智能计算体系结构课程lab6实验报告

作者：北京航空航天大学计算机学院 陈胤佳

学号：19373383

一、实验目的与要求：

本次实验目的为，使用NativeTPU实现MLP、LeNet神经网络推理。

二、实验环境：

IDE：Vivado 2019.2 / Vitis 2019.2；

Environment：ZYNQ 7020开发板及其配件；

Environment：ZYNQ 7020上的Linux系统；

OS：Windows 10 Professional

三、实验内容与步骤

1. 问题回答：

对比以下两个方案的执行时间，对比执行时间：

1. 在ARM上使用Numpy运行矩阵乘法，执行神经网络
2. 使用FPGA运行矩阵乘法，执行神经网络

请问方案a和b哪一个执行时间更快？为什么？（思路、过程）

答：从硬件结构层面分析，ARM属于控制集中型硬件，FPGA属于计算机集中型硬件，采用ARM执行神经网络时，神经网络密集的矩阵计算会耗费大量的有限的计算资源，从而降低ARM所在系统的整体性能；而利用FPGA的可编程性，将矩阵计算移植到FPGA上进行时，即可以最大程度上利用FPGA大量空闲的计算资源，效率远高于ARM上的执行效果。采用课程组所提供的系统交互方式，即ARM侧处理指令的执行、FPGA侧处理高密度矩阵计算、两侧采用AXI4LITE协议通信，达成硬件优势的最大利用。

四、实验结果

环境配置、测试成功。

五、分析与讨论

本次实验综合了前几次lab，完整实现了课程要求的智能计算体系结构，对加速神经网络的硬件交互逻辑有了详实的理解。