

引用格式:傅思扬,陈华,郁发新. 基于 RISC-V 的卷积神经网络处理器设计与实现[J]. 微电子学与计算机, 2020, 37(4): 49-54. [FU S Y, CHEN H, YU F X. Design and implementation of convolutional neural network processor based on RISC-V instruction set[J]. Microelectronics & Computer, 2020, 37(4): 49-54.]

基于 RISC-V 的卷积神经网络处理器设计与实现

傅思扬, 陈 华, 郁发新

(浙江大学 航空航天学院, 浙江 杭州 310058)

摘 要: 针对卷积神经网络对于运算资源需求的不断增长, 和传统的硬件卷积加速方案在功耗、面积敏感的边缘计算领域难以应用的问题, 设计并实现了一个低功耗嵌入式卷积神经网络加速处理器. 目标处理器基于 RISC-V 指令集架构, 内核扩展 4 条自定义神经网络指令, 并在硬件层面实现加速处理. 该卷积神经网络处理器最大程度的复用了原 RISC-V 的数据通路和功能模块, 减小了额外的功耗和芯片面积等资源开销. 目标处理器通过 RISC-V 官方标准测试集验证, 并对 MNIST 手写数据集进行识别测试, 正确率达到 97.23%. 在 TSMC 40nm 标准数字工艺下, 目标处理器面积仅为 0.34 mm², 动态功耗仅为 11.1 μw/MHz, 与同期处理器相比, 面积和功耗方面均具有一定优势.

关键词: 处理器; 卷积神经网络; 定制指令集; RISC-V

中图分类号: TP332

文献标识码: A

文章编号: 1000-7180(2020)04-0049-06

Design and implementation of convolutional neural network processor based on RISC-V instruction set

FU Si-yang, CHEN Hua, YU Fa-xin

(college of Aeronautics and Astronautics, Zhejiang University, Hangzhou, 310058, China)

Abstract: A low power embedded convolutional neural network acceleration processor is designed and implemented in response to the growing demand for computational resources in convolutional neural networks and the difficulty of applying traditional hardware convolution acceleration schemes in power- and area-sensitive edge computation applications. The target processor based on the RISC-V instruction set architecture, extend four custom neural network instructions, and accelerates the processing on the hardware architecture. The convolutional neural network processor maximizes the reuse of the original RISC-V data path and functional modules, reducing additional resource overhead. The target processor is verified by the RISC-V official standard test set, and the MNIST handwritten data set is identified and tested, with a correct rate of 97.23%. The target processor occupies 0.34mm² area and consume 11.1μw/MHz dynamic power using TSMC 40nm technology. Compared with the correlation processor, it has certain advantages in area and power consumption.

Key words: processor; convolutional neural network; custom instruction set; RISC-V

1 引言

RISC-V 是由加州大学伯克利分校设计研发的

一种新型精简指令集架构^[1], 由于其开源、轻量、优秀的拓展性等诸多优点, 受到了学术、商业等界的青睐. 目前已有多款 RISC-V 开源软核, 例如 Rock-

et^[2]、BOOM、V-scale 等。RISC-V 在提供简洁完整的基本指令集的同时,新增了大量扩展指令集,可在特定领域实现内核增强加速处理。在目前的 RISC-V 框架下,保留了 4 条指令用于实现定制扩展指令集,以实现特定计算环境下的更优性能^[3]。

卷积神经网络(convolutional neural network, CNN)是目前人工智能,机器学习等研究热点的基本框架,尤其是在图像处理方面已全面实用化。目前卷积神经网络主要依靠大规模的矩阵运算提取目标的特征值并进行分类^[3],其主要瓶颈在于卷积层消耗了大量的计算资源,且市面上的通用处理器大多

未对该类运算做加速处理。

本文基于 RISC-V 架构,设计并实现了 4 条定制 CNN 加速指令,并在原有的 RISC-V 标准处理器上进行深度架构融合,实现内核级的 CNN 加速处理能力。

2 标准 RISC-V 处理器设计

本文中设计实现的处理器采用标准的 5 级流水线,分为取指、译码、执行、访存、写回五阶段。结合分支预测、异常处理、数据前推等流水线相关技术,整体的处理器架构如图 1 所示。

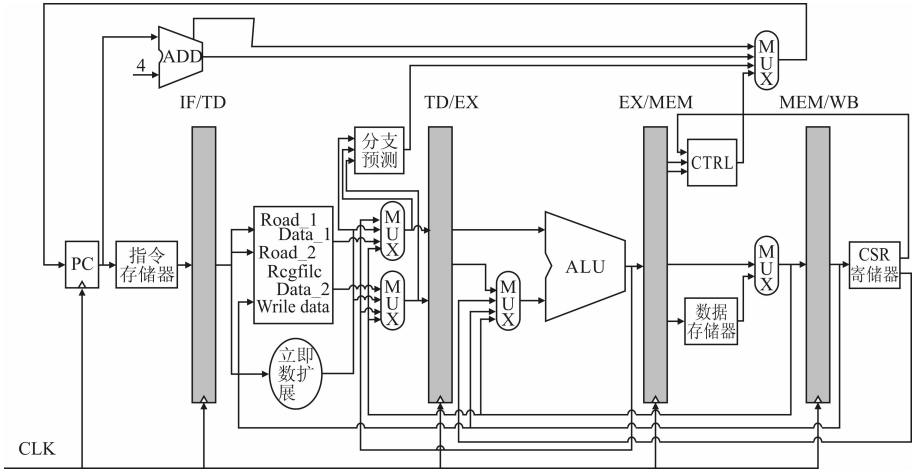


图 1 5 级流水线处理器系统架构

同时,为了验证处理器相关功能,设计并实现了小型 SOC,采用 Wishbone 总线挂载 Flash、Gpio、uart 等外围设备。在最新的 RISC-V 标准中,原本映射在 CSR 地址空间中的计数器、定时器、以及软件中断寄存器被移至外围地址空间中,所以在总线上还挂载了 Clint 模块,用于实现定时中断和软件中断,该模块处于低速时钟域中,以减少处理器功耗,用于控制中断的相关寄存器地址如表 1 所示。

表 1 中断控制寄存器地址

地址	寄存器名称	功能描述
0X4000_BFF8	mtime	64 位计数器
0X4000_4000	mtimecmp	计数器比较寄存器
0X4000_0000	msip	软件中断寄存器

3 CNN 加速处理器设计

3.1 CNN 加速指令集设计

RISC-V 标准集不仅仅支持标准扩展,还支持用户自定义四条指令用于定制特定功能。本文设计了四条指令用于实现 CNN 的加速,以此加速卷积

处理性能,减小代码空间,提高能效效率。图 2 为本文设计的扩展指令格式。

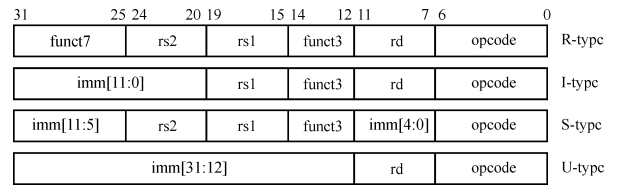


图 2 本文设计自定义指令格式

神经网络处理包括卷积、池化、激活、全连接等网络层次,对应到具体的运算类型主要为卷积运算、向量存储等操作,为此本文设计了 CONV、POOL、VLOAD、VSTORE 四条指令用于实现各层的加速。

相较于其他的卷积神经网络处理器采用 128 位长指令,需要额外的译码操作^[4-5]。本文设计的指令类型为 32 位标准指令长度,较 128 位长指令有较大的提升。同时在译码阶段,尽可能的实现了同基本整数指令集 RV32I 译码电路的复用,减小了硬件电路设计的复杂度。图 3 为 RV32I 的标准指令格式。

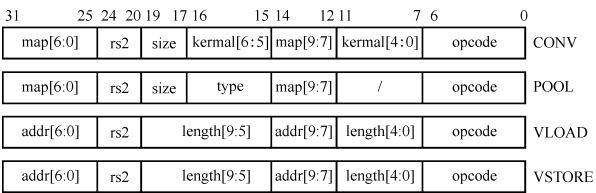


图 3 RV32I 标准指令格式

3.1.1 CONV 指令设计

卷积操作实际上是一连串的乘加操作集合,该类型操作一般占到整个卷积神经网络的 80%以上.为此,本文专门设计了 CONV 指令.译码阶段,指令依据 kernal_addr 和 map_addr 字段从片上数据暂存器读取特征图数据和卷积核数据,进行卷积操作,并将结果写入外部存储器中,外部存储器地址为寄存器 rs2 中的值.

指令中 size 字段为卷积窗口长度,本设计根据嵌入式应用的特点实现支持 2 * 2 到 5 * 5 范围的卷积操作.

3.1.2 POOL 指令设计

POOL 指令设计主要为池化层和激活函数进行加速处理.目前应用较广的池化操作主要用均值池化和最大值池化,本文设计指令可通过 type 字段自由选择两种池化方式,并将池化结果写入外部存储器中,外部存储器地址为寄存器 rs2 中的值.

在激活函数的选择上,目前使用较多的主要有 Relu,Sigmoid 两类激活函数,其中 Relu 实现简单,同时效果较 Sigmoid 更佳,已逐渐成为目前主流的激活函数选择,本文设计的激活函数默认采用 Relu 函数.

POOL 指令中[11:7]字段目前为保留位,考虑到卷积神经网络中存在少量的偏置量,在未来的设计中可以在将该字段设为通用寄存器索引,用于偏置量的读取相加,进一步减少运算开销.

3.1.3 VLOAD VSTORE 指令设计

卷积神经网络中,由于需要不停移动卷积核进行卷积运算,在这过程中需要不停的加载存储数据,其中有大量数据是重复访存,造成了极大的硬件开销浪费.为此本文设计了 VLOAD、VSTORE 两条向量访存指令.

为减少重复访存带来的周期损耗,本文在片上设计了一块独立的临时存储器,为了节省代码空间采用独立编址,通过 VLOAD 和 VSTORE 指令进行片上临时存储器和外部统一编址内存进行数据的

交互.其中外部内存地址由寄存器 rs2 中的值决定,addr 为片上临时存储器加载/存储的起始地址,length 为加载/存储的字节长度.最大支持单条 1024 字节加载/存储.

3.2 CNN 加速处理器架构设计

结合上文设计的四条自定义 CNN 加速指令,本文在原处理器架构的基础上进行了针对性的设计,主要增加了片上的临时数据存储器、卷积运算单元和针对自定义指令的流水线控制机制完善.CNN 加速处理器架构设计如图 4 所示.

同图 1 标准 RISC-V 处理器架构相比,主要添加了片上数据暂存器和卷积运算单元,同时添加了由数据存储器到片上数据暂存器的数据通路,用于支持 VLOAD、VSTORE 指令.由于 4 条定制指令大多为多周期指令,还需添加额外的流水线暂停控制信号.

3.2.1 独立编址片上 buffer 系统架构

目前 CNN 加速模块瓶颈主要在于重复访存带来的周期冗余.以基于 FPGA 的 CNN 加速方案为例,大多数待处理数据暂存于外部存储设备中,如 DRAM,其访存周期往往较大,导致处理器核的阻塞.考虑到卷积操作的窗口移动特性,即使采用各种数据复用算法,仍无法完全消除重复访存带来的影响.

Bankman^[6-7]等人的研究结果已表明,将存储器集成于片上以减少访存开销的策略是可行的,采用 SRAM 构建片上数据存储器,后续卷积操作可在单周期内取到所需的操作数,可大大减少重复访存带来的周期损耗.同时复用数据通路和运算模块,控制整体系统的资源开销.

以 5 * 5 的输入图像,3 * 3 的卷积核,步长为 1 的卷积操作为例,传统访存方案需要 81 次访存操作,而采用 SRAM 构建片上 buffer 方案只需 25 次访存操作,可有效节省访存周期.

考虑到功耗和面积的要求,本文中 buffer 采用数据位宽为 32 bit,深度为 1 024 的 SRAM 实现,并采用独立编址以减少地址编码长度从而减少代码体积.同时,独立编址的方式使得 buffer 只限自定义指令得以访问其内部数据,原 RISC-V 指令均无法对 buffer 内数据进行修改,确保在复杂的卷积操作过程中对数据本身的保护.

该 buffer 在进行数据访问时,对特征图数据和卷积核数据进行了地址划分,具体地址配置如表 2 所示:

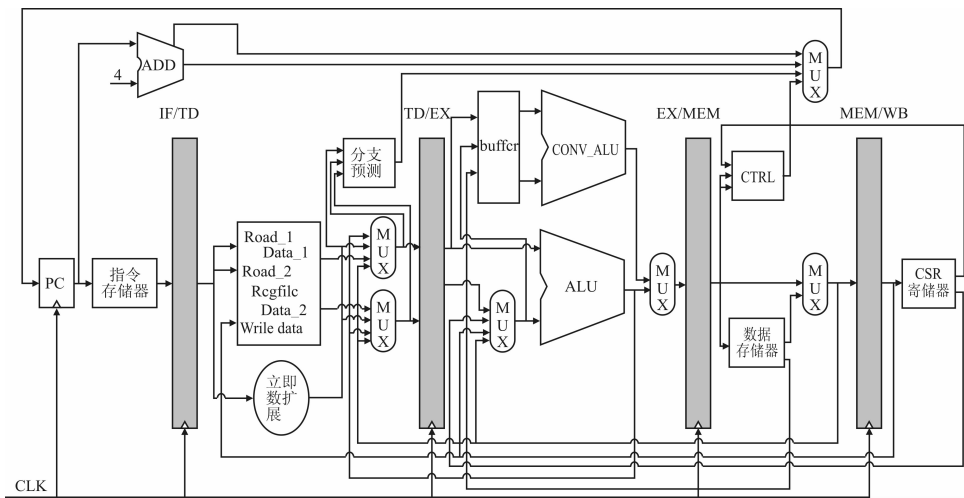


图 4 CNN 处理器系统架构

表 2 片上数据存储器地址分布

数据类型	地址空间
卷积核数据	0~127
特征图数据	128~1023

限定了卷积核数据的地址空间后,其地址编码被限制在 00000_00000 ~ 00011_11111 之间,结合图 2 中的 CONV 指令,在卷积核和特征图数据均采用一套地址的情况下,可以通过缩减卷积核编码地址高 3 位实现译码。

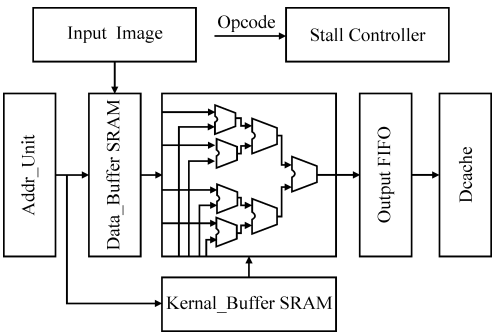


图 5 CNN 加速模块系统架构

图 5 为该系统整体架构,通过向量访存指令将片外待处理数据搬入片上 buffer 后,随即执行卷积操作,地址生成单元依据卷积单元首地址和窗口长度生成目标数据和卷积核地址,并将数据送入运算模块中,最终将运算结果输出至外部存储 DRAM 中.该系统架构采用全流水线结构,确保处理器的高速运行。

3.2.2 流水线暂停机制

由于 VLOAD、VSTORE 指令均为向量操作指令,需要花费多个周期同片外存储进行数据交互,需

要在原有的 RISC-V 处理器上添加额外的流水线暂停机制。

图 6 为本设计针对加速指令的流水线暂停机制:位于流水线第 n 阶段的指令需要多个周期执行,则暂停流水线,保持该阶段及其之前阶段的寄存器值不变,该阶段后的指令继续执行. MEM 阶段中检测到 VLOAD/VSTORE 指令的 op 字段后,产生暂停请求信号暂停流水线. CONV 指令同理,在 EX 阶段申请流水线暂停。

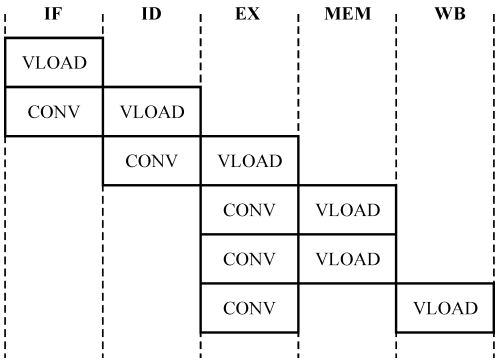


图 6 新增流水线暂停机制

3.2.3 卷积运算模块设计

目前进行卷积运算加速的实现大多基于大规模乘法器和加法树^[8-10],实现单周期内单卷积核的卷积运算,此类方法适用于 FPGA 等对于功耗和面积要求不高的场合或者用于协处理器中^[11-12],存在功耗过高、面积过大且不适用于移动嵌入式场景等问题.对于本文所实现的轻量级处理器而言,采用更为优化的方案:单周期内只对 1 行数据进行卷积操作,后接累加器进行行卷积结果累加得到运算结果,此类方法虽然在运算周期上有所增加,但是在运算资

源和功耗面积等参数上有明显的改进,同时通过流水线实现更高的运行频率。

以 5×5 大小的卷积为例,传统方案需要 25 个乘法器,同时需要大量的加法器,而本文方案只需要 5 个乘法器和少量加法器。图 7 为本文设计卷积运算模块,由于池化操作同卷积操作类似,在卷积运算模块中添加了池化运算电路,复用数据通路,进一步减少面积。

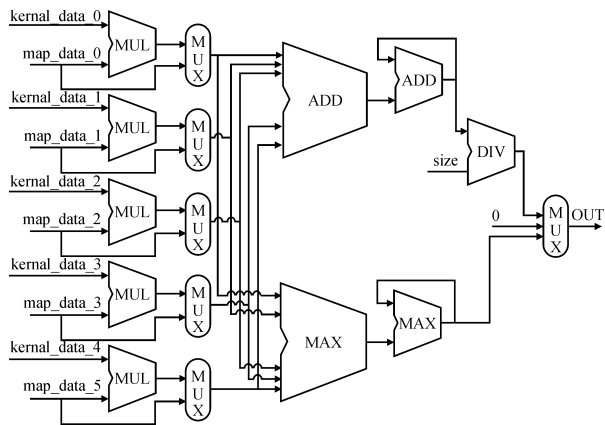


图 7 卷积运算模块

4 验证及分析

本设计的硬件验证平台为 XILINX 公司的 XC7A35T 系列 FPGA 开发板,软件验证平台为 Synopsys 公司的 Design Compiler。验证主要分为两部分:RISC-V 常规指令功能验证,CNN 加速指令功能验证。

4.1 常规 RISC-V 功能验证及分析

本文设计的常规处理器采用 UCB 官方 RISC-V 工具链中的回归测试集^[13]进行处理器常规功能验证,该测试集采用宏定义组成程序点,测试指令集架构中定义的指令,涵盖目前 RISC-V 标准指令集下所有指令,并可反映数据相关、异常处理等问题。将该测试集编译成二进制文件后,写入 FPGA 开发板板载 flash 中,进行回归测试。经测试可知,本文设计的处理器通过该测试集的验证,常规功能正常无误。

RISC-V 测试集只提供了基本的功能验证,不包含重要的中断处理验证,为此本文设计了简单的中断测试程序:处理器开放 UART 串口接收中断,接收来自外部数据后,关闭接收中断并保存数据,开启串口发送功能,将接收数据再次发送,并重新开启接收中断。外部中断功能正常。软件中断,定时器中断采用同样的方式进行验证,结果均正确。

4.2 CNN 加速功能验证及分析

TensorFlow 是目前主流的人工智能,机器学习框架,为验证 CNN 加速处理器的功能,首先在 TensorFlow 上搭建卷积神经网络,并对 MNIST 手写数字数据集^[14]进行识别训练,将训练得到的卷积核,全连接层数据进行 16 位定点化处理,并写入外部 SRAM 中,并编写汇编程序进行处理。

该卷积神经网络共包含两个卷积层。

卷积层 1 采用 4 个 3×3 的卷积核,步长为 1。输入 28×28 的标准手写数字图片,经 RELU 池化后得到 4 张 8×8 的特征图。

卷积层 2 采用 8 个 $4 \times 3 \times 3$ 的卷积核,步长为 1,输入特征图卷积池化后,输出 8 张 2×2 的特征图。

全连接层将上层特征图展开为 1×32 的向量,并与大小为 32×10 的权值矩阵进行矩阵相乘,得到 1×10 的向量,并采用分类器进行最终的图像分类。

在进行 16 位定点量化后,针对 10 000 张标准 MNIST 手写数字图片进行识别验证,正确率为 97.23%,符合预期。单张图片识别花费 35 326 个周期。同时在原架构 RISC-V 处理器上进行识别验证,结果单张图片识别花费 151 482 个周期,加速比为 4.29。

4.3 面积与功耗分析

在 Synopsys 公司的 Design Compiler 中使用 TSMC40nm 工艺库,以 50 MHz 频率和 500 MHz 分别进行综合,50 MHz 频率下,本文设计处理器面积为 0.34 mm^2 ,静态功耗为 1.42 mW,总动态功耗为 1.432 mW。500 MHz 频率下,本文设计处理器面积为 0.39 mm^2 ,静态功耗为 1.92 mW,总动态功耗为 5.73 mW。表 3 为本文设计处理器与同期处理器的参数对比。由于动态功耗同频率成正相关特性,考虑到工作频率带来的功耗差异,表 3 中采用动态功耗同频率的比值来表征处理器的实际功耗情况。从表 3 中可知,本文设计的处理器在 50 MHz 和 500 MHz 下,功耗表现良好。面积则由于添加了 CNN 加速单元,较常规处理器面积有所增加。

5 结束语

本文在标准 RISC-V 架构基础上,扩展设计了 4 条自定义卷积神经网络加速指令,并配套实现了相应的架构,通过在处理器内部创建独立编址数据暂存器以减少访存带来的性能损失,同时通过卷积加速模块以提升运算速度,实现对卷积神经网络的

加速. 目标处理器通过了 UCB 官方的 RISC-V 验证测试集,同时采用 MNIST 手写数据集验证了 CNN 加速功能. 最终本设计处理器在 TSMC40 nm 工艺

下,实现了 11.1 $\mu\text{w}/\text{MHz}$ 的低动态功耗和 0.338 mm^2 的超小面积. 可为后续人工智能、边缘计算等应用方向提供低功耗轻量化处理器平台支撑.

表 3 不同处理器的性能对比

型号	CNN 加速	指令集架构	工艺(nm)	频率	面积(mm^2)	动态功耗
本设计—50 MHz	有	RISC—V	40	50 MHz	0.34	11.1 $\mu\text{w}/\text{MHz}$
本设计—500MHz	有	RISC-V	40	500 MHz	0.39	11.4 $\mu\text{w}/\text{MHz}$
Cortex-A5 ^[15]	无	ARM	40	>1 GHz	0.53	<0.08 mW/MHz
Rocket ^[15]	无	RISC-V	40	>1 GHz	0.39	<0.034 mW/MHz
RICY5 ^[15]	无	RISC-V	65	654 MHz	—	17.5 $\mu\text{w}/\text{MHz}$

参考文献:

[1] WATERMAN A, LEE Y, PATTERSON D A, et al. The RISC-V instruction set manual, volume I: Base user-level ISA [J]. EECS Department, 2011, 7 (9): 475.

[2] ASANOVIC K, AVIZIENIS R, BACHRACH J, et al. The rocket chip generator [R]. EECS Department, University of California, Berkeley, 2016.

[3] 吕倩茹, 王彦鹏, 曹壮, 等. 面向网络报文转发的 RISC-V 压缩指令定制 [J]. 计算机工程与科学, 2018, 40(3): 381-387.

[4] 阎强. 卷积神经网络处理器的设计与实现[D]. 西安: 西安理工大学, 2017.

[5] 马珂. 具有卷积神经网络扩展指令的微处理器的设计与实现[D]. 西安:西安理工大学, 2018.

[6] BANKMAN D, YANG L, MOONS B, et al. An always-on 3.8 $\mu\text{J}/86\%$ CIFAR-10 mixed-signal binary CNN processor with all memory on chip in 28nm CMOS [C] //2018 IEEE International Solid - State Circuits Conference - (ISSCC). San Francisco: IEEE, 2018: 222-224.

[7] MOONS B, BANKMAN D, YANG L, et al. Binar-Eye: An always-on energy-accuracy-scalable binary CNN processor with all memory on chip in 28nm CMOS [C] // 2018 IEEE Custom Integrated Circuits Conference (CICC). San Francisco: IEEE, 2018: 1-4.

[8] 方睿, 刘加贺, 薛志辉, 等. 卷积神经网络的 FPGA 并行加速方案设计 [J]. 计算机工程与应用, 2015, 51 (8): 32-36.

[9] PEEMEN M, SETIO A A A, MESMAN B, et al.

Memory-centric accelerator design for convolutional neural networks[C]. // 2013 IEEE 31st International Conference on Computer Design (ICCD). Asheville, IEEE, 2013: 13-19.

[10] GUO K, SUI L, QIU J, et al. Angel-Eye: A Complete Design Flow for Mapping CNN onto Embedded FPGA [J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2017(99): 1-1.

[11] 杨一晨, 梁峰, 张国和, 等. 一种基于可编程逻辑器件的卷积神经网络协处理器设计 [J]. 西安交通大学学报, 2018, 52(7): 158-164.

[12] 杨维科, 贺光辉, 景乃锋. 基于 Rocket-Chip 开源处理器的 CNN 加速模块的设计及实现 [J]. 微电子学与计算机, 2018, 35(4): 17-21.

[13] University of California, Berkeley, RISC-V Tools (ISA Simulator and Tests) [CP/OL]. <https://github.com/riscv/riscv-tests>, 2015.

[14] Yann L, Corinna C, Christopher J C. The MNIST database of handwritten digits [DS/OL]. <http://yann.lecun.com/exdb/mnist/>, 1998.

[15] 雷思磊. RISC-V 架构的开源处理器及 SOC 研究综述 [J]. 单片机与嵌入式系统应用, 2017(2): 56-60.

作者简介:

傅思扬 男,(1995-),硕士研究生. 研究方向为数字集成电路设计.

陈 华(通讯作者) 男,(1986-),副研究员. 研究方向为耐辐照集成电路设计. E-mail:chenhua@zju.edu.cn