张津豪

(+86) 133-0945-0315 | zhangjinhao315@bupt.edu.cn | github.com/Lvhaoojieming

教育背景

- **北京邮电大学 (BUPT)** *2023.09 – 2027.06 (预计)*
  电子工程学院 · 电信工程 · 工学学士 **GPA: 3.62 / 4.00 (均分 86)**
  核心课程：C 语言 (96), 信号与系统 (95), 科研实践 (94), 深度学习 (90), 线性代数 (88), 机器学习 (88)

专业技能

- **编程与框架**：熟练掌握 **Python**, **C++**, **PyTorch**, **Transformers**；精通 **Megatron-LM**, **DeepSpeed** 进行大规模分布式训练 (Multi-node) 与显存优化。
- **强化学习 (RL)**：精通 **RLHF** (PPO) 与 **DPO** 算法；具备从 Reward Model 构建到 Policy 优化的完整经验。
- **MoE 架构**：熟悉 Sparse MoE 设计；深入理解专家路由与负载均衡，解决稀疏训练稳定性问题。
- **量化与压缩**：精通 PTQ/QAT 及 Hessian 优化 (GPTQ, AWQ)；熟悉 KV Cache 压缩、剪枝与蒸馏技术。

科研经历

中国科学院计算技术研究所 (ICT, CAS) · (客座学生) *2025.05 – 至今*

- **研究方向**：大语言模型 (LLM) 高效推理、后训练 (Post-training) 及 MoE 架构优化。
- **混合专家量化架构 (MoQE)**：
  - **问题**：解决传统量化方法在 MoE 架构上因专家激活稀疏性导致的精度崩塌问题。
  - **方法**：**主导设计**了基于 Token 敏感度的动态路由机制 (Sensitivity-aware Routing)，实现了不同位宽专家的自适应混合计算。
  - **成果**：在不同模型上，在保持推理速度不变的前提下，将量化带来的精度损失**降低了 40%**，优于现有 SOTA 方法。
- **低比特量化与自适应压缩 (CALM & HeRo-Q)**：
  - **自适应模块化 (CALM)**：**首创**基于 CKA 的层级相似度分析法，构建自动化流水线动态分配量化位宽。
  - **海森矩阵优化 (HeRo-Q)**：针对 2-bit/3-bit 极端场景，提出基于 **Hessian Matrix** 特征值的权重重构算法，有效解决了极低比特下的梯度消失与优化不稳定难题。
- **大规模训练系统工程**：
  - 基于 **Megatron-LM** 与 **DeepSpeed (ZeRO-3)**，参与了百亿参数 (10B+) 模型在 **A800 集群**上的预训练与 RL 后训练。

北京邮电大学 · 本科生科研项目 *2024.09 – 2025.06*

- **核心贡献**：在单片机上部署 YOLOv5/v8 模型；设计 PID 算法实现小车实时循迹。

代表论文

MoQE: Improve Quantization Model Performance via Mixture of Quantization Experts [2025.08]
Jinhao Zhang, Yunquan Zhang, Daning Chen | **ICML 2026** (Under Review)

CALM: A CKA-Guided Adaptive Layer-Wise Modularization Framework for LLM Quantization [2025.12]
Jinhao Zhang, Yunquan Zhang, Boyang Zhang, Daning Cheng | **ACL 2026** (Under Review)

HeRo-Q: A General Framework for Stable Low Bit Quantization via Hessian Conditioning [2026.01]
Jinhao Zhang, Yunquan Zhang, Boyang Zhang, Zeyu Liu, Daning Cheng | **ICML 2026** (Under Review)

获得奖项

◇ 北京邮电大学电子信息杯一等奖 2025.07
◇ 北京邮电大学程序设计竞赛一等奖 2024.04
◇ 全国大学生数学竞赛二等奖 2024.09