刘炼

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

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
- 导师：李涛教授，卢冶副教授
- 绩点：3.88/4，排名：1/45

研究兴趣

- LLM-centric Architecture Design, 面向 *LLM* 的架构设计
- Processing-in-Memory System, 基于存算的系统优化
- Algorithm-System Co-design, 基于算法和系统的协同优化框架

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).

DAQU: 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**, 2024. (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
Shixin Zhao*, **Lian Liu***, Yutian Zhou, Yintao He, Mengdi Wang, Yinhe Han, and Ying Wang
Submitted to **ISCA'25**, 2025. (CCF-A, In submission).

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
Submitted to **ICML'25**, 2025. (CCF-A, In submission).

DNA: A General Dynamic Neural Network Accelerator
Lian Liu, Jinxin Yu, Mengdi Wang, Xiaowei Li, Yinhe Han, and Ying Wang
Submitted to **TC'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

科研项目

基于层次化存算扩展的 LLM Serving 系统, 共同一作 2024.09 – 2025.03

低成本解决 LLM Serving 中 **KV cache 计算，访存和存储挑战**的系统设计

- 提出了**一套新颖的跨存储层异构计算策略**，以有效利用上下文局部性和层次化存算（PIM）设计
- **全栈解决方案设计**，包括计算方法，调度优化和系统扩展
- 相比于 DGX-H100 vLLM 系统，实现 **7.20**× **性能提升** & **7.17**× **计算效能优化（tokens/$）**

基于近存 DIMM 设计的 LLM 本地部署系统, 共同一作 2024.03 – 2024.10

基于近存设计的高效且低成本的大模型推理系统，以实现**高效的本地部署**

- 一个基于消费级 GPU (RTX 4090) 和近存 DIMM 设计的**异构计算系统**
- 基于异构计算系统的**两阶段（离线 & 在线）负载均衡调度**
- 实现了 13.75 **tokens/s for LLaMA2-70B**, 75.25× 加速相比于现有系统

基于聚类思路的 LLM KV cache 层次化压缩策略研究, *AMD* 实习期间 2024.07 – 2024.12

- 提出了一种基于 **KV cache 融合**的压缩策略
- 支持**灵活的压缩粒度调整**，以实现在线推理优化
- 相比于现有推理系统设计，能够实现 2.21× ∼ 3.94× 效率提升

轻量化大模型权重剪枝方法研究, *AMD* 实习期间 2024.03 – 2024.09

- **系统分析**现有剪枝策略在 LLM 剪枝中导致的**不均匀分布问题**
- 基于轻量化搜索的大模型剪枝标准的**自动优化框架设计**，**减少 2.49 的混淆度**

LLM 的细粒度混合精度量化和推理优化系统, 一作 2023.06 – 2024.06

- 提出了一个全新的细粒度混合精度量化算法来**减少计算和访存开销**
- **基于 CUTLASS** 实现了**第一个在 GPU 上支持细粒度混合精度计算的 GEMM 算子**, COMET-W4Ax
- 相比于现有推理系统，在 A100 上实现了 2.03× **加速比**

基于动态混合精度量化的 NN 加速器设计, 一作 2022.09 – 2023.09

- 面向通用神经网络模型的**动态精度量化算法**设计
- 支持灵活动态精度矩阵计算的**可配置 NPU 架构设计**
- 相比于现有 NPU 设计，**实现** 2.85× **加速 &** 3.12× **功耗节省**

通用动态神经网络加速器设计, 一作 2022.03 – 2023.11

- 基于**贪心调优策略的在线调度探索**，以实现对动态张量计算的高效处理
- 基于**关联性的预取方案设计**，以减少动态神经网络执行过程中冗余数据搬运
- 相比于面向静态神经网络的 NPU 设计，**实现** 3.48× **加速 &** 3.03× **功耗节省**

面向神经网络架构和量化策略的联合优化框架, 一作 2021.09 – 2023.03

- 新颖的**自动化搜索框架设计**，以探索最优的网络架构和量化方法组合策略
- 提出了一种**可微搜索策略**以支持快速搜索，和**精度迁移策略**以实现多精度模型之间的快速迁移
- 能够在没有引入额外精度损失的情况下，在现有低精度 NPU 上**实现** 3.5× **加速比**

获奖情况

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
三等奖，“龙芯杯”系统能力培养大赛 2019 年 8 月