he came back to work with us for an extra half year before starting his PhD at Cornell. I have closely mentored Yaohui on HAWQ-V2, ZeroQ and CoDeNet for 2 years.
2. Daiyaan Arfeen (Undergrad at Berkeley, now PhD at CMU)
   I closely mentored Daiyaan on HAWQ-V2 for 1 year.
3. Sheng Shen (MEng at Berkeley, then PhD at Berkeley, now at Meta Llama Team)
   I closely mentored Sheng when he was an MEng student, we worked together on Q-BERT for 1 year.
4. Zhangcheng (Zach) Zheng (MEng at Berkeley, now at AWS)
   I mentored Zach for 1 year on HAWQ-V3 and on his MEng thesis, Zach is currently working with the AWS group who previously collaborated with us on HAWQ-V3.
5. Eric Tan (MEng at Berkeley, now at Google)
   I mentored Eric for 1 year on HAWQ-V3 and his MEng thesis.
6. Lutfi Eren Erdogan (Undergrad at Berkeley, now at Narada.ai)
   I closely mentored Eren for 1 year on the Hot Embedding project.
7. Yizhao Gao (Exchange student at Berkeley, now PhD at HKU)
   In addition to his one year at Berkeley, Yizhao came back to work with us before his PhD at HKU. I closely mentored Yizhao on CoDeNet and HAO for 1.5 years.
8. Jannealle Brambila (Undergrad at Berkeley)
   Jannealle is a representative of my mentees at Berkeley underrepresented student mentoring program. I mentored her for half a year during which I introduced her to different AI directions, and industry/academia career development.
9. Yang Zhou (Visiting student at Berkeley, now PhD at CMU)
   I closely mentored Yang for 1.5 years, for projects DQRM, LLM Inference Survey, etc.
10. Aishani Singh (The Harker School)
    Aishani is a talented high schooler interested in AI. I mentored her for 1.5 years starting from the basics of AI to advanced research topics. She worked with us on organizing the LOVEU workshops at CVPR'23 and CVPR'24. I also mentored her on our recent VEditBench paper.
11. Yijiang Liu (PhD at NJU)
    I mentored Yijiang for 1.5 years through collaboration with his advisor Shanghang (Assistant Professor at PKU), for the projects Q-Diffusion, NoisyQuant and PromptCoT.
12. Vijay Anand Raghava Kanakagiri (Visiting student at Berkeley, now at Amazon)
    I mentored Vijay for 1 year on the Hot Embedding project and the KSort Arena project.
13. Xiuyu Li (PhD at Berkeley) I mentored Xiuyu when he was a first-year PhD in our group, on the project Q-Diffusion.
14. Mingfei Guo (MS at Stanford, now at Nvidia) I mentored Mingfei for 1 year on SANA and TASC.
15. Chenyu Wang (Undergrad at THU, now PhD at Princeton) I mentored Chenyu for 1 year on EPIM.

SKILLS

Programming languages: Python, C/C++, SQL, Matlab, Verilog

Platforms: PyTorch, JAX, Tensorflow(&Keras), MXNet, TensorRT, OpenMP/MPI/UPC++/CUDA, torch-CCL

Tools: Cadence, Gurobi, Xilinx Vivado & ISE, HSpice, Gradio, OriginLab, Modelsim