基于TTA的大型卷积神经网络处理器架构设计

An processor architecture design for Large-scale CNNs based on TTA

摘要

一、绪论 ----4000

1.1 课题背景与研究目的

1.2 国内外现状

1.3 论文主要工作

1.4 论文组织安排

二、相关技术研究 ----8000

2.1 ASIC与ASIP设计

2.1.1 ASIC设计

2.1.2 ASIP设计

2.2 TTA架构及TCE工具链

2.2.1 TTA架构

2.2.2 TCE工具链

2.2.3 设计流程与功能单元设计

2.3 卷积神经网络

2.3.1 卷积层

2.3.2 池化层

2.3.3 激活函数层

2.3.4 全连接层

2.4 本章小结

三、大规模卷积神经网络计算框架设计 ---- 4000

3.1 整体计算框架

3.2 片内存储优化方案

3.2.1 输入优化结构

3.3.2 输出优化结构

3.3 本章小结

四、卷积神经网络功能单元设计 ----8000

4.1 DDR功能单元设计

42 输入缓存功能单元设计

4.3 通用计算单元设计

3.2.1 卷积运算设计

3.2.1.1 Z型卷积计算单元

3.21.2.1 旋转存储卷积计算单元

3.2.3.1 偏置处理

3.2.2 池化模块计算单元

3.2.3 激活函数计算单元

3.2.4 全连接层计算单元

3.4输出缓存功能单元设计

3.5 本章小结

五、卷积神经网络互联网络设计 ---- 2000字

5.1 VLIW互联网络

5.2 全连接型互联网络

5.3 自定义互联网络

5.4 本章小结

六、卷积神经网络实现及性能分析 ---- 2000字

6.1 实现平台及开发环境

6.2 功能单元及互联网络实现

6.3 资源使用与性能分析

七、总结与展望 ---- 3000字

7.1 工作总结

7.2 不足与展望

7.3 本章小结

创新点：

1. 基于TTA架构实现了卷积神经网络的可配置方案。相比ASIC方式，大大提升了灵活性，在实际应用中大大减少开发成本。

2. 针对于大型卷积神经网络，提出了计算框架与优化方案，使大型卷积神经网络能够在小型FPGA上进行加速。

3. 提出了旋转存储卷积计算单元，相比传统Z型卷积计算单元，在卷积运算步长大于1时，可大幅度减少运算时间。

4. 提出了一种通用计算单元，可同时适用于卷积层、池化层、全连接层。减少了计算资源的使用。