刘学建

18844118687 | iamsonderr@outlook.com | 湖南永州
22岁 | 男
研究方向：高性能计算、编译优化、推理加速

教育经历

湖南大学 2021年09月 - 2024年06月
计算机科学与技术 硕士 超算与人工智能融合计算教育部重点实验室 湖南长沙

- 导师：全哲(湖南大学), 彭林(国防科技大学)
- **保送硕士研究生**

吉林大学 2017年09月 - 2021年06月
计算机科学与技术 本科 吉林长春

实习经历

百度 2023年06月 - 2023年09月
AI异构计算实习生
在百度深度学习平台PaddlePaddle的推理引擎Paddle Inference中进行模型部署相关的工作:

- 在Paddle Inference中编写**transfomer**中的**多头注意力机制融合算子**(FMHA)。涉及到的技术包括：**KV Cache**缓存K和V的中间结果，减少重计算的次数；**WeightOnly**在校准数据的基础上将Q、K、V**量化**成低精度进行存储和访存以降低访存压力，并**反量化**成高精度进行计算；**Back2Back GEMM**将多个GEMM融合成单个GEMM，实现Attention复用；根据体系结构设计**GEMM分块**参数，调整warp尺寸**消除shared memory bank冲突**，并基于**双缓冲**数据搬运技巧，减少数据搬运开销，进行高效GEMM优化
- 为Paddle Inference中的op编写转换至**TensorRT Layer**的converter，完善Paddle Inference对TensorRT的支持范围
- 编写收集模型推理时的**tensor shape**的pass，并将收集到的tensor shape应用于**显存复用算法**提升复用有效性和效率

澎峰科技（OpenBLAS） 2022年08月 - 2022年09月
高性能计算实习生
在**OpenBLAS发起人张先轶博士**的指导下，对多个开源的**稀疏矩阵计算库**（XTensor、Eigen、blaze、fastor和armadillo）进行了多方位对比和性能测试，形成了从**数据类型支持、向量化并行化和编译时优化**的调研报告

项目经历

面向国产处理器的多编译融合优化技术研究 2022年12月 - 2023年06月
小组负责人

- 项目描述：本项目针对现有国产化处理器平台应用程序**移植后运行性能低下**等问题，将单个程序在**源码级**以**函数或循环粒度**进行分块，并为每个程序块选择最优编译器进行融合优化，以充分结合**多编译器**的优势优化应用程序
- 工作内容：
  - 利用**rose编译器基础设施**编写对应用程序中的循环和函数进行提取形成单独的代码模块的工具
  - 使用python编写对提取出的各个代码模块进行**多编译性能分析和融合优化**的脚本
- 工作成果：实现了集成循环或函数提取、性能分析和优化的**多编译融合优化框架**，所选的编译优化领域典型benchmark在国产处理器平台上**相对于其他通用编译器平均13.4%的运行效率提升**

面向国产处理器的混合精度编译优化工具链设计与开发 2022年09月 - 2022年12月
小组负责人

- 项目描述：本项目针对**高精度浮点科学计算**应用程序**计算效率低、带宽占用高**等问题，利用**混合精度**调优技术优化应用程序用，进行程序精度和性能调优
- 工作内容：
  - 基于**LLVM编译基础设施**编写在IR层生成**精度搜索空间、精度自动转换的pass**
  - 使用python脚本实现**精度配置搜索算法**
- 工作成果：实现了集成生成精度搜索空间、精度配置搜索算法、精度自动转换、误差分析的**混合精度编译优化工具链**，多个基于国产智能处理器的数值计算应用性能经过混合精度编译优化工具链调优后相比通用编译器编译具有较大性能提升

科研经历

- (**专利**)基于深度强化学习的医疗图片分类方法、系统及电子设备.CN115761336A
- (**专利**)一种求解奇异值的精度可调节的多模式计算框架及方法.CN114861129A
- (**论文在投中**)Multi-Compilation Fusion Optimization Framework Using ROSE Compiler Infrastructure.刘学建,全哲,李磊,杜云飞,彭林.APLAS2023

荣誉奖项

获得湖南大学研究生学业一等奖学金2次，吉林大学学业一等奖学金、二等奖学金、院优秀学生各1次
吉林大学大学生数学建模竞赛一等奖 2019-10
全国大学生数学建模竞赛吉林赛区一等奖 2019-11

专业技能

- 熟悉LLVM和Rose编译器基础架构, 掌握LLVM中的Pass编写
- 熟练掌握C、C++, 熟悉Python、Shell等脚本语言
- 熟悉Linux环境, 会使用Git、Makefile管理项目
- 英语已过CET-6, 能够熟练阅读英文手册、文献