管子义

18823347376 u3008363@connect.hku.hk 深圳
https://easongzy.github.io/
1998-11 男 汉
博士 深圳 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：大语言模型优化（如量化、稀疏、微调），在CIM硬件下设计轻量级网络架构（如RRAM架构）
- 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。

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07月

- **项目动机**：优化大语言模型，减少其规模和计算需求，同时保持高性能。
- **方法与成果：**

1. 提出了LLM-Barber，一种利用块感知稀疏性的单次剪枝框架，该方法考虑了完整的注意力层和MLP层，以实现LLM剪枝的全局优化。
2. 通过识别剪枝后变得重要的非显著权重并保留这些权重，同时剪除非显著权重，重建稀疏性掩码，从而通过重新分配稀疏性来优化整体模型性能。
3. 提出了一种基于权重与梯度乘积的创新剪枝度量，使用一阶泰勒级数进行高效的重要性评估。与基于二阶hessian的方法相比，该度量提供了更准确的修剪决策并降低了计算复杂度。
4. 对LLaMa系列模型进行实验，该方法在多种后训练剪枝技术中表现出色,在困惑度和零样本任务中均达到了最先进的性能水平。该工作目前已提交AI顶会Association for the Advancement of Artificial Intelligence 2025 **(AAAI 2025)(CCF-A)**进行评审。

- **掌握技能：**大语言模型剪枝/稀疏、大语言模型分析、全局优化、度量开发。

2. 一种新型的大语言模型后训练量化方法 2023年06月 - 2023年11月

- **项目动机**：通过创新的训练后量化技术，提高大语言模型的量化效率，减小量化误差。
- **方法与成果：**

1. 提出了一种注意力感知的训练后混合精度量化，这是第一个通过将基于注意力的梯度与二阶Hessian优化相结合来量化llm的工作，从而产生了一种细微的更新机制，在整个量化过程中提高了精度。
2. 提出一种创新的Hessian跟踪驱动的混合精度量化方案，该方案根据不同层的灵敏度明智地分配高/低比特宽，在保持效率的同时优化模型性能。
3. 在LLaMa模型上进行广泛的实验，证明APTQ不仅达到了最先进的水平(SOTA)在C4数据集上的结果，但也达到了接近全精度的困惑，平均量化为4位。在零射击任务中，与SOTA方法相比，APTQ也显示出优越的性能。
4. 将工作已被第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。

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

- **项目动机**：设计适配于RRAM的轻量级高效神经网络，提高RRAM利用率和软硬件性能。
- **方法与成果：**

1. 首次提出轻量级各向同性移位点网络，其RRAM交叉条利用率接近100%。提出的PSP和SP网络模型精度和硬件指标上优于标准CNN。
2. 利用算法-硬件协同设计，利用数字域的移位操作进行空间混频，利用模拟域的点向操作进行信道混频。
3. 将工作总结成会议论文，并已被欧洲设计自动化和测试会议 Design Automation and Test in Europe Conference **(DATE 2024)(CCF-B)**接受。

- **掌握技能：**RRAM设计、轻量级网络设计、算法-硬件协同设计。

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

- **项目动机**：探索考虑硬件约束的高效神经网络搜索架构。
- **方法与成果：**

1. 采用一次性one-shot NAS来搜索适配于RRAM的网络结构，保证了通用化和定制化的稳定性。
2. 在考虑硬件约束的情况下，探索了基于延迟latency和精度accuracy之间的帕累托边界。

3. 提出了一种新的评估函数来平衡神经网络的准确率和延迟之间的权衡。
4. 将工作总结成会议论文，提交至2022年IEEE第16届国际学术会议固态集成电路技术**(ICSICT 2022)。**

- **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。

5. 老人跌倒检测项目 2020年05月 - 2020年08月

- **项目动机：**开发可靠高效的老人跌倒检测系统。
- **方法与成果：**

1. 提出了一种能够同时完成跌倒检测和姿态估计的框架。
2. 提出了一种基于时空结合点模型的网络对时间序列视频帧进行快速处理。
3. 对论文的工作进行了总结，并已经被欧洲设计自动化与测试会议 Design Automation and Test in Europe Conference **(DATE2021)(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

- **项目动机：**利用物联网和AI云平台技术提高农业环境监测和病害检测。
- **方法与结果：**

1. 开发智能农业系统，在AI云平台上完成农作物病害检测。
2. 利用物联网IOT设备实施实时环境监测。
3. 参赛作品获得**华为ICT大赛**2018-2019全球创新大赛总决赛:**全球三等奖。**

- **掌握技能：**物联网、AI云平台、实时监控、团队领导，演讲技巧。

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.

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.

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