孙爱华

男 | 年龄：21岁 | 15396897213 | aeeeeeep@proton.me

求职意向：算法工程师 | 期望城市：杭州

个人优势

1.熟悉大规模网络分布式训练框架源码及其二次开发,如 DDP,deepspeed,Megatron 等

2.掌握 GPU 异构编程,显存优化,性能调优

3.熟悉各类数据的处理与分析,并对特征有敏锐的洞察力

4.在个人服务器上管理运维博客: https://aeeeeeep.top 并在公众号 CV 技术指南 发布多篇教程

教育经历

盐城师范学院 本科 数字媒体技术 2020-2024

机器视觉实验室负责人 2021.06~至今

1.参与科研项目;

2.比赛指导带队;

3.处理招新,成员,以及日常事务。

竞赛经历

本科阶段 2020.09-2024.06

2022 微信大数据挑战赛 多模态短视频分类 全国三等奖

(参赛代码:https://github.com/aeeeeeep/2022WBDC-semi)

2022 第十八届挑战杯揭榜挂帅专项赛 RCS数据目标识别技术研究 Rank 24

kaggle - LLM Science Exam TOP 13%

kaggle: RSNA 2022 Cervical Spine Fracture Detection TOP 24%

kaggle: DFL - Bundesliga Data Shootout TOP 37%

2023 第四届全国人工智能大赛 AI+视觉特征编码赛道 Rank 30

2023 全球人工智能技术创新挑战赛 医学影像诊断报告生成 Rank 123

(参赛代码:https://github.com/aeeeeeep/2023GAIIC)

2022 MathorCup 高校数学建模挑战赛 全国二等奖

2022 五一数学建模竞赛 全国三等奖

2022 全国大学生新媒体大赛（摄影组）全国二等奖

2021 APMCM 亚太地区大学生数学建模竞赛 全国二等奖

(参赛代码:https://aeeeeeep.top/2021/11/29/2021亚太数学建模竞赛A题)

实习经历

上海壁仞智能科技有限公司 分布式训练实习生 2023.07-2024.01

内容:

基于壁仞生态,针对 NLP/CV /音视频/多模态/推广搜等场景,构建大规模的分布式机器学习系统;负责解决业务交付流程中遇到的单机多卡,多机多卡的精度、性能问题,研究行业领先的超大规模分布式策略。

实习成果:

● 分布式策略优化

如通过在 deepspeed 分区 tensor h2d 后异步 reorder ,并修改计算流程减少计算量,节省单 step 通信同步等待耗时,加速 6%;消除 dataload 时 broadcast ,单 step 加速 3%。

● deepspeed 模拟器性能守护

解析模型训练模拟器日志,实现自适应阈值守护和邮件警报功能,并在云服务器上可视化相关时间和带宽的历史折线图。

● pytorch profiler 解析工具开发

解析 pytorch profile 性能数据,计算 bubble 耗时,算子带宽与真实耗时,各模块耗时比率,峰值显存等,大大推动部门算子优化和显存优化效率。

项目经历

基于多模态特征的大黄蜂入侵物种防治算法　　算法开发　　2023.03-2023.05

内容:

大黄蜂作为一种入侵性物种，对生态系统和农业产生了严重的影响。为了有效防治大黄蜂的入侵，本项目提出了一种基于多模态特征的防治算法。通过结合图像识别、用户文本描述和上报地点等多模态信息，实现对大黄蜂的防治与繁殖预测，为后续的防治工作提供可靠的依据。

业绩:

对于图像模态,我们使用 Convnext 模型,对收集到的大黄蜂图像进行了训练和分类。识别大黄蜂的存在与否。对于文本特征,使用 Fasttext 对用户的文本描述进行情感分析和关键词提取,再编码识别。最后使用熵权法研究不同特征对大黄蜂繁殖的重要程度,进行加权判定,在测试集上达到了 99% 的准确率。

难点:样本分布极度不平衡,通过重采样,对抗训练,提示词生成稀少类别文本样本缓解。采用 Bert 模型融合方案虽然可以更好融合特征,但本项目文本过短,有过拟合现象,遂采用决策融合方案。

电力生产环境员工安检系统　　算法开发　　2022.12-2023.03

内容:

基于现场监控摄像头视频进行智能分析,作业前,进行人员着装与绝缘防护判断,对作业的配电柜与设备初始状态,做时态信息和票面信息对比研判,对错误的配电柜和错误初始状态及时告警。

业绩:

使用 Yolo 目标检测网络进行实时着装检测，使用 OpenCV 传统视觉方法对设备状态进行检测。并对算法在 Nvidia 环境下的量化， k8s 集群上的部署，实现 50 fps(96.8 acc) 的检测速度。设备检测算法实现 500 fps (99.5 acc) 的检测速度。

难点:Yolo 检测候选框重叠,使用 NMS 合并。视频传输占用带宽过高,且数据库压力过大,使用跳帧检测,并对检测结果编码压缩。

使用技术：CUDA, Tensorrt, nvidia-docker, OpenCV, Pytorch

毕业设计

基于自动并行策略的轻量化深度学习框架研究(进行中)　　2024.02-2024.05

1.选题依据 近年来,随着模型参数量的显著增大和预训练语料库的扩展,单个模型的训练规模不断扩大,伴随而来的是成本的显著增 加。特别是自2020年以来,大型语言模型的兴起已经将自然语言处理、计算机视觉和跨模态领域推向了一个超大规模的技术 竞赛。然而,现有的硬件设计和分布式通信的局限性导致了在这些集群中整体利用率的低下。因此,提高大规模分布式训练 的效能是一个具有重要意义的选题。

2.现存问题 在不同规模参数的模型配置下,目前的主流框架做到单卡有效算力 135TFLOPS~163TFLOPS,有效率用率 43%-52%。直 观上来看,相对于Nvidia A100 312TFLOPS FP16的理论性能,还是有比较大的上升空间。造成这种这种现象的原因主要 是训练框架DeepSpeed[3]并行策略依赖人工反复尝试来进行调度,导致模型并行调度中仍有很多串行计算。目前虽然已有 实现自动并行的深度学习框架,如Alpa[4]结合动态规划和整数线性规划切分模型并制定最优并行策略,但结合ZeRO- Offload[5]和Infinity[6]提出的利用host资源,ZeRO++[7]针对通信的优化,关于并行策略的自动调度仍有优化空间。

3.实施方案 实现自动微分,张量操作,网络构建,数据加载与预处理,优化器,Loss函数,模型评估等模块; 编写单元测试,守护精度性能,跑通主流模型。自动并行:将各个训练stage,调度策略,切分Tensor抽象,便于调度;基 于已知工作,将更多优化融入并行策略中,自动调度生成执行任务图。