DeclK

1605224559@qq.com | hongkun20sme@gmail.com | 上海
算法工程师 | 大模型推理优化

教育背景

南京大学 - 电子信息 - 硕士 2020.09 - 2023.06
研究方向：三维感知算法

南京大学 - 材料物理 - 本科 2016.09 - 2020.06
GPA 4.43 / 5.0，综合排名 11 / 135

工作经历

理想汽车 - 多模态大模型量化、投机采样负责人 2023.07 - 2025.11

- **VLM/VLA 模型车端交付**
  利用 mlc-llm 的机器学习编译系统，核心参与构建了车端多模态认知大模型，成为全行业首个全量推送用户的车端大模型。在 mlc-llm 框架中从 0 到 1 构建了 SigLIP、SAM、跨模态 projector 以及 diffusion head 模块，与 Qwen2 模型构建完整的 VLM/VLA。通过引入 memory bank 流式视频推理机制，显著压缩视觉编码时延，最终使 2.1B 参数的 VLM 推理延迟稳定低于 250 ms，大幅超越原定 500 ms 的基准线
- **车端大模型量化体系**
  领导打造了车端 VLM/VLA 模型的高效量化框架，该框架能够在 10 分钟内完成对模型的量化、编译以及精度评测的完整链路，以支持上游算法快速发版需求。通过优化 GPTQ 与 AWQ 算法，完成对模型的 4-bit 压缩，最终实现的模型精度损失在 0.5% 以内
- **投机采样训练以及推理框架**
  领导设计并实现了车端 VLM/VLA 模型的多模态投机采样训练以及推理框架，实现了 Medusa & EAGLE 双引擎，分别能够将 decode 时延加速 3x 和 6.9x。通过深度优化 EAGLE 方案，单步平均命中 15+ tokens，结合词表裁剪和动态 draft length，将 decode 时延降低到 50-80 ms，成为业界首个将多模态 EAGLE 投机采样部署到车端的 SOTA 方案
- **大模型推理优化服务**
  领导开发了 FQT (FullQuantTransformer) 通用量化框架，以及 EasyEagle 通用投机采样框架为整个自动驾驶部门提供大模型推理优化服务。FQT 覆盖 GPTQ/AWQ/SmoothQuant 等主流量化算法，支持 W4/W8/A4/A8/A16 多种精度，兼容 Qwen dense、MoE、ViT 等多样模型，并适配 vLLM、SGLang 等云端量化格式。EasyEagle 在训练侧提供 FSDP、torch compile、梯度累计、张量并行等高效 trick，并具备了强大的兼容性，能够将任意 transformers 库当中的模型作为 base model 进行投机采样训练与推理。最新迭代的多模态 EAGLE3 将云端 Agent 模型的投机命中率从 3.86 提升至 5.53 (+43%)

宏景智驾 - 算法实习生 2022.08 - 2022.10

- 在自动泊车场景下负责分割与检测模型的量化感知训练，将量化模型在地平线芯片以及 Orin 上高效部署

仙途智能 - 算法实习生 2022.04 - 2022.07

- 优化点云目标检测模型，集成 SC-Conv 自矫正卷积、角点辅助模块及 IoU 修正分支，提升 mAP 4.71%

项目经历

基于角点辅助网络的自集成点云目标检测 2022.01 - 2023.01

- 通过角点辅助网络与教师-学生蒸馏提升点云目标检测器性能，利用 SECOND backbone 融合角点与 BEV 特征，一致性损失提升泛化性。在 ONCE 数据集上 mAP 从 57.03 提升至 63.04。

专业技能

- 技术栈：PyTorch, TensorRT, MLC-LLM, CUDA, Transformers
- 算法方向：多模态大模型、投机采样、模型量化、点云检测、BEV 感知、时序分析

其他

- 学业奖学金一等奖（硕士）
- 人民奖学金二等奖（本科）
- 语言：英语（CET-6）
- 兴趣：网球（大网赛全国总决赛男团第六名）