管子义

u3008363@connect.hku.hk 深圳
Wx555328778 https://cn.linkedin.com/in/ziyi-guan-78a177197
https://easongzy.github.io/
27岁 男 汉族
博士应届毕业生 杭州、深圳、上海、北京、香港 大模型算法研究员

教育经历

香港大学 海外QS前100 2021年09月 - 2025年11月

Electrical and Electronic Engineering 博士 Electrical and Electronic Engineering 全日制 香港

- 导师：黄毅教授 (Prof. Ngai Wong)
- 研究方向：
  1. **大语言模型LLM的压缩算法优化研究与推理加速：**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、稀疏化等技术，**以提高模型性能并减少计算资源需求。
  2. **多模态LLM Agent研究**：参与多模态大模型**LLM GUI Agent** (如**APP Agent**)设计和研发，结合和其他LLM Agent相关**后训练Post-train**技术，如**监督微调(SFT)，强化学习(RL)**等，提升Agent的功能和效率。
  3. **面向约束硬件的轻量级神经网络架构设计：**协助设计适应**RRAM架构**等**存算一体硬件的轻量级神经网络架构**，确保在硬件约束下高效部署模型。

南方科技大学 双一流 2017年09月 - 2021年07月

微电子科学与技术 本科 深港微电子学院 全日制 中国香港

- 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。**登上华为官方公众号阅读1.6万浏览量**

工作经历

字节跳动 2025年11月 - 至今

AI Infra研究员 Seed Infra 异构硬件组 杭州

- **核心职责:** 负责公司核心业务模型在国产化AI芯片上的性能优化，涵盖训练与推理加速的压缩算法优化解决方案。负责公司核心大模型在字节内部大型推理平台的性能优化，推动模型压缩算法从实验验证到工程落地，实现稀疏注意力和低比特量化加速。
- **模型压缩与加速:** 深入研究并实践多种前沿压缩算法，包括KV Cache压缩、模型量化 (Quantization) 与稀疏化 (Sparsity) 技术，旨在极致提升模型的推理效率与部署性能。
- **前沿技术探索和落地:**
  - **自研SALS decode稀疏注意力加速：**完成 Decode 端稀疏落地，实现历史 KV page/block 的精确选择；Prefill 阶段设计 Q Group 共享、KV Block TopK、recent/fixed tail 和 skip layer 策略，实现长上下文稀疏加速；在 32K 长上下文场景下端到端吞吐提升约 11%，并保持精度基本无损。(https://arxiv.org/abs/2510.24273)
  - **自研 SALS Prefill 稀疏注意力加速：**面向长 prompt Prefill 阶段的全量 Attention 计算瓶颈，设计 Q Group 共享稀疏 pattern、KV Block TopK、head/tail dense、recent/fixed tail 与 bad layer skip 等策略，将逐 query 的稀疏选择转化为 group-level / block-level 调度；在长上下文场景下实现 Prefill 稀疏加速，核心 Attention 子项获得约 3x 级加速，整模型 Prefill 随序列长度增长获得可见收益。
  - **低比特量化：**设计并验证 2bit Weight 量化和 KV 量化方案，覆盖 FA/MoE/EP 通信路径，解决长上下文 prefill OOM 与 GPU 利用率低问题，线上模拟验证总掉分 <1 分。
  - **算子精度看护：**建立 Consumer-Aware IPD 与 KL/SNR/CosSim/AutoEval 组合指标体系，实现算子级精度风险预警和问题定位。
- **掌握技能**：LLM 推理优化、Prefill/Decode 稀疏加速、2bit weight & KV 量化、FA/MoE/EP 量化、算子精度分析、 推理框架、长上下文 continuous batching。

华为技术有限公司香港研究所 2024年11月 - 2025年09月

算法实习研究员 费马实验室 香港

- **项目背景与研究价值:** 面向 GUI 测试领域，突破 Test Agent 技术，支持鸿蒙、安卓 APP、PC 测试、游戏测试等业务的低成本测试。探索 GUI端到端的自动化能力，逐步使测试 Agent 具备真人测试员的能力。
- **主要贡献：**1）算法优化： 开发了基于测试意图和预标注UI图谱的用例路径自动生成算法，提高了核心功能测试的覆盖率。2）RAG知

识库构建：基于知识图谱和离线感知算法构建了一个强大的RAG数据库，从应用UI图谱中生成高效的路径检索，支持更精准的测试用例生成。3）框架优化: 将RAG知识库移植到真机在线决策框架中，显著降低了测试执行时间，提高了测试的可靠性，支持跨平台应用测试（Android、HarmonyOS）降低测试成本呢，提高测试效率。已将工作归纳成文章已被EMNLP 2025 main接受

- **掌握技能：**LLM GUI Agent,、RAG、知识图谱、测试自动化、团队协作

论文、专利与竞赛

1. (TCAD 2025) Ziyi Guan, et al,"APTQ+: Attention-FFN-aware Post Quantization for Layerwise LLM Acclerator on FPGA" accepted in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD 2025) (CCF-A期刊，EDA顶刊)

2. (EMNLP 2025 Main) Ziyi Guan, et al,"KG-RAG: Enhancing App Decision-Making via Knowledge Graph-Driven Retrieval-Augmented Generation" accepted to **EMNLP 2025 Main Conference (CCF-B会议,NLP顶会)**

3. (DAC 2025 Poster) Yupeng Su*, **Ziyi Guan***, et al, "LLM-Barber: Block-Aware Rebuilder for Sparsity Mask in One Shot for Large Language Models" at the 62nd Design Automation Conference, June 22-25 in San Francisco, CA **(DAC 2025) (CCF-A会议，EDA顶会，接受率23%)** (*represents equal contribution)

4. (DAC 2024 Oral) Ziyi Guan, et al."APTQ: Attention-aware Post-Training Mixed-Precision Quantization for Large Language Models"，at the 61st Design Automation Conference, June 23-27 in San Francisco, CA **(DAC 2024) (CCF-A会议，接受率20%)**

5. (DAC 2025 Poster) Dingbang Liu*, **Ziyi Guan***, et al. " A Highly Energy-Efficient Binary BERT Model on Group Vector Systolic CIM Accelerator", at the 62nd Design Automation Conference, June 22-25 in San Francisco, CA (**DAC 2025) (CCF-A会议，接受率23%)** (*represents equal contribution)

6. (DATE 2024 Oral) Ziyi Guan, et al, "An Isotropic Shift-Pointwise Network for Crossbar-Efficient Neural Network Design"，Design, Automation & Test in Europe Conference & Exhibition **(DATE 2024) (CCF-B会议，EDA顶会，接受率23%)**

7. (DATE 2021 Oral) Ziyi Guan, et al, "A Video-based Fall Detection Network by Spatio-temporal Joint-point Model on Edge Devices"，Design, Automation & Test in Europe Conference & Exhibition **(DATE 2021) (CCF-B会议，接受率23%)**

8. (ICSICT' 2022) Ziyi Guan, et al, "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国际会议)**

9. (TECS' 2022) Shuwei Li, **Ziyi Guan,** et al, "A Fall Detection Network by 2D/3D Spatio-temporal Joint Models with Tensor Compression on Edge" in ACM Transactions on Embedded Computing Systems **(TECS2022)** . **(CCF-B期刊)**

10.【国内专利】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：2024109821464，申请日期：2024年07月22日。

11.【国际专利】Cheng Chang, Xuejuan Zhu, **Ziyi Guan**, et al2024. Automatic search method and apparatus for precision and decomposition rank of Recurrent Neural Network 国际专利公开号 WO/2022/141189

项目经历

1. 基于知识图谱RAG的LLM GUI Agent决策推理框架(KG-RAG) 2024年11月 - 2025年05月

算法研究员&第一作者 华为港研所费马实验室 香港

- **项目背景：**当前的基于大语言模型（LLM）的图形用户界面（GUI）代理在执行复杂的移动应用任务时，仍然面临有限的应用特定知识的挑战。尽管应用界面过渡图（UTGs）能提供结构化的导航表示，但因其提取不完全和集成效率低下，导致其应用受到限制。
- **方法与成果：**本研究提出了KG-RAG框架，通过将不完整的UTGs转化为高质量的结构化向量数据库，优化了实时检索效率。该方法结合了意图驱动的大语言模型搜索算法，能生成针对用户需求的导航路径，从而提升agent决策能力。实验结果表明，KG-RAG在任务完成率、决策准确性和任务完成步骤数上均显著提升，并提出了涵盖30个中文APP适用于中国移动生态的CNBench基准，推动了移动应用交互领域的进步。该工作已提交**EMNLP 2025 (CCF-B)** 评审。
- **掌握技能：GUI Agent、Test Agent、知识图谱、检索增强生成（RAG）、大语言模型prompt engineering**

2. 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。该工作已被**DAC 2025 Poster接收(CCF-A)** 。
- **掌握技能：LLM Post-train、大语言模型剪枝/稀疏、全局稀疏性优化、LLM压缩优化、LLaMa模型开发**

3. 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：LLM Post-train、大语言模型量化、LLM压缩优化算法、混合精度量化、Hessian优化、LLaMa模型开发。**

4. 基于二值化和CIM加速的高效BERT模型研究 2024年02月 - 2024年10月

- **项目背景：**随着大语言模型（LLM）在边缘设备上的部署需求增加，其高计算和存储开销成为挑战。为此，我们结合二值化和CIM（计算存储一体化）技术，提出高效的BERT模型优化与加速方案，兼顾模型压缩与精度保持。
- **方法与成果：**我们提出了一种结合二值化权重分裂与知识蒸馏的优化方法，将BERT模型的权重和激活量化为BF16×1-b格式，兼顾模型压缩与精度保持。同时设计了一种基于组向量阵列的SRAM-CIM加速器，实现了32倍模型压缩和能效提升。实验显示，该方案在SST-2数据集上准确率仅损失0.5%，能效达到20.98 TOPS/W，面积效率提升10.25倍。该工作已被**DAC 2025 Poster (CCF-A) 接受。**
- **掌握技能：模型量化蒸馏、硬件架构设计与优化、算法硬件协同设计（ Algorithm-Hardware Co-design ）**

5. 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**：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（ Algorithm-Hardware Co-design ）、硬件加速。**

6. 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。**

7. 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。**

8. 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。**

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