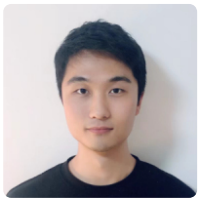
董微



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

4年工作经验 | 求职意向：算子开发 | 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 (C)Make Bazel Vscode ...

语言和库：CUDA-C++ cublas cudnn Ncu Nsys Python Pytorch Onnx G(Py)test ...

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++ 2019.04-2023.05

算子性能优化

1. 使用性能分析工具分析算子性能瓶颈
2. 给出优化方案和预期性能。
3. 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、实现和优化。
2. resnet bert rcnn 等10余网络的精度、性能测试用例。

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

2019年天池 - 航班调度算法比赛 17 名、能仿真系统的 OpenMP 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、穿越机竞速比赛第 3 名。

项目经历

Light Net 软件开发 2023.07-至今

内容:

轻量的深度学习框架

支持并行推理和基于自动求导的模型训练。

借助 Nsight compute 分析 cuda kernel 的性能瓶颈。

采用 sharedmem，warp_shuffle，vectorized，double_buffer 等方法优化算子性能。

使用 Nsight system NVTX 分析 kernel launch 等过程的性能瓶颈。

通过异步数据传输，多 stream 配置，算子融合等方法提高网络性能。

业绩:

Tensor, compute_graph, kernel, layer, loss 等模块的开发和单元测试。

部分 kernel 的优化性能与 cublas 和 cudnn 相当

kernel: Reduce Map EleWise MatMul Transpose ...

layer: Softmax Linear Relu l1loss ...

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

内容:

算子性能优化：

使用 cnperf 性能分析工具定位算子性能瓶颈，

给出性能优化方案

优化算子性能。

算子的精度性能测试用例补全。

业绩:

conv 算子性能优化：

channel 对齐优化，提升性能约 16 倍

优化指令流水掩藏延迟，提升性能约 2 倍

负载均衡优化，提升性能约 50%

性能精度测试：

补全 softmax，Batchnorm，GEMM，permute 等算子测试。

Benchamrk　　软件开发　　2020.08-2022.10

内容:

功能对标 https://github.com/mlcommons/inference

网络精度性能 benchmark 框架方案设计、开发、维护和优化

移植开源或客户的模型络到 benchamrk 框架

Bert RCNN 等网络的精度跑分用例移植（数据前后处理，backend，metrics）

- 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 工具支持快速检查，保存，更新，维护，收录 10万+ 条 baseline 数据