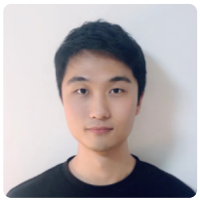
董微



男 | 年龄：28岁 | 15207499244 | 279617552@qq.com

4年工作经验 | 求职意向：深度学习 | 期望薪资：25-35K | 期望城市：上海

个人优势

通用技能：Jira Linux Shell Jenkins Git Docker make CMAKE Bazel vscode

语言和库：C++ CUDA Ncu Nsys Python Caffe Pytorch Tensorflow Onnx Gtest Pytest

算子开发，性能优化，测试经验。

工作认真，学习能力强，善于沟通。

工作经历

上海寒武纪信息科技有限公司 C/C++ 2019.04-2023.04

算子开发和性能优化

1. 使用性能分析工具分析算子性能瓶颈，写性能优化方案

2. 调优算子性能。

网络精度和性能基线框架开发

1. 推理引擎精度性能基线框架的方案设计、实现和持续优化。

2. resnet bert 等的网络的精度、性能测试。

项目经历

Light Net 软件开发 **2023.07-至今**

内容:

练习项目

从零开始实现的的深度学习网络框架。支持推理和训练。

学习和理解深度学习框架和算子的工程实现和优化方法。

项目连接 https://github.com/longer_is_better/lightnet

参考资料：

《NVIDIA CUDA》

《cuda c programming guide》

《CUDA_C_Best_Practices_Guide》

《机器学习系统：设计和实现》...

业绩:

目前已完成

Tensor ，计算图，部分 kernel ，部分算子，部分 loss 以及部分辅助工具的实现和模块测试。

部分 kernel 的性能优化，优化后性能与 cublas 和 cudnn 相近。

验证了多层感知机的正确推理和训练。

算子开发　软件开发　**2022.11-2023.05**

用性能分析工具分析算子性能瓶颈，写性能能优化方案，调优算子性能。

gemm，reduce，softmax

等算子的性能优化，精度性能测试。

设计实现 Fixture 、标准流检查宏等通用方法

Value-parameterized、Type-parameterized 测试

网络精度和性能基线　软件开发　**2020.08-2022.10**

内容:

将开源或客户的深度学习网络移植到 benchamrk 框架

包括数据集处理，前后处理，backend 添加，精度得分计算等适配工作

bert RCNN 等网络的精度跑分、数据前后处理代码，性能测试

络精度性能框架方案设计、框架开发、维护和优化

- 通用

领域：视觉、自然语言、语音等

设备：MLU、CPU、GPU

平台：云侧、边缘侧

- 高效

编译并行：离线模型编译任务可按需求，环境能力等动态配置并行度

测试并行：基于设备管理实现合理并行

设备管理：根据任务类型分配设备保证设备的高效利用

业绩:

提升代码复用率，提升执行效率，提高资源利用率

1. 并行化边缘侧用例执行逻辑，做到编译推理完全并发，提升执行效率 4 倍
2. 重构边缘侧流水执行逻辑，极大提高代码复用率，降低维护成本，提高可读性
3. 设计并实现多功能执行流水，既可以用于每日精度性能监测，也可以用于项目评测任务
4. 实现精度和性能 baseline 工具并整合入流水，支持 baseline 的快速检查，保存，更新，维护等，收录 10万+ 条 baseline 数据
5. 分析云上资源的浪费问题，并释放 50% 服务器资源

教育经历

中国人民解放军国防科技大学　硕士　系统论证与仿真评估　2016-2019

课程：离散系统仿真，数据挖掘，模式识别等

2019年天池。航班调度算法比赛 17 名

协助导师对接客户，提供驻场支持

仿真系统并行化加速

中南大学　本科　探测制导与控制技术　2012-2016

专业排名前三

课程：PLA，单片机，数模电等

长沙·国家艺术基金资助作品展志愿者

穿越机爱好者协会竞速比赛第 3 名