**密级： 保密期限：**

xm 拷贝

**硕士学位论文**



**题目：基于TTA的大型卷积神经网络处理器架构设计**

**学 号： 2016110107**

**姓 名： 张大旺**

**专 业： 信息与通信工程**

**导 师： 别志松**

**学 院： 信息与通信工程学院**

**2018年 12 月 7 日**



**A Thesis for Master Degree**

**TITLE: AN PROCESSOR ARCHITECTURE DESIGN FOR LARGE-SCALE CNNS BASED ON TTA**

|  |  |
| --- | --- |
| **Student No.:** | **2016110107** |
| **Author:** | **Dawang Zhang** |
| **Major:** | **Information and Communi-**  **cation Engineering** |
| **Supervisor:** | **Zhisong Bie** |
| **School:** | **School of Information and Communication Engineering** |

**Dec. 7st, 2018**

独创性（或创新性）声明

本人声明所呈交的论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京邮电大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：日期：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

关于论文使用授权的说明

学位论文作者完全了解北京邮电大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属北京邮电大学。学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。（保密的学位论文在解密后遵守此规定）

保密论文注释：本学位论文属于保密在年解密后适用本授权书。

非保密论文注释：本学位论文不属于保密范围，适用本授权书。

本人签名：日期：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

导师签名：日期：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

基于TTA的大型卷积神经网络处理器架构设计

**摘 要**

卷积神经网络(Convolutional Neural Network, CNN)，是神经网络的一种。由于卷积神经网络具有权值共享以及局部连接的特性，使得卷积神经网络的模型复杂度与参数数量大幅度降低。近年来，卷积神经网络发展迅速，在图像处理以及自然语言处理领域都有着广泛的应用。

卷积神经网络作为实现人工智能任务的有效算法之一，已经在各种应用场景获得广泛的应用。从云端到移动端，不同应用场景也对平台的计算能力提出了不同的需求。目前移动端的主流加速平台主要包括通用处理器以及专用集成电路 （ASIC, Application Specific Integrated Circuit）两种。前者虽然灵活性较好，但是在对实时性或者功耗要求较高的场合并不适合。后者对于某一种算法或者网络，可以在功耗和性能上可以做到最佳，但由于灵活性很差，使得这种实现方式的设计成本和功耗增加，设计周期变长。

专用指令集处理器（ASIP, Application Specific Instruction Set Processor）针对某一应用领域进行裁剪和优化。相比于通用处理器，牺牲了一定的灵活性，但使得计算效率大大增加。相比于ASIC设计方式，其牺牲了一定的性能，但在灵活性上大大增加，从而在性能与灵活性中达到了一个平衡点。传输触发架构(TTA, Transport Triggered Architecture)为ASIP中的一种架构，其核心思想是利用数据传输来触发相应功能单元的具体操作。TTA架构将寄存器单元也作为一种特殊的基本单元，它有效地减少了寄存器堆的设计压力，成为一种非常适合于专用处理器领域的处理器架构。

本文基于TTA架构，首先针对于FPGA平台片内存储不足的情况，提出了两种存储优化的方案，使大型卷积神经网络能够在较低成本的FPGA上进行加速。其次，针对于不同运算层的特性，分别设计了不同的计算单元，使加速效果最大化。同时，卷积运算单元与全连接层运算单元共用乘法器，减少了FPGA内DSP资源的使用。最后，针对于不同功能单元之间的数据互通，设计了数据传输互联网络，使处理器的复杂度进一步的降低。

**关键词：**卷积神经网络 传输触发架构 专用指令集处理器 功能单元 FPGA

**AN PROCESSOR ARCHITECTURE DESIGN FOR LARGE-SCALE CNNS BASED ON TTA**

**ABSTRACT**

Convolutional Neural Network(CNN) is a well-known deep learning architecture which extends from artificial neural network. Because of the features of weight sharing and local connectivity, the complexity of the model is greatly reduced. In recent years, CNN has developed rapidly and has been widely used in computer vision and natural language processing.

As one of the effective algorithms for realizing artificial intelligence tasks, convolutional neural networks have been widely used in various application scenarios. From the cloud to the mobile, different scenarios also put different demands on the computing power of the platform. At present, the mainstream acceleration platform of the mobile terminal mainly includes the general-purpose processors and application specific integrated circuit (ASIC). Although the former has better flexibility, it is not suitable for occasions with high real-time or power consumption requirements. The latter can achieve the best power consumption and performance for an algorithm or network, but due to the poor flexibility, the design cost and power consumption of this implementation increase, and the design cycle becomes longer.

The Application Specific Instruction Set Processor (ASIP) is tailored and optimized for an application domain. Compared to general-purpose processors, some flexibility is sacrificed, but the computational efficiency is greatly increased. Compared to the ASIC design, it sacrifices a certain performance, but the flexibility is greatly increased, thus achieving a balance between performance and flexibility. The Transport Triggered Architecture (TTA) is an architecture of ASIP. The core idea is to use data transmission to trigger the specific operation of the corresponding functional unit. The TTA architecture uses register units as a special basic unit, which effectively reduces the design pressure of the register file and becomes a processor architecture that is well suited for the field of dedicated processors.

Based on the TTA architecture, this paper first proposes two storage optimization schemes for the insufficiency of on-chip storage of FPGA platforms, enabling large convolutional neural networks to be accelerated on lower cost FPGAs. Secondly, different computing units are designed according to the characteristics of different computing layers to maximize the acceleration effect. At the same time, the convolution operation unit shares the multiplier with the fully connected layer operation unit, which reduces the use of DSP resources in the FPGA. Finally, for the data intercommunication between different functional units, a data transmission interconnection network is designed, which further reduces the complexity of the processor.

**KEY WORDS:** Convolutional Neural Network TTA ASIP Function Unit FPGA

目录

[第一章 绪论 2](#_Toc531980922)

[1.1 课题背景与研究目的 2](#_Toc531980923)

[1.2 国内外现状 2](#_Toc531980924)

[1.3 论文主要工作 2](#_Toc531980925)

[1.4 论文组织安排 2](#_Toc531980926)

[第二章 相关技术研究 3](#_Toc531980927)

[2.1 卷积神经网络 3](#_Toc531980928)

[2.1.1 卷积神经网络前向传播网络结构 3](#_Toc531980929)

1. 绪论
   1. 课题背景与研究目的
   2. 国内外现状
   3. 论文主要工作
   4. 论文组织安排
2. 相关技术研究
   1. 卷积神经网络
      1. 卷积神经网络前向传播网络结构