**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](#_Toc532313875)

[1.1 课题背景 2](#_Toc532313876)

[1.2 国内外现状与研究目的 3](#_Toc532313877)

[1.3 论文主要工作 5](#_Toc532313878)

[1.4 论文组织安排 6](#_Toc532313879)

[第二章 相关技术研究 7](#_Toc532313880)

[2.1 卷积神经网络结构 7](#_Toc532313881)

[2.1.1 卷积层结构 7](#_Toc532313882)

[2.1.2 池化层结构 7](#_Toc532313883)

[2.1.3 批量归一化层结构 7](#_Toc532313884)

[2.1.4 激活层结构 7](#_Toc532313885)

[2.1.5 全连接层结构 7](#_Toc532313886)

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，针对神经网络算法进行ASIC开发，另一种是则是使用FPGA平台进行加速。

目前，有一些较大的公司或者机构针对神经网络进行了ASIC开发。其中较为成功的有谷歌的TPU以及寒武纪的NPU。TPU的核心在于脉动阵列，原理是数据以流水的形式向下游处理单元流动，充分利用数据的重用以减少所需带宽。脉动架构结构简单，灵活性低，但却十分适合卷积操作。寒武纪的NPU有一系列的芯片，其中针对于卷积神经网络的芯片为ShiDianNao，ShiDianNao的核心优化点在于取消使用DRAM进行数据的存放，其中处理器的物理位置比较靠近图像来源，这样就避免了图像数据的存放，而卷积操作参数量较少，可以直接存放在片上SRAM中。经过这样处理，就避免了DRAM的使用，从而降低功耗。但谷歌的TPU造价高昂且并不出售，寒武纪的TPU系列至今没有提供开发平台，这使得使用这两种芯片的途径被中断。

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功能单元设计。现有的加速方案基本都基于输入特征图并行、输出特征图并行和卷积窗口内部并行三种并行方式。输入特征图并行指的是多个输入特征图同时参与计算，输出单个输出特征图，输出特征图并行指的是，单个输入特征图与多个卷积核参数进行计算得到多个输出特征图，卷积窗口内部并行指的是一个卷积窗口内的K\*K个乘加运算同时进行。现有的加速结构一般都是使用卷积窗口内部并行，再采用输入特征图并行与输出特征图并行中的一种或者两种。卷积窗口内部并行，可以通过定制功能单元的形式来完成，在功能单元内部完成多个乘加操作。目前，最常用的卷积操作功能单元为Z型功能单元，它利用卷积窗口之间的数据重用，可以大大减少计算时所需的数据带宽。也有学者针对功能单元进行优化，如一维Z型单元与二维Z型单元、将适应于3\*3卷积运算的Z型单元组合为K\*K卷积Z型单元的优化方案等。但实际上，在卷积操作步长大于1时，卷积窗口之间的数据重用变少，Z型单元将造成计算资源的浪费。

尽管已经有许多学者基于FPGA平台对一些卷积神经网络模型进行了加速，但都只是针对于特定的网络模型，而没有一种普适的解决方案。而且，在FPGA平台上进行开发，对开发者的要求较高，需要同时对软件算法以及硬件结构有一定的了解，同时对数据的输入输出以及硬件实现算法中的一些优化都需要进行深入的勘察，这使得在FPGA平台上对卷积神经网络模型进行加速的人工成本非常高昂。因此，本文基于FPGA平台上进行ASIP设计的开发，旨在提供一种在FPGA平台加速卷积神经网络的便捷方案，使用者可以通过较少的软件层次的代码编写，就可以完成卷积神经网络的加速。同时，可以根据所使用FPGA内的资源设置加速方案，从而灵活的满足需求。

* 1. 论文主要工作

本文旨在基于FPGA平台进行卷积神经网络加速方案的ASIP设计，其主要工作内容包括：

（1）、设计卷积神经网络计算框架并优化片内存储使用

卷积神经网络内部包括卷积层、池化层、批量归一化层、激活函数层、全连接层等不同形式的运算层。同时随着学术界的研究，卷积神经网络模型的深度也在不断加深。基于这两点，论文首先设计出不同运算层通用的计算框架，其不会因模型深度的增加而无法适用。再者，由于FPGA内部存储有限，往往会成为加速的瓶颈，基于上述计算框架，论文提出了两种优化方案，可以减少一半的存储资源使用。

（2）、设计特殊功能单元

卷积神经网络的不同运算层有不同特点，其中最主要的运算层为卷积层与全连接层。卷积运算为计算密集型，而全连接层为存储密集型，根据不同层设计不同的功能单元，从而达到较优的加速结果，是本文的重点所在。同时，针对原有的卷积运算的Z型单元所存在的一些缺点，本文提出了一种旋转存储方案，可以在减少存储资源使用的同时，有效的减少在卷积操作步长大于1的情况下的运算时间。

（3）、设计TTA架构的数据互联网络

TTA架构的数据互联网络也是重点之一。优异的数据互联网络可以使数据传输所需时间变少，同时减少硬件结构的复杂性，从而减少功耗。本文基于上述的特殊功能单元间的数据流动，设计了数据互联网络，并与传统的互联网络进行了比较。

* 1. 论文组织安排

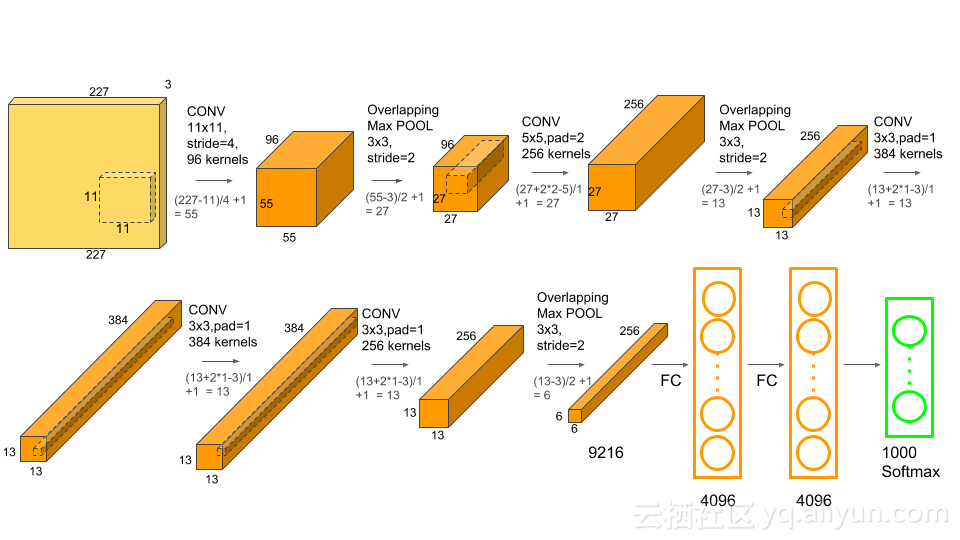
本文的组织架构如下：

第一章：绪论。主要介绍了神经网络硬件平台的研究背景和发展现状，阐述了论文的研究目的与主要工作，并安排了论文的组织结构。

第二章：相关技术研究。主要介绍了

1. 相关技术研究
   1. 卷积神经网络结构详解

卷积神经网络是一种多层神经网络，在图像识别和自然语言处理等领域都发挥着非常重要的作用。卷积神经网络一般分为两大部分，第一部分为提取特征的运算层，包括卷积层与池化层，第二部分为完成分类的运算层，包括全连接层。此外，卷积神经网络中还包括引入非线性从而提升模型表达能力的激活层，以及加快网络训练速度的批量归一化层。一个较为著名的卷积神经网络AlexNet如图X所示。下面将详细介绍各个层的计算过程。



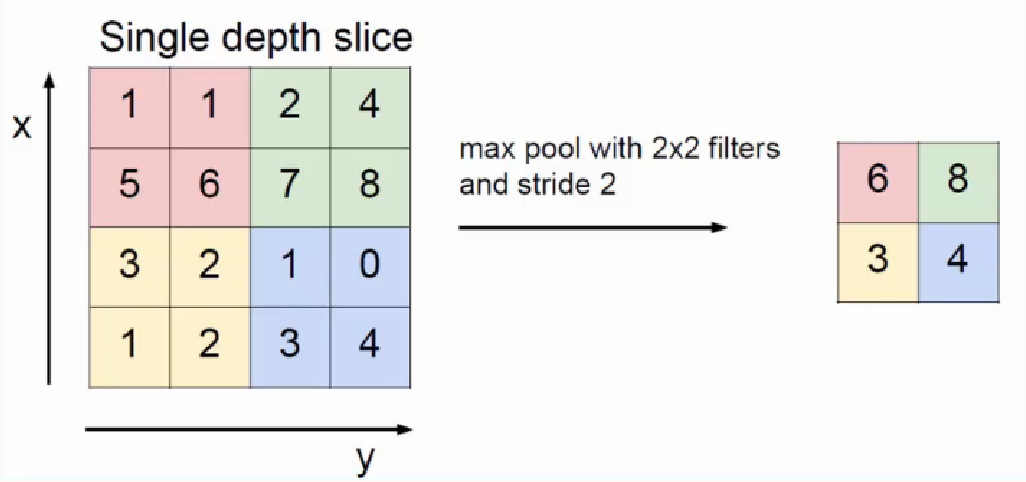
（1）卷积层

卷积运算，是将输入特征图与多个可学习的卷积核进行卷积运算，构成输出特征图的过程。一幅特征图的本质是一组离散的二维数组，对一副特征图进行卷积运算，其实际就是将一组离散的二维数据与另一组离散的二维数据进行对应位置相乘，然后进行求和。卷积运算可大概分为三个循环。最内层循环为单层输入特征图与单卷积核的单层参数之间的运算，根据卷积运算的参数，卷积窗口会不断滑动，最终产成单层输出特征图的中间结果，具体步骤如图X所示。第二层循环为对输入特征图的所有层进行遍历，输入特征图的所有层与单卷积核的所有层参数之间的运算，产生单层输出特征图。最外层循环为对卷积核进行遍历，最终产生所有输出特征图。第二层循环与最外层循环之间可以互换，而不影响最终结果的生成。



（2）池化层

池化层的主要作用是降低数据维度，从而大幅度减少计算量。池化操作与卷积操作较为相似，但主要不同的点在于，池化操作并不需要参数，本质上为下采样过程。而之所以可以进行池化操作的原因在于，即使减少了许多数据，但池化操作保留的是较为重要的一些数据，总体信息的损失程度较小。最大池化操作的具体步骤如图X所示，根据池化操作的参数设定池化窗口与池化步长，池化窗口内比较出一个最大值作为输出，不断移动池化窗口，最终得到池化结果。



（3）全连接层

卷积神经网络的卷积层以及池化层会将原始输入数据映射到特征空间，而全连接层则会将网络提取的特征映射到样本的分类空间中。全连接层的特点在于参数量较大，单个全连接层的参数往往可以达到百万级别。全连接层将卷积输出的三维特征图转化为一维向量，与多个相同长度的一维向量进行点积，最终得到一个新的一维向量。在最后一层全连接层，最终输出的一维向量长度往往为样本分类数量。

（4）激活层

激活层的作用是为神经网络引入非线性变化，提升网络的非线性表达能力。若网络中不加入激活层，则在一定程度上可以看做单层网络，因为多层的参数矩阵相乘得到的仍为一个参数矩阵。激活层是神经网络中至关重要的一层，其也影响着网络的性能，目前常用的激活函数有tanh、sigmoid、relu以及后来基于relu提出的一系列改进函数等。激活层的实现较为简单，为单输入单输出模式，输出数据维度与输入数据维度一致。

（5）批量归一化层

归一化层的作用为对输入数据进行归一化处理，将所有数据的分布转为均值0方差1的分布，使得训练数据中数值在同样量级，从而达到加速训练的效果。批量归一化层是归一化层中的一种，因为其结构简单，加速效果显著，且有一定防止过拟合的作用，在近年来的神经网络中被广泛使用。批量归一化层指的是将一个批量即一层特征图内的数据进行归一化然后输出。其过程可以使用公式表示：

计算均值：



计算方差：



归一化：



* 1. 卷积操作并行性研究

在卷积神经网络中，卷积操作是最为重要的一层，其操作结构较为复杂，计算复杂度最高，完成单层的卷积操作甚至需要亿级别的乘加操作。对卷积神经网络进行硬件加速，其中非常大的一部分任务是对卷积层完成硬件加速，而这需要深度挖掘卷积操作内部的并行结构，才能充分使用硬件进行并行计算。

在章节2.1中曾提到，卷积操作有三层循环，其正代表着卷积操作内三种并行结构，分别为：卷积窗口内并行，输入并行图并行以及输出特征图并行。

* + 1. 卷积窗口内并行

卷积窗口内并行对应着卷积操作的最内层循环，它所要实现的是同时进行卷积窗口内的K\*K次乘法以及加法，在单个时钟内输出结果。FPGA具有天然的流水结构，通过设计乘法阵列以及加法树，可以非常轻松的实现上述功能。但单个卷积窗口的计算需要同时读取K\*K个输入数据以及K\*K个核参数，所需数据带宽较高，且控制逻辑较为复杂。而且因为FPGA内的存储资源只能读取连续数据，因此卷积窗口内的K行输入数据需要K个时钟进行读取，无疑会大大增加等待时间，减少计算效率。

实际上，相邻卷积运算窗口之间会有数据被重复使用，若能把这些数据存储起来，将不需要被重新加载。对于一个规模为3\*3的卷积核，设定卷积操作步长为1，则卷积窗口与其左边的卷积窗口之间会有6个数据的复用，再加上与右面、上面、下面三个方向的卷积窗口的数据复用，最终只有1个数据需要进行加载。学者们基于相邻卷积窗口内的数据重用，设计了Z型单元，以达到简化控制逻辑与减少数据、参数带宽的目的。

Z型单元的结构如图X所示，其主要分为两个部分：数据传输模块以及运算模块。

数据传输模块包括三部分：第一部分为规模为K\*K的寄存器阵列，用来存放卷积窗口内的数据；第二部分为K-1个FIFO（First Input First Output），每个FIFO与前面一行的寄存器阵列共同存储一行数据，以保证寄存器窗口内数据的结构性；用来第三部分为参数buffer，用来存储卷积核参数。由于FIFO的深度与同行的寄存器个数相加为输入特征图的列数，因此可以保证寄存器阵列中数据为所计算的卷积窗口内的数据。在运算过程中，单个数据从寄存器阵列入口流入，寄存器阵列与FIFO中的数据依次向后移动一位，等效于卷积窗口向右移动一位，当所有数据流入完毕，得到计算结果。

运算模块包括乘法阵列以及加法树两部分，功能较简单，乘法阵列的输入为寄存器阵列以及参数buffer中的数据，将对应位置的数据与参数进行相乘。加法树的作用则是对乘法树的K\*K个结果进行相加，得到卷积结果。乘法阵列与加法树的设计一般都为流水线型，因此，在一个时钟内，就可以进行K\*K个乘法运算以及K\*K-1个加法运算，也就是说，在一个时钟内就可以输出一个卷积窗口的运算结果。

Z型单元相比只使用运算模块的设计有如下的好处。首先，利用数据的重用性，每次计算只需要加载一个数据，相比原先的K\*K个数据，大大减少了计算时所需要的数据带宽，同时也简化了数据加载逻辑。其次，卷积核的参数不需要被重复加载，在Z型单元中，参数可以预存在参数buffer中，每个卷积核的参数只需要加载一次，减少了所需参数带宽。

而Z型单元也有其不足之处。首先，对于卷积步长大于1的卷积操作，其相邻卷积窗口之间的数据复用率变低，此时可能需要加载多个数据才能进行计算，而Z型单元每个时钟只输入一个数据，导致等待时间大大增加。在卷积步长为s时，此时计算效率仅为1/s2，计算效率以平方级下降。其次，每个Z型单元将使用K个FIFO，在FPGA片内存储资源有限的情况下，可支持设计的Z型单元数据降低，加速效率可能会受到影响。基于这两个缺点，论文将在第四章设计新型单元以进行优化。

* + 1. 输入特征图并行

卷积窗口内并行解决的是单层输入特征图与单层卷积核参数之间的并行，而输入特征图有多层，结构相似，其并行性也可以进行开发。输入特征图并行方式如图X所示。输入特征图并行指的是L个输入特征图同时进行计算，也就是同时使用L个卷积窗口内并行，在当前的L个输入特征图计算完成后，切换到下一组L个输入特征图继续计算。由于L个输入特征图的计算结果都为一个输出特征图的中间结果，因此需要在L个输入特征图并行结果后加入加法树结构，再将结果暂存到输出特征图中。

输入特征图并行利用输入特征图不同层之间运算的相同性，虽然数值不同，但运算过程相同，因此，可以使用相同的控制逻辑同时对多个层同时进行计算。其具体步骤如下：

（1）切换下一组输入特征图，加载卷积核该L层对应的参数

（2）从输入buffer中同时读取L层输入特征图的数据，输入到卷积功能单元中进行计算。

（3）将L个卷积功能单元结果进行相加，中间结果存储到输出buffer中。

（4）重复（1）（2）（3），得到输出特征图的一层结果，切换卷积核。

（5）重复（4），得到输出特征图的最终结果。

* + 1. 输出特征图并行

输出特征图并行单个输入特征图与多个卷积核共同运算，生成输出特征图多层的并行方案，与输入特征图并行的思路相反。输出特征图的并行实现如图X所示。由于输出特征图并行只需要输入特征图的一层参与运算，因此L个卷积功能单元共享输入数据，L个卷积功能单元的不同之处仅在于卷积核参数的不同。

输出特征图并行的一个特点是输出中间结果较多，每一层输入特征图都对应着一部分中间结果，当下一层输入特征图计算时，需要将前面层的中间结果读出来与当前层结果进行累加，在控制逻辑上可能会略显复杂。输出特征图并行的具体计算过程如下：

（1）切换下一层输入特征图，加载L个卷积核对应参数。

（2）从输入buffer中读取该层输入特征图数，输入到卷积功能单元中进行计算。

（3）将L个中间结果暂存到输出buffer中。

（4）重复（1）（2）（3），得到输出特征图的中间结果。

（5）重复（4），得到输出特征图的最终结果。

* + 1. 并行性组合

在实际设计并行架构时，多种并行结构其实并不冲突，而且最常见的方案是设计卷积功能单元来完成卷积窗口内的并行，同时选择输入特征图并行以及输出特征图并行中的一种或者两种。

* 1. TCE开发工具链
     1. 设计流程
     2. 功能单元设计
     3. 互联网络设计
  2. 本章小结