Ziyi Guan

+86 18823347376 | u3008363@connect.hku.hk | homepage: https://easongzy.github.io/ |

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 Pruning Method for LLMs Mar. 2024 – Jul. 2024
The University of Hong Kong *Hong Kong, China*
- **Motivation:** To optimize Large Language Models (LLMs) by reducing their size and computational requirements while maintaining high performance.
- **Method:**
  1) Introduced LLM-Barber, a one-shot pruning framework that utilizes block-aware sparsity to enhance global optimization in LLM pruning.
  2) Rebuilt sparsity masks by identifying non-salient weights that become salient after pruning and retaining these, while simultaneously pruning weights that are non-salient, thus optimizing overall model performance by reallocating sparsity.
  3) Proposed an innovative pruning metric based on the product of weights and gradients, using first-order Taylor series for efficient importance evaluation, reducing computational complexity.
  4) Demonstrated the versatility and efficiency of LLM-Barber across various post-training pruning techniques, achieving state-of-the-art performance in both perplexity and zero-shot tasks. This work is currently under review for the **AAAI 2025** conference.
- **Technique adopted:** LLM weight pruning/sparsity, LLM analysis, Global optimization, Metric development

Project: A Novel Post-Training Quantization Method for LLMs June. 2023 – Nov. 2023
The University of Hong Kong *Hong Kong, China*
- **Motivation:** To enhance the efficiency and precision of LLMs through innovative quantization techniques.
- **Method:**
  1) Proposed an Attention-aware Post-Training Mixed-Precision Quantization integrating attention-based gradients with second-order Hessian optimization.
  2) Introduced a Hessian trace-driven mixed-precision quantization scheme for optimized performance.
  3) Extensive experimentation on LLaMa models demonstrated state-of-the-art results on perplexity and superior zero-shot task performance.
  4) Summarized the work into a conference paper accepted by 61st IEEE/ACM Design Automation Conference (**DAC 2024**).
- **Technique adopted:** LLM quantization, LLM analysis, Mixed-precision quantization, Hessian optimization

Project: An Efficient Hardware-friendly RRAM-based Networks Design Dec. 2022 – May. 2023
The University of Hong Kong *Hong Kong, China*
- **Motivation:** To design lightweight and efficient neural networks compatible with RRAM crossbar technology and improve RRAM utilization and hardware and software performance.
- **Method:**
  1) Designed a lightweight isotropic shift-pointwise network with near-100% RRAM crossbar utilization.
  2) Utilized an algorithm-hardware co-design for efficient spatial and channel mixing.
  3) Summarized the work into a conference paper accepted by Design Automation and Test in Europe Conference (**DATE 2024**).
- **Technique adopted:** RRAM crossbar utilization, Network design, Algorithm-hardware co-design

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

The University of Hong Kong *Hong Kong, China*

- **Motivation:** To explore efficient neural architectures considering hardware constraints for in-memory computing.
- **Method:**
  1) Adopted a one-shot NAS to ensure stability in generalization and customization.
  2) Explored the Pareto fronts between latency and accuracy considering hardware constraints and proposed a novel evaluation function.
  3) Summarized the work into a conference paper accepted by the 2022 IEEE 16th International Conference on Solid-State Integrated Circuit Technology (**ICSICT 2022**).
- **Technique adopted:** Neural architecture search (NAS), Lightweight network design, Hardware constraints analysis, Pareto front exploration

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

Southern University of Science and Technology *Shenzhen, China*

- **Motivation:** To develop a reliable and efficient fall detection system for elderly care.
- **Method:**
  1) Proposed a novel framework for simultaneous fall detection and pose estimation.
  2) Developed a network using a spatio-temporal joint-point model for fast processing of video frames.
  3) Utilized Huawei Cloud's ModelArts AI platform and Atlas accelerator to train the fall detection model.
  4) Summarized the work into a conference paper accepted by Design Automation and Test in Europe Conference (**DATE 2021**).
- **Technique adopted:** Fall detection, Pose estimation, Tensor decomposition, Spatio-temporal modeling, LSTM Compressioin

Project: Smart Agriculture based on IOT cloud platform Mar. 2019 – Jun. 2019

Southern University of Science and Technology *Shenzhen, China*

- **Motivation:** To enhance agricultural monitoring and disease detection using IoT and AI technologies.
- **Method:**
  1) Develop a smart agriculture system and completed crop disease detection on the AI cloud platform.
  2) Implemented real-time environmental monitoring using IoT devices.
  3) Successfully competed and won the Third Prize in the **Huawei ICT Competition** 2018-2019 Global Final Innovation Competition.
- **Technique adopted:** IoT, AI cloud platform, Real-time monitoring, Team leadership, Presentation Skill

Publication

1. **(AAAI'2025 Under Review)** Yupeng Su*, **Ziyi Guan***, Xiaoqun Liu, Tianlai Jin, Dongkuan Wu, Graziano Chesi, Ngai Wong, Hao Yu, "LLM-Barber: Block-Aware Rebuilder for Sparsity Mask in One-Shot for Large Language Models", In Proceedings of the AAAI Conference on Artificial Intelligence, 2025 (Under review) (*represents equal contribution)

2. **(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.

3. **(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.

4. **(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.

5. **(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.

6. **(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.

7. **(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
8. **(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*
9. **(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.
10. 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