DOI：10.7652/xjtuxb201807022

一种基于可编程逻辑器件的卷积神经网络协处理器设计

杨一晨，张国和，梁峰，何平，吴斌，高震霆
（西安交通大学电子与信息工程学院，710049，西安）

摘要：针对大数据时代下深层次大规模深度学习网络模型在预测中对运算资源和访存带宽需求指数的增长，以及业界传统CPU+GPU解决方案难以应用于日益普遍的移动嵌入式应用场景等问题，提出了一个基于可编程逻辑器件(FPGA)的卷积神经网络协处理器异构加速设计方案。该方案采用通用模型设计思想，具有可编程性，并且能够兼容多种网路模型从而实现硬件加速；方案具有可扩展性，可在硬件资源允许的范围内进行多核扩展以获得性能翻倍提升。利用硬件的并行性，数据的复用性设计的卷积运算模块提高了硬件资源利用率及运算效率；合理配置的多级缓存结构降低了协处理器对外部存储器读写频率和带宽的占用率，提升了模块内部的通信效能。在XILINX VC707评估板的上板进行实验，结果表明，MNIST-LeNet测试集的准确率高达99%，CIFAR-10可实现80%，浮点运算速度为5.511×10^{10} s^{-1}，综合性能约两倍于Intel Xeno E5-2640 V4服务器通用处理器，达到同期FPGA解决方案的主流水平。

关键词：深度学习；卷积神经网络；可编程逻辑器件
中图分类号：TP391 **文献标志码**：A **文章编号**：0253-987X(2018)07-0153-07

Design of FPGA Based Convolutional Neural Network Co-Processor

YANG Yichen, ZHANG Guohe, LIANG Feng, HE Ping, WU Bin, GAO Zhenting
(School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: In the era of big data, the demand for computing resources and memory bandwidth in deep-level and large-scale deep learning network models is increasing exponentially. Traditional industry solution CPU + GPU is not suitable to the prevalent scenarios of mobile embedded applications. To deal with this problem, we proposed a design of convolutional neural network co-processor based on FPGA programmable logic device. This solution focuses on high compatibility. It has programmability and is compatible with a variety of network models to achieve hardware acceleration. It also has scalability to allow multi-core expansion within the range of hardware resources to achieve double performance. The design of convolutional operation module focuses on hardware parallelism and data reusability, which improves the utilization of hardware resources and computing efficiency. Rationally configured multi-level buffer structure reduces the co-processor's occupancy rate of external memory's read/write frequency and bandwidth, improves the internal communication efficiency of the module. The experimental results on the XILINX VC707 evaluation board show that the accuracy of the test set is 99%, the CIFAR-10 can achieve 80%, and the peak computing capability is 5.511×10^{10}

收稿日期：2017-11-07。 作者简介：杨一晨(1994—)，男，硕士生；张国和(通信作者)，男，副教授。 基金项目：国家自然科学基金资助项目(61474093)。
网络出版时间：2018-04-20 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20180420.1538.008.html

s^{-1}, the overall performance is approximately twice that of the general-purpose processor of Intel Xeno E5-2640 V4 server. Moreover, the processing performance of our design reaches the current mainstream level of FPGA solutions.

Keywords: deep learning; convolutional neural network; programmable logic device

人工智能作为一门模拟人类能力和智慧行为的跨领域学科,机器学习是人工领域以及学术界的研究热点[1]。伴随着大数据时代的到来,深度神经网络结构愈加复杂与参数集规模愈加巨大对计算力与存储系统带来了前所未有的压力。目前,关于提供深度学习专用加速芯片成为半导体业界的热门研究方向之一,特别是在 2012 年以后涌现出大量优秀科研成果。值得注意的是,仅在 2016 年集成电路界的顶级会议 ISSCC(IEEE international solid-state circuits conference)的“下一代处理器”主题中,就有 6 篇有关深度学习专用芯片的优秀论文[2]。而 2017 年的 ISSCC 直接将主题定为了“智能世界、智能芯片[3]”。ISSCC 2018 中的 5 个全天活动围绕机器学习与推理硬件方法[4],国内各高校与研究机构也积极从事本方向的研究。中科院计算所研制的通用架构神经网络处理器“寒武纪”不但取得了从 2014 年到 2016 年间多个计算机体系结构国际会议的最佳论文[5],并顺利投入商业化。清华大学的团队开展了基于神经网络压缩技术的神经网络处理器 DPU 的设计并提供了完整的开发系统,其团队目前创立了专门进行 FPGA 加速研究的科技公司。百度公司提出了“百度大脑”计划[6],早在 2012 年就开始使用图形处理器 GPU、FPGA 进行异构加速,在数以万计的 GPU 集群中进行高性能计算,百度无人车计划也在研究相关嵌入式计算平台,华为“2012 实验室”、阿里云[7]、腾讯云也在进行关于使用 FPGA 进行加速的大规模研究,并开始提供基础服务。

本文的研究工作在课题组中属于探索性工作,目的是为课题组未来涉及机器学习与硬件相结合的研究领域建立理论与流程体系,并提供一个稳定可用的卷积神经网络协处理器 IP,成为一个更大规模基于深度学习的嵌入式计算机视觉片上系统 SoC 中的关键组件。通过对深度学习中的卷积神经网络进行算法的分析,并结合 FPGA 系统的特点进行硬件实现与优化,设计出一款基于 FPGA 的高性能、可配置的卷积神经网络协处理器。本文着重探讨了在硬件架构层级的算法实现与优化机制,并阐述了一个详细的设计方案,并全面进行了设计验证、FPGA 硬件实现与性能评估。

1 卷积神经网络分析

卷积神经网络起源于标准神经网络并提供了一种端到端的学习模型,与此同时作为神经网络领域一个重要的研究分支,卷积神经网络的特点在于其每一层的特征都是由上一层的局部区域通过共享权值的卷积核激励而得到。因此,卷积神经网络是计算机视觉领域的研究焦点,特别适用于二维数据处理的应用场景。

1.1 基本概念

卷积神经网络中,各层对应的神经元组织形式是一个三维矩阵,分别定义为长、宽、深度(也可称为通道数)。以尺寸为 32×32 的 RGB 图像作为输入层,三个维度的尺寸分别为 32、32 与 3,权值同样以三维矩阵的形式存在。因此,卷积神经网络中的每一层都完成了把输入的三维矩阵转换为输入的另一个三维矩阵的过程,卷积神经网络由一系列具有功能不同的层按序单向连接组织。当前使用的卷积神经网络结构多种多样,但均会包含以下最基本的层:卷积层;池化层;全连接层,每一层的输出还需要经过在 Alex Net[8-9] 以及各种 CNN 模型中常见的 ReLU 激活函数。

1.2 卷积层与其运算模式

卷积层是整个卷积神经网络结构的核心,对于卷积层的算法及硬件结构的优化至关重要。卷积层的输入数据是若干组分辨率相同的图像,按照卷积核的尺寸进行卷积运算,输出若干组分辨率相同的图像。图像的卷积运算通过利用一个卷积核,把对应行列的权值与输入图像的像素值进行对应的累乘及累加运算,卷积核计算的数学表达式为

$$o(n,i,j)=\sum_{m=0}^{cm-1}\left(\sum_{v=0}^{kh-1}\left(\sum_{u=0}^{kw-1}w(n,m,u,v)x(m,u+isw,v+jsh)\right)\right)$$

1.3 其他模块其运算模式

卷积层的输出往往需要经过池化或子采样。池化操作用于减小输入特征对应的每一个通道二维尺寸,进而减小了下一层计算的参数量,同时在训练效果上也可以控制过拟合[10]。以步长为 2、尺寸为 2

×2 的非交叠池化为例，二维尺寸上从每 2×2 的非交叠子区域中输出一个值，即在长宽方向上的尺寸均缩减为原特征的 1/2，数据量缩减为 1/4。全连接层的输入特征与神经元是一一连接的[11]，与标准的神经网络结构相同。全连接层的计算过程可视为普通的矩阵乘法：将输入特征延展为一维行向量，通过乘以规模比较大的二维矩阵，其输出为一个一维行向量。假设前一层的输出特征为 4×4×64，延展为 1×1×1 024 的行向量，其权值矩阵为 1 024×64 的二维矩阵，输出特征为 1×1×64。由于全连接层也可视为一个与输入特征尺寸相同的卷积核作用下的特殊卷积运算，因此现有研究中也出现了全卷积的概念，将现有网络中的全连接层全部转化为卷积进行计算。

2　协处理器概述

协处理器是辅助主处理器进行计算的模块，主要承担加速任务处理的功能。一般情况下，协处理器使用硬件设计实现几种运算复杂、耗时长的软件指令。通过简化多条指令代码为单一指令，以及直接在硬件中实现指令的方式，加速代码的计算，提升系统整体的运行效率。

2.1　卷积神经网络协处理器的意义

协处理器作为高速度和高精度的关键运算部件，其性能直接影响系统的运算能力。大数据时代的来临，卷积神经网络结构日趋复杂，大规模、深层次的网络利用了海量数据样本，其学习能力与表现能力不断提升，然而随之而来的是训练参数与计算量的成倍增加。复杂的深度学习网络的训练与预测过程需要消耗巨额浮点计算资源以及极高访存带宽，由于硬件体系结构的限制，仅利用通用架构 CPU 进行深度学习计算效率低、速度慢，难以部署大规模的计算任务。使用协处理器进行加速计算成为业内主流选择。

2.2　协处理器的通用实现

通用实现方面，CPU 可以访问加速器及存储器而加速器通过访问存储器对数据进行处理，即协处理器作为主从复合机实现整个系统的加速。通过任务内部的并行机制和自定义大小的存储器，能够为每一个计算任务优化数据路径，从而很好地处理数据。

2.3　基于 FPGA 协处理器

传统使用通用 CPU 以及 CPU＋GPU 的解决方案，结构复杂硬件开销大，并且对资源的利用率不高，而导致其工作能耗高。高端的 CPU、GPU 价格昂贵，基于 FPGA 的协处理器加速方案越来越受到学术界的关注认可。在处理模型结构复杂、数据量大、深层次的计算时，通用 CPU 仅需发送所需算法以及运算规模的计算指令给 FPGA 协处理器，由 FPGA 协处理器完成相应数据的读取以及运算实现硬件加速的功能[12]。

2.4　卷积神经网络协处理器设计思想

参考近年来新出的 CNN 模型，得到不同 CNN 模型都需要卷积层、池化层、填充单元、全连接层这些必要的基本组件，而各种模型的不同之处在于卷积层和池化层的数量以及它们的连接方式。本设计将每个组件都独立实现，并利用资源复用的思想，通过不同组件按一定顺序配置，在硬件资源允许的情况下就能实现任意结构的 CNN 模型，从而实现了卷积 CNN 的协处理器加速。

3　硬件实现细节

本协处理器设计包含了卷积处理器的完整的计算流程，具有高效能、低功耗的特点。考虑到未来应用场景对可配置性、可编程性的要求，设计了可支持不同规模，具有标准的接口、可扩展性与可定制性的卷积、池化、全连接、填充、非线性激活函数的计算模块。卷积协处理器结构图如图 1 所示。

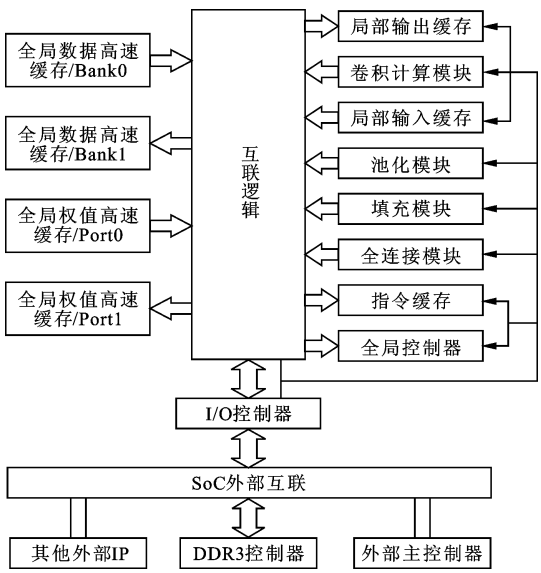

图 1　卷积协处理器结构图

卷积计算模块是整个卷积神经网络中的核心，具有计算量大、占用资源多的特点。高效能的卷积计算模块是卷积神经网络整体框架的性能分析的基础，需要着重优化卷积计算模块进而提高卷积神经网络的性能。

3.1 卷积计算特点分析

卷积计算是将卷积核和对应输入特征图的数据进行累加，其基本计算是矩阵乘法计算，由一系列的乘加计算组成。在运算中，数据会有高度的复用性以及不相关性：输入特征图数据对应多组卷积核，并且运算相互无依赖；输入特征图数据中的卷积子区域共享同一个卷积核；输入特征图像数据中的不同卷积子区域数据在空间上有大量重叠。硬件设计过程中需要在性能、控制灵活度、硬件资源使用方面进行折中权衡，片上的运算资源理论上可以全部并行使用，但其需要的数据会受到外部存储器的限制。当数据重复利用率较低的时候，运算单元需要在完成计算时立即从外部存储器中取数据，并行开启的计算单元数量会受限制于有限的外部存储器带宽单指令多数据的架构(SIMD)更适合面对大量数据的处理任务，可以大幅增强卷积神经网络的运算能力。为满足 SIMD 构架以及提升等效带宽，需要对外部存储器进行连续的读写操作，但卷积计算单元所请求的数据在存储器中的地址并不连续，会出现跨区域访存数据的情况，此时灵活的控制逻辑则是整体框架运行的保障。在硬件资源限制之内，充分利用卷积计算数据复用特性可以减小对外部存储器带宽的依赖，缩小硬件面积，提高计算能力并节约存储空间。

3.2 卷积模块硬件设计

依据上一小节所阐述的设计思路，本设计中的卷积模块架构如图 2 所示，各基本组件主要分为以下 4 个子模块：①控制模块，包含配置表与卷积单元控制器；②数据输入模块，包含输入缓存 LIB 和移位器；③计算模块，包含多组向量计算单元；④数据输出模块，包含输出缓存 LOB。

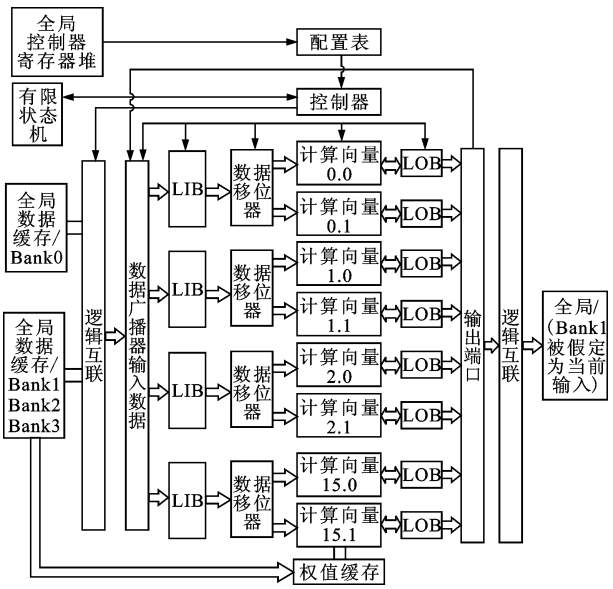

图 2 卷积协处理器结构图

计算模块由多组向量计算单元组成，在本设计中，每组向量计算单元长度为 8 个单精度浮点型，即 256 bit，其结构如图 3 所示。

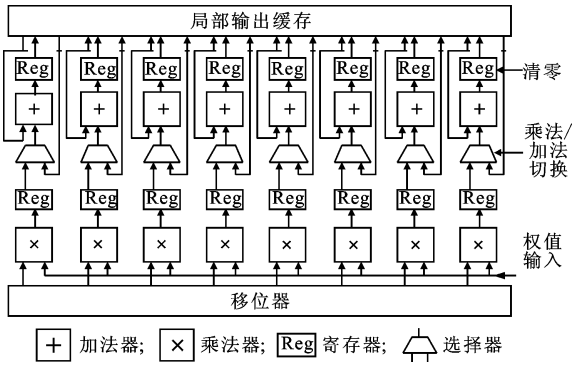

图 3 向量单元结构图示

向量计算结构中浮点加法单元，运算复杂，耗时较长。为了获得更高整体性能，并满足高频率下的时序要求，本设计选用了 3 级流水浮点加法器。由于加法器输出还要进行累加操作还需要一级的时钟节拍，所以运算结果需要 4 个时钟节拍传给下一级。由于累加计算具有依赖性，即下一组的累加计算需要等待上一组累加计算结果得出，因此需要有合理的控制逻辑保障卷积计算在流水线的正确节拍上，浮点加法流水线填充示意图如图 4 所示，其中 A_i、B_i、C_i、D_i($i=0,1,2,3$)为填入加法器流水线的数据，S_{a_i}、S_{b_i}、S_{c_i}、S_{d_i}为有效的运算结果。

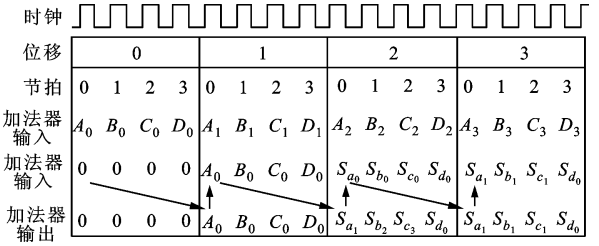

图 4 浮点流水线填充示意图

为了避免浪费流水线出现空闲节拍而造成的效率降低，在流水线的 3 个节拍上控制器需要插入 3 个属于不同输出通道的卷积核进行计算，从而保障流水线的使用效率以及整体框架的性能。输入特征将会与多组不同的卷积核进行卷积运算操作，因此输入端移位器输出的数据可被多组向量计算单元共享使用。本设计中，设置有双端口全局缓存与两组向量计算单元相连接，实现两组卷积的并行计算，同时输出两组不同通道的输出特征。如果充分利用硬

件资源，保证卷积计算单元流水线 4 个节拍不出现气泡，那么一次卷积迭代可以输出 8 组来自连续输出通道的输出特征。

3.3　其他函数单元

3.3.1　*池化单元*　池化层也称为抽象层，计算过程较为简单，分为最大池化和平均池化，计算本质并不是神经元计算，其主要作用是对特征图的数据进行采样，缩小数据规模。

根据学术界的研究成果，池化区域的尺寸多为 2×2 与 3×3，滑动步长为 2，本设计为实现更好的兼容性，同时支持这两种尺寸的池化计算。独立设计的池化单元有着更高的灵活性，实现兼容不同规模与结构的卷积神经网络。池化操作较为简单，且输入数据复用度低，在整个卷积神经网络框架中消耗时间和硬件资源较少，可以不必注重优化，故设计较为简洁。池化计算模块包含了输入数据缓冲区、控制器、最大池化计算模块、平均池化计算单元。池化操作开始时，池化控制器会从全局缓存中将特征数据读入到池化输入缓存中，池化计算模块根据计算类型计算数据，运算完成后计算结果将写回全部数据缓存。

3.3.2　*全连接单元*　全连接单元在整个构架中起到分类器的作用，全连接操作是向量的乘加操作，完成输入特征行向量与权值矩阵相乘，输出为另一个行向量的计算。根据全连接的计算特点，全连接模块支持的输入向量长度应与全局数据缓存及相应的权值缓存的带宽匹配，其性能主要受限于外部存储器的带宽。本设计中的全局数据缓存和权值的缓存位宽都是 512 bit，因此设置了 16 组浮点乘法器，其运算结果送入如图 5 所示的 16-8-4-2-1 树形结构连接的乘加结构。

浮点加法器的速度较慢，不能满足高速主时钟

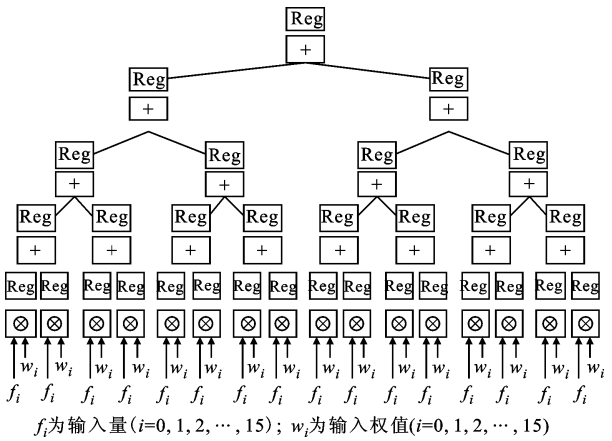

图 5　全连接单元树形乘加结构

的时序要求，这里的处理方案同前文中卷积浮点加法计算，采用了 3 级流水浮点加法器以及相应的控制器，来解决时序与数据依赖问题进而提高浮点加速器的速度。

3.3.3　*I/O 与存储体系*　本设计设置了全局数据 DDR3 SDRAM 缓存器和局部数据权值缓存的多级存储结构[13]。协处理器的工作模式是配合通用 CPU 进行计算，其外部存储器采用与 GPU 配置外部单独显存相似的解决方案，采用了 DDR3 SDRAM。

3.3.4　*I/O 控制器*　I/O 控制器完成对 DDR3 控制器与片上缓存之间的交互控制。由于片上缓存需要与 DDR3 进行频繁批量的数据交换，因此本设计的 I/O 控制器采用了 CPU 直接访存存储器方法。本设计具有独立于全局缓存器的访存控制器，具有对 DDR3 控制器批量访问读写，以及各有效信号进行控制，并实现自动完成不同字长的地址转换。将原本需要使用多种信号控制的 DDR3 控制器，简化为控制 I/O 控制器的一条读写指令。

3.4　全局控制器与指令系统

3.4.1　*全局控制器*　全局控制器实现对本设计各个模块的控制，该模块可解析指令使得各单元进入不同的工作状态，完成相应的工作。本模块包含了 8 个 32 位寄存器组成的通用寄存器堆、指令译码单元、独立指令缓存，容量为 4 kB、对应各个模块的控制接口。当外部主机或外部主控向协处理器发出启动信号后，协处理器从等待状态上线，自动进入启动状态，访问 DDR3 控制器的指令存储区域，指令指针自动从零地址开始读取指令，读取到的指令送入译码器之后执行，在执行上一条指令的同时读取下一条指令。

3.4.2　*指令系统*　控制器完成了任务的调度工作，本设计为了实现更高的灵活性兼容性，设计了指令系统来直接控制计算单元。采用位宽为 32 bit 的指令，指令的具体格式如表 1 所示。

4　实　验

本设计的硬件测试平台为 XILINX 公司的 VC707 评估套件，其搭载的 FPGA 芯片为 Xilinx Virtex-7 VX485T，在 VIVADO2016 软件上进行综合，使用 100 MHz 的系统主时钟，由时序分析得知，本设计各条关键路径满足时需要求。本设计使用 MNIST-LeNet、CIFAR-10 测试集进行测试，得到较高的准确率，MNIST-LeNet 的准确率高达 99%，CIFAR-10 可实现 80%，与标准实现准确率一致。

表 1 指令格式说明

指令名称	31～28 bit	27～25 bit	24～22 bit	21～16 bit	15～0 bit
NOP	0000	保留域	保留域	保留域	保留域
LDR	0001	寄存器号	保留域	保留域	保留域
LDEXT	0010	目的存储器	保留域	保留域	保留域
STREXT	0011	源存储器	保留域	保留域	保留域
SET	0100	层属性	保留域	保留域	保留域
EXE	0101	具体单元	读写切换	保留域	保留域
ADDi	0110	目的寄存器	立即数(0～24 bit)		
ADDR	0111	目的寄存器	源寄存器	保留域	保留域
MULTr	1000	目的寄存器	源寄存器	保留域	保留域
BNE	1001	目的寄存器	源寄存器	分支跳转地址(0～21 bit)	

表 2 不同平台的耗时比

测试集	运行时间/ms			相对运行时间比
	E5-2640 V4	GTX-1080	FPGA	
MNIST-LeNet	0.231	0.003 2	0.086	1∶72∶2.7
CIFAR-10	1.234	0.010 8	0.792	1∶114∶1.6
AlexNet	72.843	0.294 0	39.07	1∶248∶1.9

由表 2 可知，相比通用 CPU，使用基于 FPGA 卷积神经网络协处理器能效优势显著。参与对比的通用处理器为英特尔公司的服务器处理器，型号为 Intel Xeno E5-2640 V4，其构架为 64 位 CPU 结构，拥有 10 个核心，TDP 为 90 W。在 UBUNTU14.04 上利用 caffe 开源平台运行 10 000 张图片进行测试，最终统计运行的时间，并计算每幅图的平均值。从测试结果可以看出，在实验通用 CPU 上运行集的速度，仅有在 FPGA 协处理器上$\frac{1}{2.7}$，CIFAR-10 只有$\frac{1}{1.6}$，由此可看出本文提出的卷积神经网络加速器设计大幅度提高卷积神经网络的速度与效率。

本设计的卷积计算模块具有多核扩展性，在硬件资源允许的情况下，可以实现更高的运算性能。计算单元扩展，在 XiLinx VC709 评估板其核心 FPGA 芯片为 Xilinx Virtex-7 VX690T 上实现，卷积计算单元中，从每组两个计算向量扩展为每组 4 个计算向量，本设计方案与同期学术界对神经网络 FPGA 加速器的性能对比如表 3 所示。由表 3 可知，在单精度的设计方案中，本设计方案的性能达到同时期主流 FPGA 加速方案的水平。

表 3 与其他文献的性能对比

方案特性	设计年份	器件型号	时钟频率/MHz	数值格式	运算速度/10^9 s^{-1}
本设计	2016	Virtex-7 485T	100	单精度	55.11
文献[14]	2015	Virtex-7 485T	100	单精度	62.62
文献[15]	2016	Stratix -VGXA7	181	单精度	50.07
文献[16]	2016	Zynq XC7Z045	150	16 位定点	187.80

5 结束语

本文提出一种针对卷积神经网络完整的协处理器加速方案，可实现卷积、池化、全连接等的计算操作，并采用硬件资源复用的思想，可以支持不同层数的计算。本设计的卷积计算单元，在硬件资源充裕的情况下，可以进行多核扩展，实现性能翻倍。通过指令系统通过不同的指令配置实现不同规模的网络运算，相较传统固定规模的硬件设计，有着更高的灵活性和通用性。

本设计充分利用了硬件并行计算带来的优势，通过与通用处理器方案对比，体现了本设计的能效比优势；着重利用了卷积的数据复用特点，设计了卷积计算单元；通过合理的多级缓存体系的设计，使用一定的片上资源，降低了协处理器对外部存储器读写频率和带宽的占用率，使得协处理器内部各模块通信更加高效，同时降低了功耗。

本文全面验证、测试与评估了本设计，完成了行为级仿真、RTL 级仿真、FPGA 上板验证，功能符合设计预期，关键路径满足时序要求。使用经典训练集 MNIST-LeNet、CIFAR-10 测试，与标准实现准确率一致，通过对 CPU、GPU 的性能对比，显示出本设计更高的运算效率。

参考文献：

[1] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors [J]. Nature, 1986, 323(6088): 533-536.

[2] LIANG P, VERHELST M. Session 14 overview: next-generation processing [C] // Solid-State Circuits Conference. Piscataway, NJ, USA: IEEE, 2016: 252-253.

[3] DALY D, FUJINO L. ISSCC 2017: intelligent chips

for a smart world [J]. IEEE Solid-State Circuits Magazine, 2016, 8(4): 92-93.

[4] FRIEDMAN D. Hardware approaches to machine learning and inference [C]// 2018 IEEE International Solid-State Circuits Conference. Piscataway, NJ, USA: IEEE, 2018: 2376-8606.

[5] CHEN T, DU Z, SUN N, et al. DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning [J]. ACM SIGPLAN Notices, 2014, 49(4): 269-284.

[6] 陆奇. 百度大脑是百度 AI 平台核心智能云有机会颠覆云市场 [J]. 信息与电脑(理论版), 2017(14): 1-2.
LU Qi. Baidu's brain is the core of Baidu's AI platform Smart Cloud has the opportunity to subvert the cloud market [J]. China Computer & Communication, 2017(14): 1-2.

[7] 佚名. 阿里巴巴携手英特尔开发一款基于 FPGA 的解决方案 [J]. 中国电子商情：基础电子, 2017(3): 30.
Anonymous. Alibaba teamed up with Intel to develop an FPGA-based solution [J]. China Electronics Market: Basic Electronics, 2017(3): 30.

[8] 卢宏涛，张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述 [J]. 数据采集与处理, 2016, 31(1): 1-17.
LU Hongtao, ZHANG Qinchuan. Applications of deep convolutional neural network in computer vision [J]. Journal of Data Acquisition and Processing, 2016, 31(1): 1-17.

[9] ANTHIMOPOULOS M, CHRISTODOULIDIS S, EBNER L, et al. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network [J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1207.

[10] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems. Piscataway, NJ, USA: IEEE, 2012: 1097-1105.

[11] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2015: 3431-3440.

[12] 余奇. 基于 FPGA 的深度学习加速器设计与实现 [D]. 合肥：中国科学技术大学, 2016: 5-20.

[13] 陈建英，杨宪泽，张楠. 面向大规模分布式系统的多级缓存信息结构研究 [J]. 西南民族大学学报(自然科学版), 2012, 38(3): 457-460.
CHEN Jianying, YANG Xianze, ZHANG Nan. Research on multi-level structure of cache information in large-scale distributed system [J]. Journal of Southwest University for Nationalities, 2012, 38(3): 457-460.

[14] ZHANG C, LI P, SUN G, et al. Optimizing FPGA-based accelerator design for deep convolutional neural networks [C]// Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. New York, USA: ACM, 2015: 161-170.

[15] WANG D, AN J, XU K. PipeCNN: an OpenCL-based FPGA accelerator for large-scale convolution neuron networks [EB/OL]. [2018-03-16]. http://pdfs.semanticscholar.org/8d6d/df21989e9b5bd15e4bff972e2370d5dc47d2.pdf.

[16] QIU J, WANG J, YAO S, et al. Going deeper with embedded FPGA platform for convolutional neural network [C]// Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. New York, USA: ACM, 2016: 26-35.

[本刊相关文献链接]

王亚昆，吴爱国，董娜. 吸收式制冷机组逆神经网络设定优化. 2018, 52(1): 123-128. [doi: 10.7652/xjtuxb201801018]

曹玉良，明廷锋，贺国，等. 基于深度学习的离心泵空化状态识别. 2017, 51(11): 165-172. [doi: 10.7652/xjtuxb201711023]

王丽华，谢阳阳，张永宏，等. 采用深度学习的异步电机故障诊断方法. 2017, 51(10): 128-134. [doi: 10.7652/xjtuxb201710021]

章军辉，李庆，陈大鹏. 基于 BP 神经网络的纵向避撞安全辅助算法. 2017, 51(7): 140-147. [doi: 10.7652/xjtuxb201707020]

李晓娟，孙文磊，袁逸萍，等. 扰动环境下作业车间网络多瓶颈识别方法研究. 2016, 50(12): 64-72. [doi: 10.7652/xjtuxb201612011]

孙泽宇，伍卫国，曹仰杰，等. 无线传感器网络中能量均衡参数可控覆盖算法. 2016, 50(8): 77-83. [doi: 10.7652/xjtuxb201608013]

张小栋，郭晋，李睿，等. 表情驱动下脑电信号的建模仿真及分类识别. 2016, 50(6): 1-8. [doi: 10.7652/xjtuxb201606001]

姜涛，黄伟，王安麟. 多路阀阀芯节流槽拓扑结构组合的神经网络模型. 2016, 50(6): 36-41. [doi: 10.7652/xjtuxb201606006]

张鹏，陈湘军，阮雅端，等. 采用稀疏 SIFT 特征的车型识别方法. 2015, 49(12): 137-143. [doi: 10.7652/xjtuxb201512022]

（编辑　赵炜）