Ziyi Guan

+86 18823347376 | u3008363@connect.hku.hk |

Education Background

Southern University of Science and Technology Sep. 2017 – Jul. 2021

B.Eng. in Microelectronics, School of Microelectronics *Shenzhen, China*

- Cumulative GPA: 3.68/4.0

The University of Hong Kong Sep. 2021 – Sep. 2025 (Expected)

Ph.D. candidate in Department of Electrical and Electronic Engineering *Hong Kong, China*

Research Interest

- **Large Language Model Optimization (such as quantization, sparsity, low-rank adaption, finetune)**
- **Design lightweight network architecture under constraint hardware (such as CIM architecture).**

Research Experience

Project: A Novel Post-Training Quantization Method for LLMs June. 2023 – Nov. 2023

The University of Hong Kong *Hong Kong, China*

- Propose an Attention-aware Post-Training Mixed-Precision Quantization, which is the first work to quantize LLMs by integrating the attention-based gradients with second-order Hessian optimization, leading to a nuanced update mechanism that enhances the precision throughout the quantization process.
- Propose an innovative Hessian trace-driven mixed-precision quantization scheme that wisely allocates high/low bitwidths across different layers based on their sensitivity, optimizing model performance while maintaining efficiency.
- Conduct extensive experimentation on the LLaMa models and prove that APTQ not only achieves state-of-the-art (SOTA) results on the C4 dataset but also attains near full-precision perplexity at an average quantization of 4 bits. In zero-shot tasks, APTQ also demonstrates superior performance compared to the SOTA approaches.
- Summarizing the work into a conference paper and have been accepted to the 61st IEEE/ACM Design Automation Conference. (DAC), (**DAC 2024**).

Project: An Efficient Hardware-friendly RRAM-based Networks Design Dec. 2022 – May. 2023

The University of Hong Kong *Hong Kong, China*

- We are among the first to design a lightweight isotropic shift-pointwise network with near-100% RRAM crossbar utilization. The proposed PSP and SP networks outperform standard CNNs in model accuracy and hardware metrics.
- We utilize an algorithm-hardware co-design to exploit shift operation in digital domain for spatial mixing and pointwise operation in the analog domain for channel mixing.
- Summarizing the work into a conference paper and have been accepted to the Design Automation and Test in Europe Conference (**DATE 2024**).

Project: A Novel Neural Architecture Search for In-Memory Computing Dec. 2021 – Jun. 2022

The University of Hong Kong *Hong Kong, China*

- Adopted a one-shot NAS to ensure the stability of generalization and customization
- Explored the Pareto fronts between latency and accuracy considering hardware constraints;
- Proposed a novel evaluation function to balance the tradeoff between the accuracy and latency of neural networks.
- Summarized the work into a conference paper and submitted to the 2022 IEEE 16th International Conference on Solid-State Integrated Circuit Technology **(ICSICT 2022)**

Project: Fall Detection System of Elderly People Jun. 2020 – Aug. 2020

Southern University of Science and Technology *Shenzhen, China*

- Proposed a novel framework that can complete the fall detection and pose estimation tasks simultaneously.
- Developed a network by *spatio-temporal joint-point model* to process the time-series video frames fast.
- Make a summary of the work to the paper and have been accepted to Design Automation and Test in Europe Conference (**DATE 2021**).

Project: Smart Agriculture based on IOT cloud platform Mar. 2019 – Jun. 2019
Southern University of Science and Technology *Shenzhen, China*

- Worked as a leader in a group of three and won the Huawei ICT Competition 2018-2019 Global Final Innovation Competition: Third prize worldwide.
- Completed the crop disease detection on the AI cloud platform.
- Realized the real-time monitoring of environment by the IOT board.

Publication

1. **(DAC'24) Ziyi Guan**, Hantao Huang, Yupeng Su, Hong Huang, Ngai Wong and Hao Yu, "APTQ: Attention-aware Post-Training Mixed-Precision Quantization for Large Language Models", In Proceedings of DAC 2024: *61st IEEE/ACM Design Automation Conference. (DAC)*, San Francisco, CA, June 23-27, 2024.
2. **(DATE'24) Ziyi Guan**, Boyu Li, Yuan Ren, Muqun Niu, Hantao Huang, Graziano Chesi, Hao Yu and Ngai Wong, " An Isotropic Shift-Pointwise Network for Crossbar-Efficient Neural Network Design", *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, March 25, Valencia, 2024.
3. **(DATE'24)** Zikun Wei, Tingting Wang, Chenchen Ding, Bohan Wang, **Ziyi Guan**, Hantao Huang, and Hao Yu, " FMTT: Fused Multi-head Transformer with Tensor-compression for 3D Point Clouds Detection on Edge Devices", *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, March 25, Valencia, 2024.
4. **(DATE'23)** Changhai Man, Cheng Chang, Chenchen Ding, Ao Shen, Hongwei Ren, **Ziyi Guan**, Yuan Cheng, Shaobo Luo, Rumin Zhang, Ngai Wong and Hao Yu, "RankSearch: An Automatic Rank Search towards Optimal Tensor Compression for Video LSTM Networks on the Edge",*Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2023.
5. **(ICSICT'2022) Ziyi Guan**, Wenyong Zhou, Yuan Ren, Rui Xie, Hao Yu, and Ngai Wong. 2022. "A Hardware-Aware Neural Architecture Search Pareto Front Exploration for In-Memory Computing." in *2022 IEEE 16th International Conference on Solid-State Integrated Circuit Technology (ICSICT).* IEEE, 2022, pp. 1–4.
6. **(TECS'2022)** Shuwei Li, **Ziyi Guan**, Changhai Man, Ao Shen, Wei Mao, Shaobo Luo, Rumin Zhang, and Hao Yu. 2022. "A Fall Detection Network by 2D/3D Spatio-temporal Joint Models with Tensor Compression on Edge." in *ACM Transactions on Embedded Computing Systems (TECS)* vol. 21, no. 6, pp. 1–19, 2022
7. **(DAC'2022 Workshop) Ziyi Guan**, Yuan Ren, Wenyong Zhou, Rui Xie, Quan Chen, Hao Yu, Ngai Wong, "XMAS: An Efficient Customizable Flow for Crossbarred-Memristor Architecture Search." in *59th Design Automation Conference (DAC) Engineering Track*
8. **(DATE'2021) Ziyi Guan**, Shuwei Li, Yuan Cheng, Changhai Man, Wei Mao, Ngai Wong, and Hao Yu,"A Video-based Fall Detection Network by Spatio-temporal Joint-point Model on Edge Devices", *Design, Automation & Test in Europe Conference & Exhibition (DATE).* IEEE, 2021, pp. 422–427.
9. Cheng Chang, Xuejuan Zhu, **Ziyi Guan**, Kai Li, Laimin Du, Wei Mao, and Hao Yu, 2020. An automatic searching method and device for precision and rank decomposition of Recurrent Neural Network. PCT International Patent PCT/CN2020/141379, filed December 2020.

Awards and Honors

World Awards May. 2019

- Huawei ICT Competition 2018-2019 Global Final Innovation Competition: Third Prize
  Huawei ICT Competition was held by Huawei Technologies Co, attracting more than 100,000 students from more than 1,600 colleges and universities in 61 countries. Our work "Smart Agriculture based on Huawei Internet of Things Cloud Platform" was awarded as the third prize (ranked fourth in all teams).

School Honors

- Jiangbolong Merit Student Scholarship (Valued 10000 RMB), Jun. 2021
- Merit Student Scholarship 2019-2020: Third Prize, Oct. 2020
- Merit Student Scholarship 2018-2019: Second Prize, Oct. 2019
- Innovation Contest 2019: Best Application Award, Jul. 2019

Computer Skills

Python, Pytorch, Tensorflow.v1, Tensorflow.v2, C/C++, Java, MATLAB, LaTex