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 1500

1.2 国内外现状与研究目的 2000

1.3 论文主要工作 300

1.4 论文组织安排 200

二、相关技术研究 ----8000

2.1 卷积神经网络介绍 1500

2.2 卷积操作并行性研究 2500

2.2.1 卷积窗口内并行

2.2.2 输入特征图并行

2.2.3 输出特征图并行

2.24 并行性组合

2.3. TTA及TCE开发工具集 4000

2.3.1 TTA架构介绍

2.3.2 TCE工具集介绍

2.3.3 TCE设计流程

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 计算框架设计 2000

3.2 片内存储优化方案 2000

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、基于TTA架构的卷积神经网络功能单元及互联网络设计 ----8000

4.1 DDR功能单元设计 700

4.2 输入缓存功能单元设计 2000

4.2.1 旋转存储结构

4.2.2 顺序存储结构

4.3 通用计算功能单元设计 3000

3.2.1 卷积运算操作设计

3.2.2 池化运算操作设计

3.2.3 激活函数操作设计

3.2.4 批量归一化操作设计

3.2.5 全连接层操作设计

4.4 输出缓存功能单元设计 700

4.5 互联网络设计 1500

4.5 本章小结 200

五、卷积神经网络实现及性能分析 ---- 5000

5.1 自定义卷积神经网络实现 3500

5.1.1 卷积层实现

5.1.2 池化层实现

5.1.3 归一化层实现

5.1.4 激活函数层实现

5.1.5 全连接层实现

5.2 资源使用与性能分析 1500

5.3 本章小结

六、总结与展望 ---- 2000

6.1 工作总结 1000

6.2 不足与展望 1000

6.3 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。