本科毕业设计（论文）任务书

课题名称：基于 RISC-V 存算一体芯片的编译器关键技术研究

学 员 姓 名：	简泽鑫	学 号：	202102001019
首次任职专业：	无	学历教育专业：	计算机科学与技术(计算机系统方向)
命 题 学 院：	计算机学院	年 级：	2021 级
指 导 教 员：	曾坤	职 称：	副研究员
所 属 单 位：	计算机学院微电子与微处理器研究所		

国防科技大学教育训练部制

一、 课题主要任务

本课题将面向 RISC-V SRAM 存算芯片修改 LLVM 编译器，实现对存算指令的支持。通过深度研究 LLVM 编译器架构和工作原理，分析 RISC-V SRAM 存算一体芯片的特性，探索如何在 LLVM 中添加对 RISC-V CIM 的支持。这包括但不限于对指令集的扩展、内存模型的适配、NPU 指令的智能识、指令调度别以及优化策略的调整等，以实现应用算子的自动映射和正确指令流的生成，从而更好地协调计算部件，挖掘芯片内部的计算并行性，为 RISC-V 存算一体芯片提供有利的编译支持。

二、 课题要求

(一) 理论研究

1. 研究现有的深度学学习编译器、深度学习加速器以及 LLVM 编译器。
2. 深入研究 LLVM 编译器架构和工作原理，分析 RISC-V SRAM 存算一体芯片的特性，探索在 LLVM 中添加对 RISC-V CIM 的支持。

(二) 编译器扩展

1. 基于理论研究，扩展 LLVM 对 RISC-V CIM 的支持。这包括指令集的扩展、内存模型的适配、NPU 指令的智能识别、指令调度别以及优化策略的调整等，需明确其核心思想、预期效果以及可能的创新点。
2. 考虑底层 RISC-V CIM 硬件架构，确保其能够应用 NPU 加速核来加速计算，同时确保编译器的功能性正确。

(三) 代码实现

1. 使用 C/C++在软件层面实现 NPU 指令的智能识别、指令调度以及后端指令集扩展。代码应当具有良好的可读性和可维护性。
2. 对编译器进行初步测试，选取预训练神经网络模型进行优化，验证其功能性。
3. 记录并分析测试结果，包括 NPU 利用率、NPU 能耗等指标，同时记录深度学习常见算子利用 NPU 和不利用 NPU 的加速比。

（四） 编译器实现与测试

1. 基于 FASHION MNIST 数据集完成自定义网络模型的推理任务，对编译器的功能性进行测试。
2. 对于深度学习中常见的算子在利用 NPU 和不利用 NPU 的情况下，对编译器的性能进行测试。
3. 记录并评估对应的测试数据，包括 NPU 利用率、NPU 能耗等指标。

（五） 文档撰写

1. 撰写详细的毕业设计报告，内容包括课题背景、理论研究、LLVM 编译器后端扩展、NPU 指令的智能识别、指令调度、编译器测试结果、分析与讨论等。
2. 报告应该图文并茂，内附有清晰的编译器架构图、测试数据表、分析图表等。

三、 完成形式

C++实现代码：完成 C/C++源代码，辅以必要的说明文档。

毕业设计报告：提交一份格式规范、内容完整的毕业设计报告文档。报告应当详细记录课题的研究过程、编译器设计思路、实现细节、测试结果与分析。

毕业设计答辩：在答辩会上，就课题的研究内容、实现细节、创新点、测试结果等方面进行口头汇报，并回答评审老师问题。

四、 进度安排

序号	各阶段内容	时间安排
1	调研 AI 编译器、存算一体芯片等相关理论和研究现状，形成调研报告	2024.11.01-2024.12.20
2	配置 LLVM 开发环境，熟悉 RISC-V 开发工具链，并安装必要的模拟器和调试工具，撰写本科毕业设计开题报告	2024.12.21-2025.01.19

3	分析 LLVM 编译器架构同时进行扩展，并进行内存模型适配，准备中期检查相关材料	2025.01.20-2025.03.30
4	调整编译器优化策略并进行模拟器继承与测试，验证编译器的正确性和性能提升	2025.04.01-2025.04.20
5	梳理总结研究成果，撰写毕业设计报告，准备毕业论文答辩工作	2025.04.21-2025.06.10

参考文献

[1] Tianqi Chen,Thierry Moreau,Ziheng Jiang,et al. TVM: An Automated End-to-End Optimizing Compiler for Deep Learning[J]. 2018,.

[2] Tillet, Philippe,Kung,et al. Triton: an intermediate language and compiler for tiled neural network computations[C]//MAPL 2019: Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages. 2019.

[3] Daniel Snider,Ruofan Liang. Operator Fusion in XLA: Analysis and Evaluation[J]. 2023,.

[4] Zhao, Jie,Li,et al. AKG: automatic kernel generation for neural processing units using polyhedral transformations[C]//PLDI 2021: Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation. 2021.

[5] Jared Roesch,Steven Lyubomirsky,Logan Weber,et al. Relay: A New IR for Machine Learning Frameworks[J]. 2018,.

[6] Siyuan Feng,Bohan Hou,Hongyi Jin,et al. TensorIR: An Abstraction for Automatic Tensorized Program Optimization[J]. 2022,.

教研室（研究室、实验室）意见： 领导签名：　　年　　月　　日
系（研究所、重点实验室）意见： 领导签名：　　年　　月　　日
学院教学科研处（教务处）意见： （公　章） 年　　月　　日

注：任务书由指导教员填写，经各级审核后下达给学员，答辩结束后由指导教员交系（所、室）。