刘炼

liulian211@mails.ucas.ac.cn · (+86) 159-8257-6242 · https://leliyliu.github.io/ ·

个人简介

本人博士研究方向主要集中于面向神经网络模型的软硬件协同优化，尤其是针对于**大语言模型推理系统**的协同优化，包括：模型量化剪枝算法研究，异构近存大模型推理系统，大模型推理算子的协同优化，异构系统的内存管理等。本人在人工智能和计算机体系结构领域**顶级国际会议期刊（CCF-A）**发表 **6 篇一作论文**，并拥有从算法设计，系统优化到芯片架构设计的全栈优化经验。

教育背景

计算技术研究所, 中科院, 中国科学院大学, 北京 2021.09 – 至今

- 在读博士研究生 计算机系统结构, 预计 2026 年 6 月毕业
- 导师：王颖研究员，李华伟研究员

南开大学, 天津 2017.09 – 2021.06

- 学士 物联网工程　GPA: 3.88/4, Rank: 1/45
- 导师：李涛教授，卢冶副教授

科研项目

基于异构存算扩展的 LLM Serving 系统研究，负责人 2024.09 – 至今

项目背景：LLM Serving 不仅对服务器算力以及数据访问带宽有较大需求，还对容量有极大的需求。此外，Serving 过程的动态性（不同用户 requests 的差异性）也对服务效率产生了极大影响。本项目尝试通过**异构存算扩展**来解决上述问题。

研究方法：本项目利用存算芯片来处理 LLM 推理过程中的访存受限算子，如 Attention。并通过内存层次的异构扩展，利用 Attention 计算的冷热特性，实现容量扩展。此外，为了解决不同用户 requests 差异性导致的资源竞争，本项目还提出了一个**解耦池化策略**，通过构建结构的 NPU 池和 PIM 池，并使用动态调度策略，来进一步提升 Serving System 的吞吐量。

项目效果：通过有效的存算扩展，相比于现有系统，实现了 7.17× 计算效能优化。相关工作计划投稿（包含已投稿）至 MICRO'25, HPCA'26,ASPLOS'26。

面向端侧 LLM 部署的稀疏算法-系统优化协同研究，负责人 2023.09 – 2024.11

项目背景：端侧 LLM 部署存在资源受限的挑战。稀疏性技术（参数、激活、注意力稀疏）可降低计算与内存需求，但现有方案**面临算法-硬件脱节、显存瓶颈及异构系统管理复杂等问题**。

研究方法： 1. 量化压缩：动态混合精度量化（W4A4/W4A8），开发 COMET-W4Ax 算子；2. 参数卸载：基于激活稀疏性划分冷热参数，结合 GPU 与 NDP-DIMM 近存扩展，实现 Hermes 异构系统；3. 异构管理：UVM2 统一内存管理，混合 CPU-GPU 并行计算，动态负载均衡优化 KV Cache 冷热迁移。

项目效果：通过利用稀疏特性和异构扩展，在端侧实现了大幅度的性能提升，相关工作发表于 **ASPLOS'25, HPCA'25**，后续工作准备投稿至 EuroSys'26。

面向神经网络动态特性的算法-硬件协同设计，负责人 2021.09 – 2023.06

项目背景：利用软硬件协同设计来实现神经网络的高效部署是一个热点话题，然而，之前的工作主要关注于利用模型的静态稀疏性来提高整体的运行效率。受到动态神经网络思路的影响，本项目尝试通过探索神经网络中的动态特性，来实现模型的高效压缩以及推理效率的优化。

研究方法：该项目首先探索了量化算法与模型架构的协同设计，设计了一个自动化优化框架来实现统一探索。在此基础上，进一步提出动态量化策略，并设计具有灵活数据流映射的专用硬件来解决动态量化中存在的拥塞问题。最后，我们构建了动态神经网络的形式化描述，并根据此分析在 NPU 上如何通过扩展实现通用动态神经网络加速，设计了轻量化在线调度策略和关联性的预取方案。

项目效果：该项目在静态神经网络模型压缩的基础上，通过软硬件协同设计，进一步取得了从 $2.12\times \sim 4.23\times$ 的性能提升和能耗节省。相关工作发表于 **TCAD'24, DAC'24, TC'25**.

实习经历

大语言模型稀疏算法研究，*AMD 算法实习生* 2024.01 – 2024.12

基于权重稀疏性的 LLM Post-training 剪枝算法研究：该研究通过分析现有大语言模型剪枝算法中存在的不足，包括使用一次性 pruning mask 以及现有 pruning mask 选择存在的明显偏好性问题，设计了基于轻量化搜索的**大语言模型剪枝标准的自动优化框架**，相比于之前的剪枝后 LLM 有效地减少了 2.49 的混淆度。相关成果发表于 **COLING'25, ICML'25**。

基于 LLM KV Cache 融合稀疏性算法研究：在长上下文本下 KV Cache 的存储量会成为 LLM 推理的瓶颈。为了解决这一问题，研究提出了一套**基于 token 相似性的 KV cache 融合**压缩策略，在少量精度损失情况下，支持 2.21× ∼ 3.94× 的 KV Cache 压缩。

论文

COMET: Towards Practical W4A4KV4 LLMs Serving
Lian Liu, Long Cheng, Haimeng Ren, Mengdi Wang, Xiaowei Li, Yinhe Han, and Ying Wang
Accepted by **ASPLOS'25**, 2025. **(CCF-A)**.

Make LLM Inference Affordable to Everyone: Augmenting GPU Memory with NDP-DIMM
Lian Liu*, Shixin Zhao*, Bing Li, Mengdi Wang, Xiaowei Li, Yinhe Han, and Ying Wang
Accepted by **HPCA'25**, 2025. **(CCF-A)**.

BaWA: Automatic Optimizing Pruning Metric for LLMs with Balanced Weight and Activation
Lian Liu, Xiandong Zhao, Guanchen Li, Dong Li, Mengdi Wang, Yinhe Han, and Ying Wang
Accepted by **ICML'25**, 2025. **(CCF-A)**.

DNA: A General Dynamic Neural Network Accelerator
Lian Liu, Jinxin Yu, Mengdi Wang, Xiaowei Li, Yinhe Han, and Ying Wang
Accepted by **TC'25**, 2025. **(CCF-A)**.

An Automatic Neural Network Architecture-and-Quantization Joint Optimization Framework for Efficient Model Inference
Lian Liu, Ying Wang, Weiwei Chen, Xiandong Zhao, Huawei Li*, Xiaowei Li, and Yinhe Han
Accepted by **TCAD'24**. **(CCF-A)**.

Drift: Leveraging Distribution-based Dynamic Precision Quantization for Efficient Deep Neural Network Acceleration
Lian Liu, Zhaohui Xu, Yintao He, Ying Wang, Huawei Li, Xiaowei Li, and Yinhe Han
Accepted by **DAC'24**, 2024. **(CCF-A)**.

PAM: Processing Across Memory Hierarchy for Efficient KV-centric LLM Serving System
Lian Liu*, Shixin Zhao*, Yutian Zhou, Yintao He, Mengdi Wang, Yinhe Han, and Ying Wang
Submitted to **MICRO'25**, 2025. (**CCF-A**, In submission).

Enhanced One-Shot Pruned Pre-trained Language Models through Sparse-Dense-Sparse Mechanism
Guanchen Li, Xiandong Zhao, **Lian Liu**, Zeping Li, Dong Li, Lu Tian, Jie He, Ashish Sirasao, Emad Barsoum
Accepted by **COLING'25**, 2025. **(CCF-B)**.

On-Line Fault Protection for ReRAM-Based Neural Networks
Wen Li, Ying Wang, Cheng Liu, Yintao He, **Lian Liu**, and Huawei Li
Accepted by **TC'23**, 2023. **(CCF-A)**.

* Co-first author

其他

- 语言能力: 大学英语六级证书
- 其他：熟练使用 Python, C++ 等编程语言，了解面向“算法-系统-体系结构”全栈优化

获奖情况

中科院计算所 所长优秀奖	2023 年 12 月
南开大学 优秀毕业生，优秀毕业论文	2021 年 6 月
南开大学 国家奖学金	2019 年 10 月
“龙芯杯”系统能力培养大赛 全国三等奖	2019 年 8 月