刘昊

hao.liu@kaust.edu.sa · (+86) 13699121719 · linkedin

教育背景

阿卜杜拉国王科技大学 (KAUST), 沙特 2022 – 至今

在读博士研究生 计算机科学, 预计 2026 年 7 月毕业

北京航空航天大学, 北京市 2019 – 2022

硕士 网络空间安全

北京航空航天大学, 北京市 2015 – 2019

本科 电子信息工程

科研经历

面向近边缘加速器的 ViT 协同推理机制

2025 年 6 月 – 2026 年 1 月

KAUST, 导师: Prof. Suhaib A. Fahmy

- **提出协同推理框架：**针对 Vision Transformers (ViTs) 提出了一种通用的协同推理框架，利用轻量级边缘模型配合多个近边缘专家模型 (Near-edge Experts)，在边缘端资源受限的情况下显著提升了推理效率与精度。
- **设计鲁棒路由机制：**利用边缘模型的 Top-k 预测结果设计了零开销的路由决策机制，能够精准判断并激活所需的专家模型。
- **渐进式专家训练策略：**提出 Progressive Specialist Training 策略，在保持模型通用特征提取能力的同时，强制提升各专家模型在特定领域的专业性。
- **系统实测：**搭建包含真实边缘设备与近边缘节点的测试平台，并在 CIFAR-100 数据集上完成完整的性能验证与分析。

建模近边缘加速器上的 DNN 分割推理

2024 年 6 月 – 2025 年 5 月

KAUST, 导师: Prof. Suhaib A. Fahmy

- **通信压缩机制优化：**实验证明基于自编码器 (Autoencoder) 的特征图压缩方案，在精度保持和压缩率方面均优于传统的量化 (Quantization) 和张量低秩分解方法。
- **精确性能建模：**结合 GPU 数据摄入 (Data Ingestion) 特性与基于包的网络传输机制，提出精确的计算与通信性能模型，模型预测与实际部署测量值高度吻合。
- **真实系统验证：**构建多加速器硬件系统，使用 VGG16 和 ResNet50 模型在 CIFAR-100 数据集上进行广泛测试，验证不同 Batch Size 下，单分割点方案的有效性。
- **模型指导搜索：**利用该性能模型指导了多分割点 (Multi-split)、动态带宽分配及多租户场景下的最优配置搜索。

面向分割计算的网络架构搜索

2023 年 6 月 – 2023 年 9 月

暑期实习 清华大学, 导师: 汪玉教授 (Prof. Yu Wang)

- 学习研究网络架构搜索框架 *aw_nas* (`https://github.com/walkerning/aw_nas`)。
- 探索在分割计算约束下，如何自动化搜索最优的神经网络分割点与子结构。

深度神经网络的分割计算研究

2022 年 10 月 – 2024 年 5 月

KAUST, 导师: Prof. Suhaib A. Fahmy

- **问题建模：**将 DNN 多点分割问题建模为综合考量精度、计算延迟及传输开销的联合优化问题。
- **多自编码器分割方案：**提出通过插入多个自编码器对 DNN 进行切分，并将不同分区调度至异构设备执行，实现端到端延迟与系统能耗之间的最佳权衡。
- **实验验证：**基于 ResNet50 和 VGG16 模型，在 CIFAR-100 和 ImageNet 数据集上方案验证。

基于滤波器秩的卷积神经网络剪枝方法

2020 年 6 月 – 2021 年 12 月

北京航空航天大学, 导师: 关振宇教授

基于物理不可克隆函数 (PUF) 的物联网设备认证 2019 年 3 月 – 2019 年 12 月

北京航空航天大学, 导师: 关振宇教授

面向态势感知与可信协同的可穿戴自组网设备 2018 年 9 月 – 2019 年 2 月

北京航空航天大学, 导师: 关振宇教授

论文发表

1. **Hao Liu**, Suhaib A. Fahmy. "Ask the Expert: Collaborative Inference for Vision Transformers with Near-Edge Accelerators". *Under Review*, 2026.
2. **Hao Liu**, Mohammed E. Fouda, Ahmed M. Eltawil, Suhaib A. Fahmy. "Practical Modeling for Split DNN Inference on Near-Edge Accelerators". *Under Review*, 2026.
3. **Hao Liu**, Mohammed E. Fouda, Ahmed M. Eltawil, Suhaib A. Fahmy. "Split DNN Inference for Exploiting Near-Edge Accelerators". In *IEEE International Conference on Edge Computing and Communications (IEEE EDGE)*, 2024.
4. **Hao Liu**, Zhenyu Guan, Peng Lei. "A Filter Rank Based Pruning Method for Convolutional Neural Networks". In *IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2021.
5. Zhenyu Guan, **Hao Liu**, Yuyao Qin. "Physical unclonable functions for IoT device authentication". *Journal of Communications and Information Networks*, 2019.
6. Zhenyu Guan, Jiawei Li, **Hao Liu**, Dawei Li. "A Wearable Ad Hoc Device for Situational Awareness and Trusted Collaboration". In *Smart Blockchain: Second International Conference*, 2019.

技术项目

NNStreamer 部署与自定义插件开发 [Docker] [Code]

C++, GStreamer 边缘计算协作推理

- 在 NVIDIA Jetson 边缘平台上搭建基于 NNStreamer 的分布式推理框架。
- 使用 C/C++ 开发自定义 GStreamer 插件实现分布式推理，优化了跨多设备的 DNN 协同推理性能。

ResNet 加速方案对比：Vitis-AI 与 hls4ml [Code]

FPGA, Xilinx Vitis-AI 性能评估

- 在 Xilinx Alveo U55C 加速卡上基于 Xilinx Vitis-AI 流程实现了 ResNet 推理加速。
- 对比分析该方案与 hls4ml 实现方案在推理性能与硬件资源利用率上的差异。

FPGA 游戏设计：Block the Brick [Code]

Verilog HDL 数字逻辑设计

助教经历

ECE 265P: AI Training (助教, 沙特内政部项目) 2025 年 4 月
CS256: Digital Design and Computer Architecture (助教, KAUST) 2022 年 9 月
数字电路与系统 (助教, 北航) 2020 年 9 月

获奖情况

研究生学业二等奖学金 2019, 2020
第十一届全国大学生信息安全竞赛二等奖 （队长） 2018
第二十八届"冯如杯"学生学术科技作品竞赛二等奖 2018
北航大学生科研训练计划 (SRTP) 二等奖 （获 3 万元项目资助） 2018
第八届"蓝桥杯"单片机设计与开发大赛三等奖 2017