**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

人工智能（AI, Artificial Intelligence），是通过计算机来达到扩展人的智能的一门科学，与空间技术和能源科技并称为世界的三大顶尖科技。近年来，人工智能领域取得了极大的进展与成就，使得人工智能成为最引人注目的热点领域。其也影响着诸如制造业、出行、教育行业、金融行业、新闻行业等一系列我国的产业应用，人工智能已成为推动我国发展的新动力[1]。深度学习作为人工智能领域中的一大分支，近几年在许多事件中都体现出了其强大的解决问题能力，具有很高的学术研究价值。

卷积神经网络（Convolutional Neutral Network，CNN），是深度学习领域中的一类神经网络模型。由于其具有权值共享以及局部连接的特性[2]，相比较传统深度神经网络，网络模型的复杂度以及参数数量都大大降低。该优点在网络的输入是多维图像时表现得更明显，使图像可以直接作为网络的输入, 避免了传统识别算法中复杂的特征提取和数据重建过程。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。

神经网络的硬件计算平台可以分为训练以及使用两个场景分别进行探讨。

相比神经网络的使用场景，神经网络的训练需要更大的计算量。在最初始阶段，学术界使用灵活性较强的CPU来进行神经网络的训练，由于单CPU计算性能较差，因此选择使用CPU集群来进行加速。谷歌曾在2012年使用1000台含有16核CPU的机器来训练模型，训练时间长达3天模型得以收敛。由于卷积神经网络的训练所需要的计算量过于庞大，学术界转而寻找其它硬件以代替CPU。由于GPU（Graphics Processing Unit）采用SIMD（Single Instruction Multiple Data），在矩阵运算中具有天然优势，因此学者们将点积运算转为矩阵运算，得以在GPU上进行网络的加速。GPU的流行，也促使了许多深度学习框架的诞生，如Caffe、Tensorflow、Torch等，使得深度学习的发展进一步加快。

针对于神经网络的使用场景，其一般都具有高实时性、低功耗的需求。传统通用处理器CPU以及ARM，虽然擅长于逻辑控制，但计算资源以及效率不足，无法满足实时性的特点。而GPU虽然具有大量的计算单元，但一般的GPU功耗都在百瓦以上，无法满足低功耗的需求。此外，GPU采用SIMD架构，在训练时可以将多幅图片同时进行计算来提升计算资源的使用效率，但在应用场景中，只需对一副图片进行计算，无法使用所有计算资源，使得GPU的计算资源的使用效率变得很低。

除了通用处理器，在应用场景中，最主流的设计方式为ASIC（Application Specific Integrated Circuit）。ASIC是一种为了专门目的而进行设计的集成电路，在设计者的能力较强的情况下，ASIC一般可以具有小体积、低功耗、高可靠性、低成本等优点，可以说是特定场景中最适用的硬件平台。但由于ASIC目的性过强，因此需要设计者完成所有电路设计，所需人力成本较高，开发周期较长。同时ASIC设计方式在算法层面上灵活性较差，在对算法有少量修改时，仍需重新设计。因此ASIC适用于算法确定、开发周期较长且有大规模需求的场景。

ASIP（Application Specific Instruction Set Processor）设计是介于通用处理器以及ASIC设计的一种设计方式。ASIP面向某一领域，对该领域进行分析，从而达到功耗、速度、成本、灵活性等多方面的权衡。ASIP面向某一特定领域进行特殊功能单元的设计，在通用处理器上需要多条指令完成的任务，在ASIP上可能只需要一条指令调度特殊功能单元来完成。因此相比于通用处理器，ASIP在功耗、速度上可以达到更优。而相比ASIC设计，ASIP的逻辑控制通过指令的调度完成，算法的变化只需对指令进行更改而不涉及到电路，在灵活性上具有更大的优势。ASIP更适用于需要对算法进行更新与优化的场景。

FPGA（Field－Programmable Gate Array），即现场可编程门阵列，是一种可以通过软件来编程内部连接结构，完成既定设计功能的数字集成电路。FPGA被称为“万能芯片”，它并不是一种设计方式，而是一种设计平台。在ASIC与ASIP的设计过程中，往往需要设计硬件电路，但这些电路被固化后无法进行修改，一旦失败则需要重新设计，导致开发周期的延长。而FPGA作为一种可编程的硬件电路，具有可编程灵活性高的特点，在FPGA平台上进行ASIC或ASIP的设计，可以使得设计人员在软件层面来进行硬件结构的更改，而无需替换和设计芯片，加快开发时间。

本文所研究的内容为卷积神经网络的硬件平台设计，由于神经网络算法的多样性，且深度学习领域发展迅速，算法的迭代较为快速。因此相比于ASIC设计方式，ASIP设计方式在应用场景中更为适合。同时为了减少应用中硬件的设计成本，本文选择FPGA平台来进行ASIP的设计实现，达到对卷积神经网络计算加速的效果。

* 1. 国内外现状与研究目的

神经网络的实现通常以软件方式实现，谷歌、微软等公司以及一些学者也专门开发了针对于神经网络的开源软件框架，如Tensorflow、Caffe、Pytorch等。GPU近年来发展迅速，由于GPU的高计算吞吐量，同时采用SIMD架构，可以同时多个输入数据同时进行计算，使它成为加速神经网络最好的硬件平台之一。现有的软件框架也陆陆续续开始支持在GPU上进行加速，达到CPU+GPU的异构框架模式。2012年，多伦多大学的Hinton与Alex使用GPU对卷积神经网络Alexnet进行加速，拿到了当年的ImageNet图像识别比赛的冠军，拉开了深度学习热潮的帷幕，也使得GPU加速神经网络也成为共识。然而在工业界的应用中，由于GPU存在高价格、高功耗、在推断阶段低效率等一些问题，在实际使用中并不是很理想的硬件平台。

神经网络的硬件实现由来已久，早在19世纪60年代，感知机和ADALINE神经网络模型被提出，当时有几家研究机构试图使用硬件来实现神经元。在实际实现中，感知机的突触强度是通过电动马达的转动来进行改变，这种方法有着体积大、不可靠等缺点。ADALINE模型的实现则使用电阻来改变突触强度。1974年，LouisGilstrap和Roger Barton开发出第一块神经元芯片。这种芯片只包含一个神经元，具有多个输入且互联强度可以自适应改变，将多个芯片插板构建在一起，就可以构建出具有多个神经元的网络。而随着硬件以及算法的发展，如今神经网络的神经元数量往往在十万甚至百万级别，发展新的硬件平台成为需要。

现有的神经网络硬件设计方案主要分为两种，一种是针对于某一种网络算法，使用FPGA平台进行加速，另一种是则是针对神经网络算法进行ASIC开发。

在FPGA平台上对某种特定的神经网络进行加速的研究成果较多。赵文来等学者首先发表了在FPGA平台加速LeNet网络的论文。LeNet网络是一种结构非常简单的卷积神经网络，由于其参数量较小，因此可以将所有参数放入到FPGA的内部缓存中，可以减少卷积神经网络中的数据吞吐。李慧敏等学者则在FPGA平台上对AlexNet网络进行了加速，AlexNet网络规模较大，由于FPGA内部BRAM资源有限，因此无法将参数存入FPGA中。该论文提出了将参数放入外部存储DDR中，在FPGA内部只进行中间计算数据的存储，同时为了减少所需外部存储的带宽，在计算全连接层时，采用批处理的方式，大大减少了参数加载所需带宽。之后，姚颂等学者在FPGA平台上对大型卷积神经网络VGGNet进行加速，该论文提出，卷积操作是计算密集型运算，全连接层是访存密集型计算。对于卷积操作，设计了多个可并行计算的基本单元来进行加速，对于全连接层，则设置了较大的输入带宽来保证参数加载的速度。该论文还对大型的卷积神经网络加速提出了新的方案，通过将大型卷积操作进行分解，对小型卷积操作加速的方式来完成加速，从而减少FPGA内部缓存资源的使用。

实际上，在FPGA平台上设计加速方案的一大难点是卷积操作的基本单元设计。

深鉴科技、谷歌TPU、寒武纪芯片、类脑芯片

* 1. 论文主要工作

卷积神经网络，

* 1. 论文组织安排

1. 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