李文

leeven8888@163.com · (+86) 150-3520-2407

教育背景

浙江大学 (985, 杭州, 推免) 硕士, 大数据技术与工程 2023.09 – 2026.03

电子科技大学 (985, 成都) 学士, 计算机科学与工程学院 2019.09 – 2023.06

实习经历

字节跳动 AI Lab Research / 视频生成模型项目 2024.10 – 2025.03

背景介绍: 实习所在部门为视频生成模型算法工程团队，负责模型的数据预处理、大规模训练、模型推理优化等任务。在实习期间，**本人负责支持模型的推理优化策略迭代**，包括但不限于加速：dit 视频生成模型、text embedding 模型、video caption 模型，通过复现开源推理优化技术、针对自研模型进行改进，实现了从算子、模型结构、并行策略的全方位加速。在实习期内，我们负责加速优化的模型在公司内部竞争中脱颖而出，且加速完成的模型成功上线“即梦 AI”和火山方舟。

- 针对自研 DiT 视频模型，进行激活权重量化、dit cache、triton 算子融合等加速策略的调研与落地，提升模型端到端推理速度；
- 使用创新 cache 方法提升 DiT 平均推理速度约 1.1 倍，在 vbench、ssim、psrn 等指标上均优于现有 cache 方法；
- 自研 vlm 类型 video caption 模型针对每个请求使用唯一且重复的 token 计算 logits，为了提升模型推理吞吐，优化 page attention 算子，适配 vllm 推理框架；
- 自研 text embedding 模型采用双向注意力的类似 gpt attention，为了提升 prefill 运算速度，优化并适配 tensorrt-llm 框架。

美团 基础研发平台 / 原生多模态大模型项目 2024.05 – 2024.10

背景介绍: 实习所在部门为原生多模态大模型算法工程团队，负责模型的大规模训练、推理优化等任务。在实习期间，**本人负责支持 mllm 的训练框架迭代及支持业务 dit 模型的推理优化**。在实习期内，我们负责训练加速的自研模型成功点火启训；加速后的业务文生图模型上线 wow app；沉淀研发公司内部大模型多模态推理框架，为业务模型推理部署提供一键式服务。

- 在 sdpa 出现前，模型自定义 attention mask 没有现成接口。优化 flash attention kernel，并接入 megatron 训练框架，相较使用 pytorch 裸写，mfu 从 27% 提升到 54%；
- 对业务文生图模型进行推理优化，主要使用 tensorrt，提升推理速度约 1.3 倍，并完成模型压测、上线；
- 参与公司内部多模态大模型推理框架研发，集成针对小模型的自动 tensorrt 优化以及针对大模型的 ulysses、xdit 等并行、融合算子替换等功能，并成功使用该框架优化部署多个自研视觉模型。

项目经历

基于文本提示的图像编辑系统 2022.10 – 2023.6

- 项目描述：复现并重新训练图像编辑模型 StyleCLIP，使用 triton inference server 对预训练模型进行部署，搭建网站提供编辑服务
- 职责与成果：实现了响应快、延迟低的图像编辑系统。相较于直接调用模型进行推理，响应时间缩短了约 40%。本科毕业设计成绩 90+

技能与荣誉

- 了解 Triton、Cuda 编程和并行计算；了解 TensorRT、TensorRT-LLM、vLLM 等推理方案；
- 了解 C/C++，面向对象等，了解 Python 等语言；熟悉常见数据结构，数组、栈与队列、树等；
- 熟悉 Linux 操作系统，熟练使用 git、docker、cmake 等。
- 本科曾获：校一等奖学金 × 3，校优秀毕业生