Peiyan (Peggie) Dong

+1-6179597749 | dong.pe@northeastern.edu | Google Scholar | LinkedIn | GitHub | Boston, USA

EDUCATION

Ph.D. Candidate in Computer Engineering *Spring 2020 – May 2024 (expected)*
Northeastern University, Boston, MA, USA GPA: 4.0/4.0
Research Advisor: Yanzhi Wang

M.S. in Operation Research *January 2017 – December 2019*
Northeastern University, Boston, MA, USA GPA: 3.9/4.0

Bachelor's in Information Engineering *August 2012 – June 2016*
Beijing Institute of Technology, Beijing, China GPA: 3.8/4.0

Research Focus: *General AI System, Hardware and Software Co-design for DNN Architecture, Inference-Efficient/Energy-Efficient Artificial Intelligence Systems, Efficient Deep Learning on Superconducting Devices, Placement and Routing on Superconducting Devices, Emerging Deep Learning Systems.*

PUBLICATIONS

Summary: There are 16 first/co-first author publications ranging from:

(I) EDA, solid-state circuit, and system conferences such as DAC (6), ICCAD (2), ISSCC (1), ASP-DAC (2), RTAS (1), MLSys (1).

(II) Architecture and computer system conferences such as MICRO (1), HPCA (1), ICS (1), HPCA (1 under review).

(III) Machine learning algorithm conferences such as NeurIPS (2), ICML (1), CVPR (1), AAAI (2), ECCV (2), IJCAI (1), AAAI (1 under review).

(IV) Journal publications including TCAD, Advanced Intelligent Systems, TCASI, TECS, TPAMI.

Selected Conference Publications (*** Equal Contribution**)

1. [**23'HPCA**] Peiyan Dong, Mengshu Sun, Alec Lu, Yanyue Xie, Zhenglun Kong, Xin Meng, Xue Lin, Zhenman Fang, Yanzhi Wang, "**HeatViT: Hardware-Efficient Adaptive Token Pruning for Vision Transformers**", to appear in the 2023 IEEE International Symposium on High Performance Computer Architecture.
2. [**23'ASP-DAC**] Peiyan Dong, Changdi Yang, Yi Sheng, Yanyu Li, Lei Yang, Xue Lin, Yanzhi Wang, "**FF-Medical: Fast and Fair Medical AI on the Edge through Hardware-oriented Search for Hybrid Vision Models**", in the Proceedings of the 28th Asia and South Pacific Design Automation Conference.
3. [**22'DAC**] Peiyan Dong, Hongjia Li, Yanyue Xie, Olivia Chen, Mengshu Sun, Nobuyuki Yoshikawa and Yanzhi Wang, "**TAAS: A Timing-Aware Analytical Strategy for AQFP-Capable Placement Automation**", in Proceedings of the 59th Annual Design Automation Conference.
4. [**20'DAC**] Peiyan Dong, Siyue Wang, Wei Niu, Chengming Zhang, Sheng Lin, Zhengang Li, Yifan Gong, Bin Ren, Xue Lin, Dingwen Tao, "**Rtmobile: Beyond real-time mobile acceleration of rnns for speech recognition**", in Proceedings of the 57th Annual Design Automation Conference.
5. [**23'NeurIPS**] Peiyan Dong, Zhenglun Kong, Xin Meng, Pinrui Yu, Yanyue Xie, Yifan Gong, Geng Yuan, Fei Sun, Hao Tang, Yanzhi Wang, "**HotBEV: Hardware-oriented Transformer-based Multi-View 3D Detector for BEV Perception**", in Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS, 2023).
6. [**23'NeurIPS**] Peiyan Dong, Lei Lu, Chao Wu, Cheng Lyu, Geng Yuan, Hao Tang, Yanzhi Wang, "**PackQViT:**

Faster Sub-8-bit Vision Transformers via Full and Packed Quantization on the Mobile", in Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS, 2023).

7. [**23'ICML**] Peiyan Dong, Zhenglun Kong, Xin Meng, Peng Zhang, Hao Tang, Yanzhi Wang, Chih-Hsien Chou, "**SpeedDETR: Speed-aware Transformers for End-to-end Object Detection**", to appear in the 2023 Fortieth International Conference on Machine Learning.
8. [**22'ECCV**] Zhenglun Kong*, Peiyan Dong*, Xiaolong Ma, Xin Meng, Mengshu Sun, Wei Niu, Bin Ren, Minghai Qin, Hao Tang, Yanzhi Wang, "**SPViT: Enabling Faster Vision Transformers via Soft Token Pruning**", European Conference on Computer Vision 2022 (ECCV 2022).
9. [**20'ICS**] Runbin Shi*, Peiyan Dong*, Tong Geng, Yuhao Ding, Xiaolong Ma, Hayden K-H So, Martin Herbordt, Ang Li, Yanzhi Wang, "**CSB-RNN: a faster-than-realtime RNN acceleration framework with compressed structured blocks**", in the Proceeding of the International Conference on Supercomputing.
10. [**23'MICRO**] Zhengang Li, Geng Yuan, Tomoharu Yamauchi, Zabihi Masoud, Yanyue Xie, Peiyan Dong, Xulong Tang, Nobuyuki Yoshikawa, Devesh Tiwari, Yanzhi Wang, Olivia Chen, "**SupeRBNN: Randomized Binary Neural Network Using Adiabatic Superconductor Josephson Devices**", to appear in the Proceeding of 56th IEEE/ACM International Symposium on Microarchitecture.
11. [**23'ICCAD**] Changdi Yang*, Yi Sheng*, Peiyan Dong*, Zhenglun Kong, Yanyu Li, Pinrui Yu, Lei Yang, Xue Lin, "**Fast and Fair Medical AI on the Edge through Neural Architecture Search for Hybrid Vision Models**", in 2023 IEEE/ACM International Conference on Computer Aided Design.
12. [**23'DAC**] Zhengang Li*, Yanyue Xie*, Peiyan Dong*, Olivia Chen, Yanzhi Wang, "**Invited: Algorithm-Software-Hardware Co-Design for Deep Learning Acceleration**", in Proceedings of the 60th Annual Design Automation Conference.
13. [**23'DAC**] Changdi Yang*, Yi Sheng*, Peiyan Dong*, Zhenglun Kong, Yanyu Li, Pinrui Yu, Lei Yang, Xue Lin, "**Fast Fair Medical Applications? Hybrid Vision Models Achieve the Fairness on the Edge：Late Breaking Results**", in Proceedings of the 60th Annual Design Automation Conference (DAC).
14. [**22'ICCAD**] Zhirui Hu, Peiyan Dong, Zhepeng Wang, Youzuo Lin, Yanzhi Wang, Weiwen Jiang, "**Quantum Neural Network Compression**", 2022 IEEE/ACM International Conference on Computer Aided Design (ICCAD).
15. [**21'ASP-DAC**] Qin Li*, Peiyan Dong*, Zijie Yu, Changlu Liu, Fei Qiao, Yanzhi Wang, Huazhong Yang, "**Puncturing the memory wall: Joint optimization of network compression with approximate memory for ASR application**", in the Proceedings of the 26th Asia and South Pacific Design Automation Conference.
16. [**21'RTAS**] Geng Yuan*, Peiyan Dong*, Mengshu Sun, Wei Niu, Zhengang Li, Yuxuan Cai, Jun Liu, Weiwen Jiang, Xue Lin, Bin Ren, Xulong Tang, Yanzhi Wang, "**Work in Progress: Mobile or FPGA? A Comprehensive Evaluation on Energy Efficiency and a Unified Optimization Framework**", to appear in the Proceedings of RTAS 2021.
17. ［**23'AAAI**] Zhenglun Kong, Haoyu Ma, Geng Yuan, Mengshu Sun, Yanyue Xie, Peiyan Dong, Xuan Shen, Hao Tang, Minghai Qin, Tianlong Chen, Xiaolong Ma, Xiaohui Xie, Zhangyang Wang, Yanzhi Wang, "**Peeling the Onion: Hierarchical Reduction of Data Redundancy for Efficient Vision Transformer Training**", in the Proceedings of the AAAI Conference on Artificial Intelligence (Oral).
18. [**23'IJCAI**] Xuan Shen, Zhenglun Kong, Minghai Qin, Peiyan Dong, Geng Yuan, Xin Meng, Hao Tang, Xiaolong Ma, Yanzhi Wang, "**Data Level Lottery Ticket Hypothesis for Vision Transformers**", in Proceedings of the 32nd International Joint Conference on Artificial Intelligence.
19. [**22'ECCV**] Geng Yuan, Sung-En Chang, Qing Jin, Alec Lu, Yanyu Li, Yushu Wu, Zhenglun Kong, Yanyue Xie, Peiyan Dong, Minghai Qin, Xiaolong Ma, Xulong Tang, Zhenman Fang, Yanzhi Wang, "**You Already Have It: A Generator-Free Low-Precision DNN Training Framework Using Stochastic Rounding**", European Conference on Computer Vision 2022 (ECCV 2022).
20. [**21'MLSys**] Yanyu Li, Geng Yuan, Zhengang Li, Wei Niu, Pu Zhao, Peiyan Dong, Yuxuan Cai, Xuan Shen,

Zheng Zhan, Zhenglun Kong, Qing Jin, Bin Ren, Yanzhi Wang, Xue Lin, "**A Compiler-aware Framework of Network Pruning Search Achieving Beyond Real-Time Mobile Acceleration**", to appear in the Proceeding of Fourth Conference on Machine Learning and Systems.

21. [**20'AAAI**] Ao Ren, Tao Zhang, Yuhao Wang, Sheng Lin, Peiyan Dong, Yen-kuang Chen, Yuan Xie, Yanzhi Wang, "**DARB: A Density-Adaptive Regular-Block Pruning for Deep Neural Networks**", in the Proceedings of the AAAI Conference on Artificial Intelligence.
22. [Under Review] Peiyan Dong, Jinming Zhuang, Zhuoping Yang, Shixin Ji, Dongkuan Xu, Yanyu Li, Heng Huang, Jingtong Hu, Alex K. Jones, Yiyu Shi, Yanzhi Wang, Peipei Zhou, "**EQ-ViT: Algorithm-Hardware Co-Design for End-to-End Acceleration of Real-Time Vision Transformer Inference on Versal ACAP**", submitted to HPCA 2024.
23. [Under Review] Peiyan Dong, Lei Lu, Chao Wu, Changdi Yang, Wei Niu, Geng Yuan, Yanzhi Wang, "**LogicViT: Low-bit Quantization and Concatenation for Boosted Vision Transformer on the Edge**", submitted to HPCA 2024.
24. [Under Review] Pinrui Yu*, Peiyan Dong*, Pu Zhao, Zhenglun Kong, Xin Meng, Fei Sun, Hao Tang, Yanzhi Wang, Xue Lin, "**Q-TempFusion: Quantization-aware Temporal Multi-sensor Fusion on Bird's-Eye View Representation**", submitting to CVPR 2024.
25. [Under Review] Xuan Shen*, Peiyan Dong*, Lei Lu, Zhenglun Kong, Zhengang Li, Ming Lin, Chao Wu, Yanzhi Wang, "**Agile-Quant: Activation-Guided Quantization for Faster Inference of LLMs on the Edge**", submitted to AAAI 2023.
26. [Under Review] Zhengang Li, Alec Lu, Yanyue Xie, Zhenglun Kong, Mengshu Sun, Hao Tang, Peiyan Dong, Yanzhi Wang, Xue Lin and Zhenman Fang, "**Quasar-ViT: Hardware-Oriented Quantization-Aware Architecture Search for Vision Transformers**", submitted to AAAI 2023.
27. [Under Review] Yanyue Xie*, Peiyan Dong*, Geng Yuan, Zhengang Li, Chao Wu, Sung-En Chang, Xufeng Zhang, Olivia Chen, Nobuyuki Yoshikawa, Yanzhi Wang, "**SuperFlow: An RTL-to-GDS Design Automation Flow for AQFP Superconducting Devices**", submitting to DATE 2023.
28. [Under Review] Peiyan Dong, Zhirui Hu, Tianlong Chen, Zhangyang Wang, Weiwen Jiang, Yanzhi Wang, "**A Task-agnostic Quantum Transformer for Quantum Ground State Preparation**", submitting to DATE 2023.
29. [Under Review] Masoud Zabihi, Yanyue Xie, Zhengang Li, Peiyan Dong, Geng Yuan, Olivia Chen, Yanzhi Wang, "**A Life-Cycle Energy and Inventory Analysis of Adiabatic Quantum-Flux-Parametron Circuits**", submitting to DATE 2023.

Journal Publication (*** Equal Contribution**)

1. [**TCAD**] Peiyan Dong, Mengshu Sun, Yanyue Xie, Xue Lin, Zhenman Fang, Yanzhi Wang, "**HetaViT: Hardware-Efficient and Token-Aware Joint Compression with Pruning and Quantization for Vision Transformers**," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**Impact Factor 2.9**).
2. [**TPAMI**] Wei Niu, Zhengang Li, Xiaolong Ma, Peiyan Dong, Gang Zhou, Xuehai Qian, Xue Lin, Yanzhi Wang, Bin Ren, "**GRIM: A General, Real-Time Deep Learning Inference Framework for Mobile Devices based on Fine-Grained Structured Weight Sparsity**," in IEEE Transactions on Pattern Analysis and Machine Intelligence (**Impact Factor 23.6**).
3. [**TCASI**] Qin Li, Changlu Liu, Peiyan Dong, Yanming Zhang, Tong Li, Sheng Lin, Minda Yang, Fei Qiao, Yanzhi Wang, Li Luo, Huazhong Yang, "**NS-FDN: Near-Sensor Processing Architecture of Feature Configurable Distributed Network for Beyond-Real-Time Always-on Keyword Spotting**", accepted in the IEEE Transactions on Circuits and Systems I: Regular Papers (**Impact Factor 5.1**).
4. [**TECS**] Yuan Geng*, Peiyan Dong*, Mengshu Sun, Wei Niu, Zhengang Li, Yuxuan Cai, Yanyu Li et al.

“**Mobile or FPGA? A Comprehensive Evaluation on Energy Efficiency and a Unified Optimization Framework**”, ACM Transactions on Embedded Computing Systems (2022) (**Impact Factor 2.0**).

5. Runze Han, Peng Huang, Yachen Xiang, Hong Hu, Sheng Lin, Peiyan Dong, Wensheng Shen, Yanzhi Wang, Xiaoyan Liu, Jinfeng Kang, “**Floating Gate Transistor-Based Accurate Digital In-Memory Computing for Deep Neural Networks**”, Advanced Intelligent Systems (2022) (**Impact Factor 7.4**).

Workshop Publication (*** Equal Contribution**)

1. [**23’DAC**] Yanyue Xie*, Peiyan Dong*, Geng Yuan, Zhengang Li, Chao Wu, Sung-En Chang, Xufeng Zhang, Olivia Chen, Nobuyuki Yoshikawa, Yanzhi Wang, “**SuperFlow: An RTL-to-GDS Design Automation Flow for AQFP Superconducting Devices**”, accepted in the DAC 2023, Work-in-Progress.
2. [**23’DCAA**] Peiyan Dong, Mengshu Sun, Alec Lu, Yanyue Xie, Zhenglun Kong, Xin Meng, Xue Lin, Zhenman Fang, Yanzhi Wang, “**Hardware-Efficient Adaptive Token Pruning for Vision Transformers**”, to appear in the 2023 Workshop on DL-Hardware Co-Design for AI Acceleration.
3. [**22’CVPR**] Zhenglun Kong*, Peiyan Dong*, Xiaolong Ma, Xin Meng, Mengshu Sun, Wei Niu, Bin Ren, Minghai Qin, Hao Tang, Yanzhi Wang, “**Enabling Faster Vision Transformers via Soft Token Pruning**”, in T4V: Transformers for Vision, CVPR 2022 (Spotlight).
4. [**21’DAC**] Changlu Liu, Qin Li, Peiyan Dong, Yanming Zhang, Minda Yang, Fei Qiao, Yanzhi Wang, Huazhong Yang, “**A hardware-adapted joint optimization of pruning & quantization for energy-limited speech applications**”, accepted in the DAC 2021, Work-in-Progress.
5. [**21’ISSCC]** Qin Li, Changlu Liu, Peiyan Dong, Yanming Zhang, Tong Li, Minda Yang, Fei Qiao, Yanzhi Wang, Li Luo, Huazhong Yang, “**A 22.3 nJ/Frame low-Memory beyond-real-Time keyword Spotting Chip with Configurable Feature Extraction and Distributed Perceptual Computation**”, accepted in International Symposium on Solid-State Circuits (ISSCC) SRP, 2021.
6. [**20’BARC**] Runbin Shi*, Peiyan Dong*, Tong Geng, Martin Herbordt, Hayden So, and Yanzhi Wang. “**CSB-RNN: A Super Real-time RNN Framework with Compressed Structured Block**”, in Boston Area Architecture Workshop.

Patent Publication

1. [**US Patent**] Yanzhi Wang, Peiyan Dong, Zhengang Li, Bin Ren, Wei Niu, “**Computer-implemented methods and systems for compressing recurrent neural network (rnn) models and accelerating rnn execution in mobile devices to achieve real-time inference**”, US Patent App.

INVITED TALKS

4TH ROAD4NN WORKSHOP, DAC 2023, San Francisco, CA, USA
■ TALK: **Software-Hardware Co-Design: Towards Ultimate Efficiency in Deep Learning Acceleration**

TINYML: BRING DEEP LEARNING MODELS TO TINY DEVICES, DAC 2023, San Francisco, CA, USA
■ TALK: **Algorithm-Software-Hardware Co-Design for AI Acceleration**

HONORS & AWARDS

2023 The EECS Rising Stars Workshop
⇒ **EECS Rising Star**

2022 The 37th Annual AAAI Conference on Artificial Intelligence
⇒ **Oral Paper Award**

2022 The IEEE / CVF Computer Vision and Pattern Recognition Conference Workshop

⇒ **Spotlight Paper Award**

2015 The Mathematical Contest in Modeling (MCM)/The Interdisciplinary Contest in Modeling (ICM)
⇒ **M Award**

2015 China National College Students Math Modelling Competition
⇒ **1st Award (National Level)**

2014 The Mathematical Contest in Modeling (MCM)/The Interdisciplinary Contest in Modeling (ICM)
⇒ **H Award**

2014 China National College Students Math Modelling Competition
⇒ **1st Award (Provincial Level)**

2014 BIT Math Modelling Competition of Beijing Institute of Technology
⇒ **1st Award**

2013 BIT Math Modelling Competition of Beijing Institute of Technology
⇒ **2nd Award**

SELECTED RESEARCH

Efficient DNN Inference for the Deployment on Diverse FPGA Platforms *2019 – Now*

- Propose a hardware-efficient and image-adaptive token pruning framework for efficient yet accurate ViT acceleration on embedded FPGAs.
- Propose an end-to-end acceleration framework with novel algorithm (quantization with optimized nonlinear kernels) and architecture co-design features to enable real-time ViT acceleration on AMD Versal Adaptive Compute Acceleration Platform (ACAP).
- Propose an optimized full-stack RNN framework with a novel compressed structured block (CSB) pruning technique. And implement it on embedded FPGAs with a dedicated compiler.
- Perform a comprehensive qualitative and quantitative comparison of the energy efficiency between FPGA-based and mobile-based (GPUs) DNN executions and provide an in-depth analysis.

Real-time Execution for Various DNNs on Mobile Devices and IoT Devices *2019 – Now*

- Propose a low-bit quantization framework with hardware-oriented nonlinear kernels to boost ViT inference on the edge. Also, design low-bit SIMD-based multipliers to support the practical sub-8-bit computation.
- Propose an activation-guided quantization framework for popular Large Language Models (LLMs) and implement an end-to-end accelerator on mobile CPUs and Raspberry Pis.
- Participate in a mobile inference acceleration framework design that is general to both CNNs and RNNs and leverage fine-grained sparse model structure and compiler optimizations for mobiles.
- Propose a near-sensor processing architecture of feature-configurable distributed network for always-on keywords spotting applications. And the chip is fabricated in a low power 65nm CMOS process.

Efficient Deep Learning on AQFP Logic Family and Related Design Automation *2021 – Now*

- Design an AQFP-based randomized BNN acceleration framework to support feasible BNN execution on AQFP devices.
- Design and release a placement automation tool on AQFP circuits: *AQFP_Placement_v1.0*.
- Design and release a timing analysis tool on four-phase AQFP circuits: *AQFP_Timing_Analysis_Tool.*

Efficient Inference from Model Sparsity and Hardware-oriented Methodology *2019 – Now*

- Propose a speed-aware transformer for end-to-end object detectors, achieving fast on-device latency.
- Propose a framework that utilizes a novel block-based pruning approach and compiler optimizations to RNN-based speech recognition on mobile devices.

PROFESSIONAL EXPERIENCE

- **Futurewei Technologies**, Santa Clara, CA, U.S.A *06/2022 – 09/2022*

Research Internship

- **Northeastern University** *01/2021 - 05/2021*

***Teaching Assistant**, Course: Digital Logic Design Laboratory*

- **Beijing Institute of Technology**, Beijing, China *07/2016 – 12/2016*

***Team Organizer** in Mathematical Modeling Training Center*

** From 01/2020 to the present, all other periods not included above were covered by Research Assistant under the supervision of Prof. Yanzhi Wang.*

COURSES STUDIED

Core Courses

Object Oriented Programming C++
Optimization and Complexity
Data Mining in Engineering
Computer Vision/Pattern Recognition
Advances in Deep Learning

Other Courses

Deterministic Operation Research
Probabilistic Operation Research
Applying Probability & Stochastic Probability