陶然

25岁 | 男 | 汉
18896792735 | 1045204687@qq.com

教育背景

2022-09 ~ 至今 **东南大学** **人工智能（硕士）**
专业排名： 2/49 **主修课程：** 自然语言处理、深度学习、计算机视觉、算法设计、模式识别等等。

2017-09 ~ 2021-06 **宁波大学** **数学与应用数学（本科）**
专业排名： 4/36 **主修课程：** 随机过程、离散数学、精算数学、复利数学等等。

实习经历

2024-01 ~ 至今 **英特尔** **ai软件开发实习生**
部门的主要业务是开发维护ipex的代码，支持pytorch计算llm等大模型在cpu服务器上的高性能运转。开源的地址如下
GitHub - intel/intel-extension-for-pytorch: A Python package for extending the official PyTorch that can easily obtain performance on Intel platform
1.主要负责weight only quant模块的gemm kernel开发，实现了dequant weight upfront的功能，在精度测试结果不变的情况下，在gpt-j-6b上测试latency减少了5%，目前已经提pr，预计下一版ipex会上线。
2.对原来的代码进行移植优化，提升代码复用率，提高代码的可读性，将原先cpu_devkit代码仓库的代码移植到ipex仓库，减少开发上线成本。

2023-09 ~ 2024-01 **大疆车载高性能部门** **智能驾驶-高性能计算实习生**
部门的主要业务是基于感知算法，设计出一套高性能的软件框架，使得自动驾驶算法可以更高效的运行。主要的工作内容有在链路上的联调，整体的loading、内存优化，业务代码的开发和优化等。
1. 参与自动驾驶感知算法和软件在嵌入式平台上的链路联调，移植优化，系统集成工作，完成了freespace模块在高通、vh平台地形识别、路面预瞄等链路的联调，功能按期交付上线，同时发布老化报告，系统loading无明显上涨。
2. 负责软件单元测试环境维护，基于软件详细设计，完成软件单元测试自动化脚本的设计与开发，完成了GoogleTest的用例编写累计5000余行。
3. 负责感知软件AutoSar的coverity修复，对代码的实现细节进行重构，累计修复漏洞3000余条，增强代码的安全性规范性。
4. 负责功能安全、故障注入部分的功能开发，实现了12条故障注入用例的编写和链路验证。

2023-05 ~ 2023-08 **中国航天科技集团** **高性能计算实习生**
部门主要的项目是为卫星数据传输设计一套加密算法，并使用GPU并行算法为其加速，拟采用国密sm4算法CTR模式并加入一些秘籍进行加密，目标是在六张卡的运行下达到40Gb/s的吞吐量。
1.协助完成对sm4加密算法的gpu实现,完成把多个密文帧发给GPU,对多个密文帧同时进行解密，产生对应明文帧,并回传给CPU的一个完整的过程。
2.通过性能分析得知,数据在从CPU转向GPU的过程中消耗了大量的时间,尝试过锁页内存，共享内存、寄存器分块、数据预取的做法对此进行优化。
3.进行对算法的加速优化,通过若干优化的方法，如分解rkey为四个小数组,输出密文的帧头帧尾的数据赋值拷贝移至程序开头CPU端操作等，使得吞吐量变为原来的240%。

项目经验

GEMM优化
项目介绍： 矩阵乘（GEMM）的优化是一个非常重要的课题。这些年涌现了一系列的深度学习模型。模型里面最耗时的东西，包括卷积、全连接层、attention，都可以转换成GEMM操作，出于学习的目的，实现了GEMM在cuda上的优化。
解决的问题： 通过向量化访问、数据预取、共享内存、消除bank conflict等为优化，性能可以达到cuBLAS 70%的水平。

实现dropout kernel
项目介绍： Dropout可以作为训练深度神经网络的一种trick供选择。在每个训练批次中，通过忽略一半的特征检测器（让一半的隐层节点值为0），可以明显地减少过拟合现象。
解决的问题： 通过算子融合两个kernel fuse成一个大的kernel以及向量化访问为优化，由此dropout性能可以达到和PyTorch-CUDA dropout的性能相当的情况。

专业技能

- 熟练掌握数据结构和算法、操作系统、计算机组成原理
- 熟悉C++、Java、Python，拥有良好的编程习惯，熟悉 Linux 开发环境
- 了解 Nvidia GPU/ARM CPU 处理器体系结构，理解软件与硬件底层映射关系
- 熟悉计算机体系结构、并行计算基本技术以及 GPU 并行计算基本原理