江云飞

年龄：26岁 | 18683202286 | jiangyunfei23@163.com

5年工作经验 | 求职意向：算法工程师

教育经历

西南石油大学（双一流） 本科 计算机科学与技术 2016-2020

在校经历：

曾多次获得一等奖学金、优秀学生干部，参加互联网+大赛，获得校级优秀奖

在校期间加入老师团队, 参与一些实验室项目，接触过一些图像算法相关的知识

已通过英语CET-6，具有一定的英文读写能力

专业技能

1. 熟悉Transformer及MOE模型的模型架构，例如Qwen、Llama和DeepSeek等
2. 熟悉常用的注意力机制，比如MHA、MQA、GQA、MLA、FlashAttention等
3. 熟练使用一些开源的大模型训练框架，如transformers、llamafactory、peft、trl等
4. 熟悉模型训练过程中的前向传播和反向传播流程，并对LoRA微调技术有较深的理解
5. 熟悉常见的分布式训练工具的原理和使用方式，比如DDP、DeepSpeed、FSDP、Accelerate等
6. 熟悉常见的并行优化策略，比如数据并行DP，模型并行PP、TP等
7. 熟悉常见的训练优化技术，如梯度累积、梯度检查点、混合精度训练、量化训练等
8. 熟悉常用推理框架vLLM以及SGLang，对SGLang中的KV Cache管理以及Scheduler机制较为了解
9. 熟悉基本的图像算法原理，比如图像分类、目标检测、语义分割、关键点检测等
10. 熟悉Pytorch和Tensorflow深度学习框架，能够熟练使用python完成算法开发或论文复现等
11. 熟悉大模型训练中RLHF的相关训练方法及训练流程，包括PPO、DPO、ORPO等
12. 了解序列并行SP，专家并行EP以及上下文并行CP等
13. 了解大模型应用中的Agent和RAG相关技术，比如OpenManus、Computer-use Agent等
14. 了解一些多模态相关知识，比如文生图、文生视频、文生语音、VLM模型（Qwen2.5 VL）等
15. 了解一些编译优化技术，比如torch compile、triton、cuda graph等

工作经历

北京九章云极科技有限公司 算法工程师 2022.10-至今

AI平台算子开发，大模型训练框架开发，大模型部署推理等

行为科技（北京）有限公司 图像算法 2022.08-2022.10

负责加油站烟火识别项目的全链路实现

成都玻尔兹曼科技有限公司 图像算法 2020.08-2022.06

图像算法设计实现、文档编写以及算法模块维护等

项目经历

Transtech训练框架开发

项目描述

自主研发的具备高扩展性的大语言模型训练框架，支持多源异构数据接入、异步流式数据加载、配置化训练流程控制与分布式断点重训，兼容 HuggingFace / Accelerate / FSDP 等生态，适用于科研与生产环境中的大模型预训练与微调任务

个人工作

1. 主导开发了模型断点保存与恢复模块，支持 HuggingFace中的PreTrainedModel与PEFT微调模型的统一处理，封装FSDP、DeepSpeed、DDP的训练状态保存/加载逻辑，适配单卡及各种分布式训练场景
2. 实现多类型训练状态持久化与恢复，包括模型权重、优化器、学习率调度器、数据加载器、随机种子等，保障断点重训一致性并且保存的权重支持在accelerate以及transformers框架中加载使用
3. 结合训练代码做该模块的内部功能测试以及分布式场景下训练时断点保存与恢复的效果测试

大模型训练微调服务

项目描述

在公司的AlayaNew平台上，参与LM Lab产品中大模型训练微调模块的设计和开发工作，为用户提供完整的大模型训练示例pipeline，用户只需提供自定义数据集或者选择预置的数据集就可实现基础大模型训练或者基础大模型微调

个人工作

1. 借助transformers及peft库构建大模型训练微调模块，兼容目前市面上主流的大模型，如llama、qwen等
2. 在训练微调模块支持pretrain和sft，集成混合精度训练、梯度检查点和梯度累积等各种加速策略
3. 结合aim包实现对训练过程的全程追踪监控，包括运行参数、训练指标、资源消耗以及运行日志等
4. 评估模块开发：基于instruct_eval实现对模型在 MMLU、BBH 和 HumanEval 等评估任务中的性能测试
5. 编写模块运行脚本，支持一键式做单卡、单机多卡以及多机多卡训练，能够适配不同规模的训练需求

大模型推理部署服务

项目描述

在公司的AlayaNew平台上，通过部署DeepSeek-R1模型向客户提供API服务，供客户直接使用或者构建第三方应用，从而达到算力消纳的目的

个人工作

1. 调研支持DeepSeek系列模型部署的框架，如vLLM、SGLang
2. 使用SGLang框架部署模型以及做Benchmark性能测试
3. 构建MCP服务来测试部署好的模型的工具调用能力
4. 根据性能测试结果，找到较优的并发设置、并行方式以及优化参数（fa3、dp_attention等）
5. 按照最佳参数组合部署上线以及推理加速探索：PD分离、TensorRT-LLM框架测试等

APS平台算子开发

项目描述

APS是一个支持自动建模和工作流拖拽式建模的算法开发平台，涵盖了算法开发的各个环节：包括数据录入、数据处理、pipeline初始化、建模训练、评估以及pipeline归档使用等，其中最核心的部分就是建模训练中使用的算子开发

个人工作

1. 使用Vision Transformer对传统cv算子进行改造，包括图像分类、目标检测、语义分割等；具体使用的模型是VisionTransformer、SwinTransformer、DETR以及Segformre等
2. 关键点检测算子开发：Keypoint_RCNN, Keypoint_CID；各个算子都支持单机和分布式训练
3. 逐层量化的模块的开发：主要针对目标检测（Yolo）和语义分割（Unet）模型以实现推理加速
4. 结合主动学习+图像分类相关技术开发图像分类数据的智能标注工具
5. 使用SAM-HQ模型优化分割的智能标注工具，支持使用点、矩形框和掩码作为辅助信息来提升标注质量
6. 开发数据增强模块，支持对批量图像做串行或者并行处理，增强方式包括仿射变换、马赛克、调节色彩等

医学药品名称标准化

项目描述

由于各个医生的开药习惯不同，同一个药品在医院数据库中存储的名字各不相同，这就导致了这些数据库中的数据难以被查询分析以及复用；该项目的主要目的就是将各个药品的非标准名转换为标准名称，实现医院数据库的数据治理

个人工作

1. 给到的数据量是三十多万条，对应的标准名称是三千多个，看似是一个分类问题，但是类别数目过多做分类不好做，并且在每次新增标准药品名称的时候可能都需要重新训练，维护成本高
2. 参考人脸识别的思路，定性为一个特征比对问题，只需要训练一个特征提取器
3. 具体实现是：一维卷积+TransformerEncoderLayer+MLP，使用一维卷积减小词向量维度以减小特征向量在Transformer层的计算量，借助自注意力机制可以在训练时捕获更多的上下文信息
4. 损失选择用ArcLoss，优化特征向量和类别向量之间的夹角，并通过在夹角中添加一个角边距m，使得同一个类别的特征向量和类别向量之间的夹角更小，自然不同类别向量之间的夹角也就变大了，从而增加模型对于各个类别输入数据之间的区分度
5. 模型训练好后就会将标准药品名的特征向量录入数据库作为标准，通过比较相似度来映射各个非标药品名

工业芯片缺陷检测

项目描述

一家芯片厂商需要做自动检测流水线，其中一个需求是要实现芯片缺陷检测，并根据缺陷等级来实现芯片样本的分类分拣操作

个人工作

1. 缺陷的主要类型是划痕和油污，形状不规则，当即定性为一个分割问题
2. 给到的图片只有一百多张，有点少，采取了三种方案实现数据增样的效果，分别是对缺陷图片做仿射变换、缺陷迁移（ps、opencv）以及使用瓷砖缺陷数据集做预训练
3. 选择的分割模型是u2netp，分割效果与u2net相当，但是模型权重较小，利于后续的实时检测
4. 而缺陷分级可以通过opencv来实现，借助分割得到的缺陷二值图来做缺陷提取以及面积计算等
5. 由于前景背景的像素数量差异过大，存在样本不均衡问题，因此损失函数采用的是交叉熵+FocalLoss，再去做训练和评估等