张昊

150-3679-8729 | 2120230710@mail.nankai.edu.cn

模型压缩、大模型推理优化加速

教育背景

南开大学（985）	计算机技术	硕士（保送）（前 10%）	2023.09 月——至今
中国矿业大学（211）	电子信息科学与技术	学士（校级优秀毕业生）（前 3%）	2019.09 月——2023.06 月

学术成果

- **Unveiling Super Experts in Mixture-of-Experts Large Language Models** **ICLR 2026（CCF-A），第三作者**
  - ◇ 在 DeepSeek-R1、Qwen3-30B-A3B、Mixtral-8x7B 等 MoE 大模型上完成大规模实验评估；
  - ◇ 实现并运行 Super Expert 剪枝实验，分析其对模型推理能力的影响；
  - ◇ 在 AIME、MATH500、GPQA、HumanEval 等 benchmark 上进行性能评测；
  - ◇ 通过 activation profiling 分析 expert 激活异常与 attention sink 机制。

实习经历

- **北京三快在线科技有限公司（美团） M2CA 团队** **2025.04 月——2025.08 月**
  - ◇ **LLM_plat：FP8 Block 自动量化业务支持**：支持 Friday 平台 LongCat / FlashCat（32K 长上下文）系列模型高效量化转换与部署；采用分块动态 FP8 量化方案并集成至 LLM_plat 实现自动化量化转换。
    效果：推理吞吐提升 30%–40%，vLLM 启用 FP8 Block 内核生成后吞吐提升 25%–40%，精度损失 <1%。
  - ◇ **Qwen3 / FlashMoE 自动量化部署工具链开发（vLLM / SGLang）**：基于 MTPQ 支持 W8A8 INT8 / FP8 / FP8 Block / W4A16 等多种量化模式，集成 GPTQ / Quarot / AWQ / SQ 等主流算法，形成可复用 MoE 量化工具链。
  - ◇ **MoE-Quant 量化验证**：对 DeepSeek 类模型进行 MoE 量化与效果验证。

项目经历

- **大语言模型解码阶段 KV Cache 联合压缩优化研究（硕士毕设）** **2025.08 月——2026.04 月**
  - ◇ 在前期 AIDCS 工作基础上，进一步针对长上下文解码阶段 KV Cache 的存储与访存瓶颈，提出**基于 KV Cache 稀疏与量化的联合压缩解码加速方法（Joint Sparsification and Quantization of Key-Value Cache, JSQKV）**，包含差分稀疏、双窗口在线执行、Hadamard 稳定化 Per-Token-Tile 量化及 Bitmap-Based Sparse-Quant 数据格式与解码算子。
  - ◇ 以 Meta-Llama-3-8B 为例，在输入 4096、输出 256、**70%** KV 稀疏 + **2-bit** 配置下，Batch Size 为 2 / 4 / 8 时端到端吞吐达到 **27.06 / 47.12 / 75.71 tokens/s**，其中 Batch Size 为 4 时较 Dense 基线提升约 **44.2%**。
  - ◇ 与 AIDCS 联合部署验证表明，两类方法在作用路径上具有较强正交性；在相同长上下文设置下，联合方法 Batch Size 为 4 时端到端吞吐较 Dense 基线提升约 **44.9%**。
- **基于输入特征动态稀疏的大模型解码推理优化研究（横向课题）** **2023.11 月——2025.06 月**
  - ◇ **AIDCS：基于输入特征动态稀疏化的推理加速方法（Adaptive Input-Driven Computation Sparsity）**：针对解码阶段线性层权重访存瓶颈，提出输入特征动态稀疏方案，结合**核密度估计（KDE）**、**桥函数**、逐块重构、逐层重构与关键 Token 保护机制，在总体稀疏度达到 **50%** 时精度损失不超过 **1%**。
  - ◇ **解码性能收益**：以 Llama2-7B 为例，AIDCS 在 Batch Size 为 1 / 2 时解码吞吐较 Dense 基线提升约 **1.6**× / **1.35**×。
  - ◇ **Triton 高性能稀疏 Kernel**：基于 Triton 实现高性能稀疏向量-矩阵乘内核，加速大语言模型解码推理，实现端到端推理在 BatchSize 为 1 时加速 **1.8 倍**、BatchSize 为 10 时加速 **1.2 倍**。
- **面向 Transformer 类智能模型的高能效计算架构研究（166 工程项目）** **2022.11 月——2024.06 月**
  - ◇ **基于硬件感知的多目标优化模型剪枝与混合比特位宽量化**：针对 Transformer 类网络模型，研究基于硬件感知的多目标优化结构化剪枝方法，以计算量和参数量作为硬件资源敏感的代理约束，使用遗传算法对各层进行精细化化剪枝率和比特位宽配置。通过优化实现模型存储空间压缩至全精度网络的 1/6，且平均量化比特位宽**不高于 8 比特**，同时模型精度损失控制在 **2%** 以内。
  - ◇ **面向国产 FPGA 的 Transformer 硬件加速电路设计与优化**：针对国产 FPGA 资源受限的特点，设计并优化**乘累加计算内核**和 **Softmax 计算内核**，以加速多头注意力机制的各个模块。研究基于**多部查找表（Multipartite Table）**的拟合方法和硬件电路设计，实现低精度量化非线性函数的快速近似拟合，显著提升计算效率和硬件资源利用率。
  - ◇ **模型随机失活量化与全量化优化**：研究随机失活重构算法与残差算法，研究 LayerNorm 算子融合的全整数量化方案，减少数据移动开销，优化计算内核设计。通过实现全整数量化，使得 LayerNorm 推理过程可在板端完成，无需移动到 host 端，进一步提升 FPGA 硬件平台的推理效率。

专业技能

- 熟悉 Python、C++ 等编程语言；了解 Cuda、Triton 等异构编程模型；
- 熟悉常用深度学习算法、业界常用模型压缩、加速方法等；
- 熟悉 Linux 系统；熟悉 Pytorch、Onnx、llama.cpp、vLLM、SGLang 等框架，了解 Lmdeploy、Triton 等框架；

荣誉奖项

- 本科：三次校级优秀学生、校级一等奖学金、华为智能基座奖学金、上海能源奖学金；
- 研究生：新生奖学金、公能奖学金；昇腾 AI 创新/算子挑战赛等奖项。