HONGJIE WANG

Department of Electrical and Computer Engineering, Princeton University, NJ, USA
+1 609-480-4317 | Google Scholar link | e: hongjiewang@princeton.edu

EDUCATION

Princeton University — Princeton, NJ, USA
Ph.D. candidate in Electrical and Computer Engineering — Sep 2021 – present
M.A. in Electrical and Computer Engineering — Sep 2021 – Aug 2023

- GPA: 4.00/4.00
- First-Year Fellowship in Natural Sciences and Engineering

Peking University — Beijing, China
B.S. in Physics with honor — Sep 2017 – Jul 2021

- GPA: 3.84/4.00 (rank top 5%); admitted based on performance on Chinese Physics Olympiad (20/300000)
- Selected awards: Excellent Graduate of Peking University and Beijing (top 5%), National Scholarship (top 1%), May 4th Scholarship at Peking University (top 5%), First Prize in China Undergraduate Mathematical Contest in Modeling
- Selected to UGVR Program at Stanford University (top 0.5%, based on outstanding research performance)
- Visiting student at Stanford University and Rice University

PUBLICATIONS

1. **Hongjie Wang**, Difan Liu, Yan Kang, Yijun Li, Zhe Lin, Niraj K. Jha, Yuchen Liu, "Attention-Driven Training-Free Efficiency Enhancement of Diffusion Models", accepted to *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024.*
2. **Hongjie Wang**, Bhishma Dedhia, Niraj K. Jha, "Zero-TPrune: Zero-Shot Token Pruning through Leveraging of the Attention Graph in Pre-Trained Transformers", accepted to *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024.*
3. Haitong Li, Wei-Chen Chen, Akash Levy, Ching-Hua Wang, **Hongjie Wang**, Po-Han Chen, Weier Wan, Win-San Khwa, Harry Chuang, Y.-D. Chih, Meng-Fan Chang, H.-S. Philip Wong, Priyanka Raina. "SAPIENS: A 64-kb RRAM-Based Non-Volatile Associative Memory for One-Shot Learning and Inference at the Edge." *IEEE Transactions on Electron Devices (2021)*.
4. Haitong Li, Wei-Chen Chen, Akash Levy, Ching-Hua Wang, **Hongjie Wang**, Po-Han Chen, Weier Wan, H.-S. Philip Wong, Priyanka Raina. "One-Shot Learning with Memory-Augmented Neural Networks Using a 64-kbit, 118 GOPS/W RRAM-Based Non-Volatile Associative Memory" *IEEE Symposia on VLSI Technology and Circuits (VLSI) 2021*.
5. **Hongjie Wang**, Yang Zhao, Chaojian Li, Yue Wang, Yingyan Lin. "A New MRAM-based Process In-Memory Accelerator for Efficient Neural Network Training with Floating Point Precision" *IEEE International Symposium on Circuits and Systems (ISCAS) 2020*. (oral)

RESEARCH EXPERIENCE

Meta GenAI — Menlo Park, CA, USA
Research Scientist Intern mentored by Dr. Xiaoliang Dai — May 2024 – present
Efficient Long Video Generation with Linear Scaling

- Details of this ongoing project is still confidential.

Adobe Research — San Jose, CA, USA
Research Scientist Intern mentored by Dr. Yuchen Liu — May 2023 – Nov 2023
Attention-Driven Training-Free Efficiency Enhancement of Diffusion Models

- Proposed AT-EDM, a training-free efficiency enhancement framework that leverages rich information from attention maps to accelerate pre-trained Diffusion Models (DMs).
- Generalized the Weighted Page Rank (WPR) algorithm to make it compatible with cross-attention to deduce importance score distribution of tokens.
- Developed a novel similarity-based copy method to recover pruned tokens for convolution.
- Designed a cross-denoising-step heterogeneous pruning schedule that exploits the contribution difference of different denoising steps to the generated image quality.

- Proposed a cross-denoising-step attention-reusing schedule to further reduce the generation overheads.
- Deployed the proposed AT-EDM on SD-XL and conducted both qualitative and quantitative evaluations. AT-EDM reduces the FLOPs cost of SD-XL by 40% with negligible FID score increment and CLIP score loss.
- As the first author, published a paper in ***CVPR 2024.***

Princeton University (Department of Electrical and Computer Engineering) Princeton, NJ, USA
Research Assistant to Prof. Niraj K. Jha Jan 2022 – Apr 2023
Zero-Shot Token Pruning through Leveraging of the Attention Graph in Pre-Trained Transformers

- Developed Zero-TPrune, a zero-shot token pruning method that efficiently utilizes the feature identification ability (i.e., attention graph and embedding vectors) existing in pre-trained Transformers.
- Implemented the Weighted Page Rank (WPR) algorithm that uses the token correlation and coupling hidden in the attention graphs to deduce importance score distribution of tokens.
- Proposed the Emphasizing Informative Region (EIR) aggregation to distinguish the informative regions in different heads.
- Developed Variance Heads Filter (VHF) to prevent the detrimental effect of noisy or meaningless heads.
- Implemented a dynamic training schedule to exploit the learning ability of the proposed Zero-TPrune method.
- Evaluated the proposed Zero-TPrune method on multiple Transformer backbones (on ImageNet dataset). Zero-TPrune improves the throughput of DeiT-S by 45.3% with only 0.4% accuracy loss.
- As the first author, published a paper in ***CVPR 2024.***

Stanford University (Department of Electrical Engineering) Stanford, CA, USA (remote from Beijing, China)
Research Assistant to Prof. Priyanka Raina Jun 2020 – Sep 2020
One-Shot Learning using RRAM-Based Associative Memory

- Developed an L1-distance-based one-shot learning algorithm based on previous meta-learning methods.
- Proposed a quantization method to compress the feature vectors outputted from the pre-trained feature extractor.
- For complex datasets, developed the adaptive assigned weights method to improve the computation precision supported in the RRAM memory array significantly.
- For simple datasets, proposed a new computation strategy to introduce computation sparsity to improve the supported computation precision in the RRAM memory with high efficiency and low latency.
- Evaluated the accuracy loss reduction after applying proposed encoding methods and strategies. Compared with the straightforward encoding method, experimental results show that accuracy loss could be reduced by up to 99%.
- As one of the authors, published a paper in ***VLSI2021*** and a paper in ***IEEE Trans on Electron Devices.***

Peking University (Department of Electrical Engineering and Computer Science) Beijing, China
Research Assistant to Prof. Huailin Liao Jan 2020 – Apr 2021
Process In-Memory based Gated Recurrent Unit (GRU) Accelerator Design for a Low Power Hand Gesture Recognition System

- Proposed a Fast Fourier Transformation (FFT) based algorithm for feature extraction from raw radar signals.
- Developed a fine-grained pipeline at the GRU cell level for higher computation parallelism.
- Developed a coarse-grained pipeline at the GRU layer level to decouple the inputs of GRU for higher computation throughput.
- Evaluated the improvement of the proposed GRU accelerator compared with state-of-the-art works in terms of throughput and energy consumption.

Rice University (Department of Electrical and Computer Engineering) Houston, TX, USA
Research Assistant to Prof. Yingyan Lin Jul 2019 – Nov 2019
SOT-MRAM based Process In-Memory Accelerator for Neural Network Training with Floating-Point Precision

- Developed a new SOT-MRAM memory cell favoring more efficient PIM-based DNN training.
- Proposed a new computation dataflow for 1-bit full addition that requires fewer computing steps to finish addition operations compared with the state-of-the-art design.
- Developed efficient designs of the dominant floating-point multiplication and addition in DNN training.

- Integrated the aforementioned design to realize a SOT-MRAM PIM based DNN training accelerator. Experimental results show that the proposed design offers improvement in terms of energy, latency, and area over a state-of-the-art PIM based DNN training accelerator.
- As the first author, published a paper and delivered a lecture in ***ISCAS 2020***.

Peking University (Department of Electrical Engineering and Computer Science) Beijing, China
Research Assistant to Prof. Huailin Liao Apr 2019 – Jan 2020
Nuclear Magnetic Resonance (NMR) based Non-invasive Blood Glucose Monitor Design

- Simulated the distribution of magnetic fields induced by PCB coils based on finite element analysis.
- Optimized the design of PCB coils to introduce homogeneous enough magnetic fields for NMR.
- Proposed an algorithm with DNN to identify different materials in the solution via the NMR spectrum.

Peking University (Department of Electrical Engineering and Computer Science) Beijing, China
Research Assistant to Prof. Huailin Liao Oct 2018 – May 2019
In-Memory Computing based Neural Network Accelerator with Phase Domain Computing

- Proposed a concept of phase domain computing to support high-efficiency dot-product for quantized neural network processing.
- Developed a new efficient data mapping methodology for phase domain computing.
- Proposed a new loss function for training and inference to improve robustness.
- Designed a mixed-signal in-memory core for efficient depth-wise separable convolution processing.
- Customized the phase domain processor under TSMC 40nm PDK to verify our mapping methodology and architecture.

SELECTED AWARDS AND HONORS

- First-Year Fellowship in Natural Sciences and Engineering (Princeton University) 2021
- Excellent Graduate of Beijing 2021
- Excellent Graduate of Peking University 2021
- PKU Scholarship in Physics 2020
- Xiaomi Scholarship at Peking University 2020
- Student of Merits for 2019-2020 at Peking University 2020
- National Scholarship 2019
- Student of Merits for 2018-2019 at Peking University 2019
- First Prize in the 35th National University Student Physics Competition 2018
- First Prize in China Undergraduate Mathematical Contest in Modeling 2018
- May 4th Scholarship at Peking University 2018
- Student of Merits for 2017-2018 at Peking University 2018
- Gold Medal in 33rd Chinese Physics Olympiad 2016

ADDITIONAL INFORMATION

Course Projects

- **Machine Learning Algorithm**: Zero-shot token pruning with the attention probability in pre-trained Transformers; Privacy leakage and its mitigation in large language models
- **FPGA**: Depth-wise separable quantized CNN acceleration on FPGA
- **ASIC:** DES (Data Encryption Standard) processor design for data encryption and decryption; CNN accelerator with mixed-signal charge domain in-memory computing

Research Interests

- Efficient Machine Learning Algorithms for Low-Power Edge Devices
- Software and Hardware Co-design for High-Efficiency Computing at Edge

Graduate Courses

- **Algorithm**:
  - [ECE 535] Machine Learning and Pattern Recognition
  - [COS 598D] Advanced Topics in Computer Science: Systems and Machine Learning
  - [COS 597G] Advanced Topics in Computer Science: Understanding Large Language Models
  - [ECE 538A] Special Topics in Information Sciences and Systems: Graph Signal Processing and Learning
  - [COS 434] Machine Learning Theory
  - [COS 516] Automated Reasoning about Software

- **Architecture & Circuit Design**:
  - [ECE 575] Computer Architecture
  - [ECE 562] Design of Very-Large Scale Integrated (VLSI) Circuits

Undergraduate Core Courses

- **Architecture and Machine Learning**: Computer Vision and Deep Learning, Introduction to Computer Systems, Introduction to Machine Learning
- **Circuit Design**: Fundamentals of reconfigurable systems (VLSI Design and Its Implementation on FPGA), Principles of Digital Integrated Circuits, Principles of Analog Integrated Circuits, Advanced Principles of Analog Integrated Circuits, Integrated Circuit Design Laboratory, Fundamentals of Electronic Circuits and Experiments
- **Physics and Device**: Semiconductor Physics, Solid State Physics, Electrodynamics, Quantum Mechanics, Device Design and Applications for Modern Integrated Circuits

Skills

- Programming Language: Python, MATLAB, C++, C
- Deep Learning Framework: PyTorch, TensorFlow
- VLSI Design: Verilog
- Mixed-signal Design: Cadence Virtuoso