管子义

电话: 18823347376 丨邮箱: u3008363@connect.hku.hk 丨现居城市: 深圳
微信: Easongzy 丨个人网站: https://easongzy.github.io/
年龄: 25岁 丨性别: 男 丨民族: 汉
当前状态: 博士 丨意向城市: 深圳 丨求职意向: 大模型算法工程师

教育经历

香港大学 - 电子科学与技术 博士 Electrical and Electronic Engineering 2021年09月 - 2025年09月

海外QS前100

- 导师：黄毅教授 (Prof. Ngai Wong)
- 研究方向：专注于大语言模型（LLM）和多模态大模型的优化与推理加速，涵盖模型压缩小型化技术的多个领域，包括但不限于量化、蒸馏、剪枝、稀疏化等
- 获得香港大学博士全额奖学金

南方科技大学 - 微电子科学与工程 本科 深港微电子学院 2017年09月 - 2021年07月

双一流

- GPA：3.68/4.0 (专业前10%)
- 荣誉奖项：**华为ICT大赛2018-2019全球创新大赛决赛:三等奖**、2018-2019校优秀学生奖学金二等奖、2019-2020校优秀学生奖学金三等奖。
- 本科在EDA顶会Design,Automation & Test in Europe Conference & Exhibition **(DATE 2021)(CCF-B)**发表一作论文，获得直接前往香港大学电子系PhD攻读的offer。

论文、专利与竞赛

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. **（CCF-A会议，接受率23%）**

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. **（CCF-B会议，接受率23%）**

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. **（CCF-B会议，接受率23%）**

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. **（CCF-B会议，接受率25%）**

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. **（CCF-B 期刊）**

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. **（CCF-B会议，接受率24%）**

10.【专利】管子义，黄洪，罗少波，王锦萍，2024. 一种基于层级的无训练的大模型混合量化方法及系统. 中国发明专利申请号：243655-PI-237-JK，申请日期：2024年07月22日。

科研项目经历

1. 一种新型的大语言模型后训练剪枝(Post-Training Pruning)方法 2024年03月 - 2024年08月

- **项目背景：**随着大语言模型（LLM）的规模不断扩大，计算和存储成本随之激增，剪枝技术成为减小模型规模的关键。但现有的剪枝方法通常只在局部层面上优化，忽视了全局的稀疏性分布问题，导致性能下降。
- **方法与成果：**为解决这一问题，我们提出了LLM-Barber，一种基于块感知稀疏性的剪枝框架，能够全局优化自注意力层和MLP层的稀疏性分布。该方法通过识别并保留在剪枝后仍然重要的权重，重构稀疏性掩码，并采用一阶泰勒级数进行重要性评估，极大地

降低了计算复杂度。实验结果显示，该方法在LLaMa模型的剪枝中，不仅在困惑度上实现了SOTA性能，还在零样本任务中超越了当前的后训练剪枝方法。该工作已提交**AAAI 2025 (CCF-A)** 评审。

- **掌握技能：大语言模型剪枝/稀疏、全局稀疏性优化、一阶泰勒级数、LLaMa模型开发。**

2. 一种新型的大语言模型后训练量化(Post-Training Quantization)方法 2023年06月 - 2023年11月

- **项目背景**：大语言模型在硬件上的高效部署常常受限于其庞大的参数量和高计算需求，量化技术可以有效降低计算负担，但现有的量化方法在保证精度的同时，往往无法兼顾效率，且传统的量化方法再极低比特量化下精度损失严重。
- **方法与成果：**我们设计了一种注意力感知的后训练混合精度量化方法APTQ，通过结合基于Hessian的跟踪优化，细化了不同层的比特宽分配。APTQ创新性地利用Hessian矩阵追踪来评估不同层的灵敏度，在LLaMa模型上实验中，实现了4位量化的精度，与全精度模型的困惑度几乎一致，且在零样本任务中也超越了SOTA水平。该研究成果已被第61届 IEEE/ACM 设计自动化大会 Design Automation Conference. (**DAC 2024) (CCF-A)**会议接受。
- **掌握技能：大语言模型量化、混合精度量化、Hessian优化、LLaMa模型开发。**

3. 一种高效的硬件友好型RRAM网络设计 2022年12月 - 2023年05月

- **项目背景：**RRAM（可变电阻式存储器）作为新兴的存储技术，能够大幅提升计算效率，但传统的神经网络架构未能充分利用RRAM的硬件特点，导致计算和存储效率低下。
- **方法与成果：**我设计了一种轻量级各向同性移位点网络，通过算法-硬件协同设计，最大化了RRAM的交叉条利用率，达到接近100%的硬件利用率。该网络不仅在硬件资源上实现了最优配置，还在多个计算任务中超越了标准的CNN架构，显著提升了性能。该研究已被**DATE 2024 (CCF-B)** 录用。
- **掌握技能：RRAM神经网络、轻量级网络设计、算法-硬件协同设计、硬件加速。**

4. 面向内存计算的新型神经网络架构搜索 2021年12月 - 2022年06月

- **项目背景**：在面向内存计算的硬件环境下，神经网络架构的设计必须平衡延迟与精度，现有的架构搜索方法未能充分考虑硬件约束，导致在特定硬件上的适应性差。
- **方法与成果：**我们开发了一种基于一次性NAS的架构搜索方法，通过引入硬件延迟和精度的帕累托优化，确保架构在RRAM硬件下的通用性和稳定性。实验结果显示，该方法在硬件资源受限的情况下依然能维持较高的模型精度，并具有良好的扩展性。该工作已被**ICSICT 2022**录用。
- **掌握技能：神经网络架构搜索 (NAS)、硬件约束分析、帕累托边界探索、RRAM。**

5. 老人跌倒检测项目 2020年05月 - 2020年08月

- **项目背景：**老年人跌倒是全球老龄化社会中的重大健康风险，现有的检测系统存在计算复杂度高、实时性差的问题。
- **方法与成果：**提出了一种能够同时完成跌倒检测和姿态估计的框架，利用基于时空结合点的模型对时间序列视频帧进行快速处理，提高了检测速度和精度。该研究的成果被**DATE 2021 (CCF-B)** 录取。
- **掌握技能：跌倒检测、姿态估计、时空建模、LSTM。**

6. 基于物联网云平台的智慧农业 2019年03月 - 2019年06月

- **项目背景：**农业环境监测和病害检测是智慧农业的核心应用，传统的检测方法通常需要大量人力，且监测实时性差。
- **方法与结果：**设计并开发了智能农业系统，结合物联网设备和AI云平台，实时监测农作物环境并检测病害。该系统的创新设计在**华为ICT大赛**2018-2019全球创新大赛中获得**全球三等奖。**
- **掌握技能：物联网、AI云平台、农业环境监测、团队领导，演讲技巧。**

技能与其他

- **技能：** 熟悉 Python, Pytorch, Tensorflow.v1, Tensorflow.v2, C/C++, Java, MATLAB, LaTex 等
- **语言：** 英语（CET-6）雅思 （IELTS）6.5