

语音识别中卷积神经网络的 FPGA 实现

屈家丽, 贺光辉, 王国兴

(上海交通大学 电子信息与电气工程学院, 上海 200240)

摘 要: 针对应用于语音识别中的卷积神经网络, 为了提高能耗比, 将网络在 FPGA 中进行定制化实现, 并通过 PCIe 完成了 FPGA 与主机的交互. 对该网络定点仿真结果表明, 整体采用 16 位定点计算可以在保证精度的情况下有效地减少数据存储量和带宽要求. 采用多种复用方式进行卷积层与全连接层的速度优化, 并与流水线结构相结合, 提高了系统的吞吐率, 在系统 150 MHz 的时钟频率下达到了 3 715 fps 的速度.

关键词: 语音识别; 卷积神经网络; 现场可编程门阵列; 硬件加速

中图分类号: TP39

文献标识码: A

文章编号: 1000-7180(2018)09-0037-05

FPGA Implementation of Convolutional Neural Network in Speech Recognition

QU Jia-li, HE Guang-hui, WANG Guo-xing

(School of Electronic Information and Electrical Engineering, Shanghai Jiao
Tong University, Shanghai 200240, China)

Abstract: A convolutional neural network used in speech recognition is customized in FPGA in order to improve the energy efficiency. PCIe is used to transfer data between the FPGA and PC. A 16-bit fixed data quantization is adopted after analysis and simulation. A variety of data multiplexing methods is used for speed optimization in convolutional layer and fully-connected layer. Pipe-line helps to improve the throughput of the system. This FPGA programs with 150MHz frequency and achieves the speed of 3715fps.

Key words: speech recognition; convolutional neural network; FPGA; hardware acceleration

1 引言

随着智能硬件的快速发展, 以语音技术为基础的语音识别技术越来越成为众人机交互形式中的一种重要手段^[1]. 卷积神经网络(CNN)^[2-3]通过将权值共享与降采样相结合的方式, 有效地减少了模型的复杂度与网络参数的数量^[4], 对于有局部相关性和平移不变性的输入有很好的识别作用, 由于语音信号同时具备这两个特性, 卷积神经网络被证明能够有效应用在语音识别中^[5]. 另一方面, 由于卷积神经网络的运算密集性, 如何高效快速地进行卷积神经网络的计算显得尤为重要. CPU 无法有效利用 CNN 的并行性进行计算, GPU 的出现使得计算速度更快^[6], 但其功耗较高. FPGA 具有高度并行和低

功耗等优点, 因此其能显著提高针对 CNN 计算的能效比.

本文利用 CNN 的高度数据重复性以及规整的运算方式等特点, 实现了一个语音识别中的卷积神经网络, 通过分析模型特点, 采用了流水线^[7]结构, 并且对网络各层进行定单独设计, 在 DE5a-NET 开发平台上对模型进行了定制化实现, 通过卷积层内复用、全连阶层共享控制模块、多个输入共同计算等方式, 在 150 MHz 的时钟频率下, 达到了 3 715 fps 的计算速度.

本文首先简单介绍了卷积神经网络中的计算, 然后在 FPGA 平台上实现一个语音识别中的 CNN 算法, 并且将实现结果与该网络在 GPU 上的运行结果进行了比较.

2 卷积神经网络模型基础

一个典型的 CNN 包括一系列按顺序计算的层,按照计算方式的不同被分为卷积层与全连接层.

2.1 卷积层与全连接层模型

通常,卷积层包含两个子层:卷积层与降采样层;卷积层通过将卷积核在输入层上滑动进行卷积计算后与该层的偏置相加得到该层的卷积结果,在一个卷积层内包含多个卷积核,每个卷积核与输入向量的卷积将作为该层卷积输出向量的一个平面,通常卷积层可以用如下表达式^[8]表示:

$$\begin{aligned} &O[z][u][x][y] \\ &= B[u] + \sum_{k=0}^{C-1} \sum_{i=0}^{R-1} \sum_{j=0}^{S-1} I[z][k][Uz + i][Uy + j] \\ &\times W[u][k][i][j] \\ &0 \leq z < N \\ &0 \leq u < M \quad E = (H - R + U) / U \\ &0 \leq y < E \quad F = (W - S + U) / U \\ &0 \leq x < F \end{aligned}$$

式中, M 代表卷积核数量; N 代表输入向量数量; C 代表输入层或卷积核的深度; $H/W/R/S/E/F$ 分别代表输入层向量、卷积核和输出层向量的高度和宽

度; U 代表将卷积核在输入平面上滑动一步的步长; $B[u]$ 为每一个卷积核的偏置.

降采样层又叫作池化层(Pooling),可以分为最大池化(Max Pooling)、均值池化(Average Pooling)等,本文实现的模型采用了最大池化,即在所选取数据范围内保留最大值作为输出,该操作能显著减小模型的体积与计算量.

对卷积层的输出进行非线性运算后作为下一层的输入,非线性运算并不会改变计算尺寸的大小,ReLU 函数为其中一种,表示方式如下:

$$\text{ReLU}(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

全连接层为向量与权值矩阵的乘积再与偏置相加后得到的结果与非线性运算相结合.若全连接层的输入尺寸为 $1 \times M$,输出层的尺寸为 $1 \times N$,则权值向量的尺寸为 $M \times N$.

2.2 本文实现的 CNN 拓扑结构

本文实现的卷积神经网络模型如图 1 所示.该网络是一个语音识别中声学模型的识别网络,整个网络共有 7 层,其中 2 层卷积层 5 层全连接层,各层包含的计算如下.

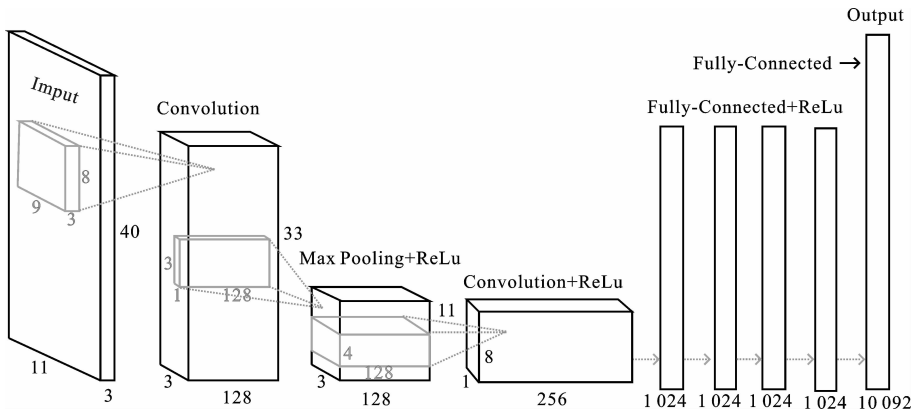


图 1 本文所实现的卷积神经网络模型

第一层卷积层输入层尺寸为 $40 \times 11 \times 3$,其中 40 为每一个声学单元提取的特征向量,11 表示 1 个中心帧与前后 5 帧的组合,3 表示当前帧的静态特征加一阶、二阶动态特征的组合.该向量与 128 个尺寸为 $9 \times 8 \times 3$ 的卷积核进行卷积运算后得到尺寸为 $3 \times 33 \times 128$ 的卷积结果,经过尺寸为 1×3 的最大值池化操作后再进行 ReLu 得到卷积层尺寸为 $3 \times 11 \times 128$ 的输出向量.

第二层卷积层的输入为第一层卷积层的输出,与 256 个尺寸为 $3 \times 4 \times 128$ 的卷积核进行卷积得到

尺寸为 $1 \times 8 \times 256$ 卷积结果,直接经过 ReLu 后得到第二层卷积层输出.

全连接层第一层的输入为第二层卷积层的输出,将其展开为尺寸为 $1 \times 1\,024$ 的列向量后与 $1\,024 \times 1\,024$ 的权值矩阵相乘得到一 $1 \times 1\,024$ 的列向量,ReLU 后得到该层的输出.

全连接层第二至四层结构相同,均是尺寸为 $1 \times 1\,024$ 的列向量输入与 $1\,024 \times 1\,024$ 的矩阵相乘得到 $1 \times 1\,024$ 的结果后经过 ReLu 得到输出.

第五层全连接层:尺寸为 $1 \times 1\,024$ 的输入列向

量与 1024×10092 的矩阵的相乘后加上对应偏置即得到整个网络的输出,尺寸为 1×10092 . 代表的是输入帧可能对应 10092 个声学单元中的每一个的概率。

3 卷积神经网络的 FPGA 实现

3.1 网络的整体分析

在 FPGA 中使用短的定点位数而非长的浮点数能有效减小对于存储空间和数据带宽的要求,但是过小的定点位数会导致准确率的严重下降^[6]. 为了在保证一定精度的要求下尽可能地压缩数据位数,本文首先在 Matlab 中进行了定点位数的仿真,将所有数据以不同的定点位数表示,通过输入测试集进行计算,统计在不同的位数下网络的准确率. 当定点位数大于 14 之后,准确率已经无明显上升趋势,达到与在 CPU 上运算相接近的准确率. 由于 2 的整数次幂的数据长度更适合在硬件系统中进行传输与计算,因此将数据位数定在 16 位。

整个网络在 FPGA 内部的实现采用流水线的方式,以此来提高系统吞吐率. 通过分析网络特点可知,卷积层以少量的数据消耗大量的计算量而全连接层需要消耗大量的存储空间,因此在网络的实现中将卷积层的卷积核参数与网络的偏置存储在片上存储单元中,将全连接层的权重参数存储在 DDR 存储器中。

图 2 所示为系统的整体结构图. 在数据准备阶段,在 PC 端将数据通过 PCIe 以 DMA 方式传输至 DDR. 待数据准备好之后,PC 将输入向量通过 PCIe 以 DMA 方式传输至片上存储单元,待得到计算结果之后,再通过 PCIe 传输回 PC 端。

图 3 所示为一个包含了能完成乘累加,池化和 ReLu 操作的数据处理模块的结构框图. 在 FPGA 内部通过控制逻辑(Control)向存储单元(Memory)发送数据读取命令,数据从存储单元送入数据处理模块,在处理模块内进行卷积层或全连接层的计算,计算完成之后将每层的输出数据写回至存储单元,作为下一层卷积层或全连接层的输入。

模块内各部分的主要作用如下:

1、DSP 阵列负责进行输入数据与权值的乘法;
2、加法树(Adder Tree)将前级的乘积进行累加,同时偏置经过移位对齐之后与累加结果相加. 至此在完成所有的乘累加操作。

3、池化与 ReLu 模块负责将前级得到的和进行进一步处理后将输出数据进行截位后写回存储

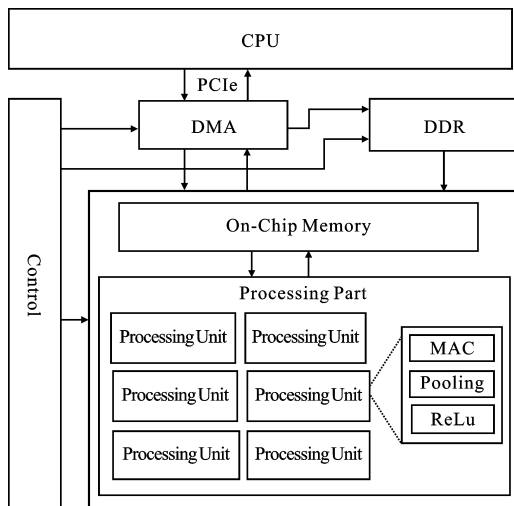


图 2 整体系统架构

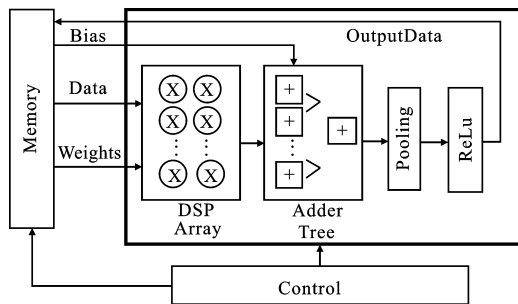


图 3 数据处理模块

单元。

为了能尽量提高能效比,在卷积计算中要着重利用这种计算方式的高度并行性与数据重复性,它们体现在:

- (1) 多个数据的乘加可以并行进行计算。
- (2) 若有 n 个输入向量,每个卷积核与全连接权值至少被重复使用 n 次。
- (3) 在一层卷积层的计算中,若有 m 个卷积核,则该层的卷积输入向量至少被重复利用了 m 次。
- (4) 在一次从输入至输出的计算中,为了得到一个卷积层的输出层,若输出层的尺寸为 $E \times F$,则每个卷积核被重复利用了 $E \times F$ 次。

在进行计算时若能充分用上述特点,则可以减少对存储单元的读取次数并且提高计算速度。

3.2 卷积层的实现

本文实现卷积层的方式如图 4 所示^[8],通过控制逻辑每次分别取输入层与对应卷积核的一列送入乘法器阵列中,将每一列乘法器得到的乘积相累加得到对应列的部分和(P_{SUM})。乘法器数量由输入向量和卷积核向量的行数决定,其中,乘法器阵列的行

数等于卷积核向量的行数,列数则等于输入向量的行数减去卷积核向量的行数加一,这样的数据映射将一系列输入向量中的数据以 L 形顺序排列在卷积核阵列的第一列与最后一行,再以对角线的方向进行数据的复用。

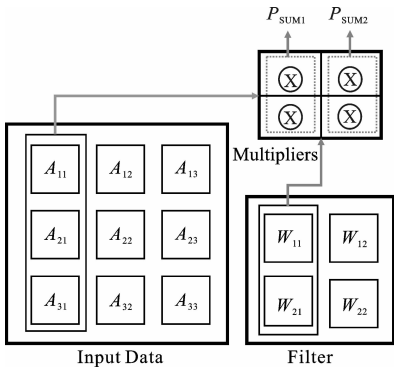


图 4 数据处理单元中的卷积模块

通过这样的数据映射,相当于只读取数据一次就达到了卷积核的一列数据在输入向量的列方向滑动的目的,实现了数据的重复利用。

以 3×3 的输入平面、 2×2 的卷积核平面为例,卷积核滑动步长为 1,卷积计算完成之后的输出为一个 2×2 的向量。卷积核行数为 2,输入向量行数为 3,则乘法器阵列尺寸为 2×2 ,如图 5 所示,将输入平面的第一列 A_{11} 、 A_{21} 与 A_{31} 沿乘法器阵列边沿送入乘法器阵列中,再以对角线的方式复用数据 A_{21} (由于 A_{11} 与 A_{31} 刚好处于左上与右下的顶点位置,不需要复用),卷积核平面的第一列 W_{11} 、 W_{21} 以水平复用的方式送入乘法器阵列,将得到的乘积以列的形式相累加得到两个部分和。

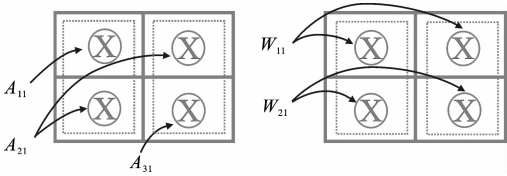


图 5 数据在乘法器阵列中的排列方式

图 6 所示为不同时刻乘法器阵列的输入及处理过程:

在 t 时刻:输入向量的第一列(A_{11} , A_{21} , A_{31})与卷积核的第一列(W_{11} , W_{21})被读取,经乘法器阵列计算后累加得到部分和 $P_{sum1,1}$ 、 $P_{sum2,1}$ 。

在 $t+1$ 时刻:将输入平面的第二列(A_{12} , A_{22} , A_{32})作为输入,卷积核部分保持上一时刻的数据,计算后得到 $P_{sum1,2}$ 与 $P_{sum2,2}$ 。

在 $t+2$ 时刻:保持输入平面的第二列数据,卷

积核的第二列(W_{12} , W_{22})被读取,将计算得到的部分和与 t 时刻得到的 $P_{sum1,1}$ 与 $P_{sum2,1}$ 相加得到 $sum_{1,1}$ 、 $sum_{2,1}$ 。

在 $t+3$ 时刻:将输入平面的第三列(A_{13} , A_{23} , A_{33})作为输入,卷积核部分保持上一时刻的数据,将计算得到的部分和与 t 时刻得到的 $P_{sum1,2}$ 与 $P_{sum2,2}$ 得到 $sum_{1,2}$ 、 $sum_{2,2}$,完成卷积计算。

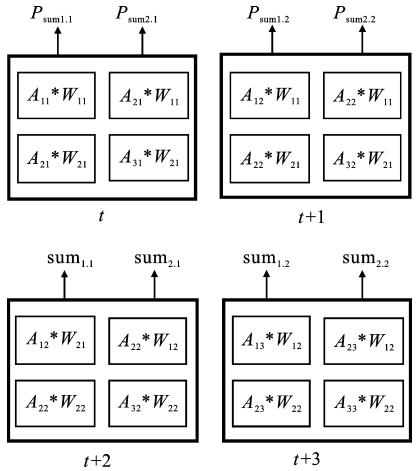


图 6 卷积乘法器阵列的处理过程

3.3 全连接层的实现

第二、三、四层全连接层具有相同的结构,即除了处理的数据不同,它们具有相同的处理过程,因此它们可以共享控制逻辑。

图 7 所示为第二到四层全连接层共享控制逻辑的模块结构,这三层被分配有各自的输入存储单元与权值存储单元,通过同一个控制模块发出统一的控制信号,向三个输入存储模块与权值存储模块读取数据,数据传输至数据处理模块进行全连接层的计算,当数据处理完成之后写入至各自的输出存储模块中,通过这样的方式节省了两层全连接层的控制逻辑所需要的硬件资源。

3.4 数据的重用

当有多个输入向量时,可以通过对计算参数与控制逻辑的重用来减少存储器的读取次数并降低资源利用。如图 8 所示为同时计算 2 组输入向量时的结构,2 个数据处理模块共享一组数据读取控制逻辑,每个数据处理单元均被分配有各自的输入存储模块,控制模块向权值存储器与输入存储器发出读取信号,读取数据之后输入数据传输至对应的数据处理单元,而权值参数则作为统一的输入进行复用,通过这样的方式每层将会进一步减少 50% 的控制逻辑所需要的硬件资源,同时计算速度将提高一倍。

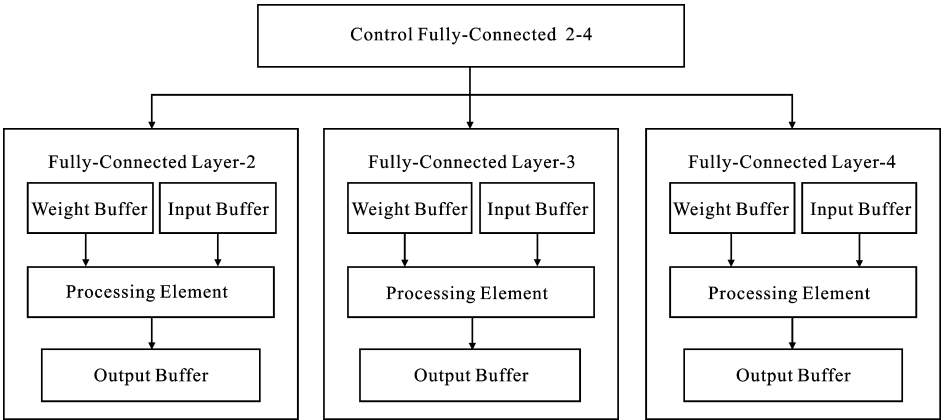


图 7 第二到四层全连接层共享控制逻辑

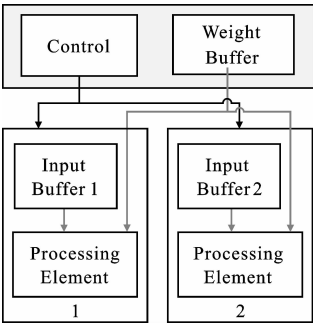


图 8 当输入向量个数为 2 时,参数与控制模块的重用

4 FPGA 资源利用总结与比较

本文使用 Intel(原 Altera)的 Quartus Prime 工具进行硬件开发与测试,使用 DE5a-Net 作为实现平台,该平台使用的 FPGA 为 Arria-10 115N2E1. 表 1 对比了不同的 CNN 在 FPGA 上实现的性能情况. 在时钟频率为 150 MHz 的情况下,本文所实现的 CNN 的吞吐率为 159.5GOPS.

表 1 与其他 CNN 在 FPGA 上实现的比较

	Ma ^[9]	Qiu ^[6]	Ours
FPGA	Stratix-V GXA7	Zynq-7000 XC7Z045	Arria-10 115N2E1
频率/MHz	100	150	150
数据精度/bits	8-16	16	16
网络参数数量/10 ⁶	7.59	50.18	15.99
吞吐率/GOPS	117.3	136.97	159.5

在表 2 中总结了 FPGA 电路板上的资源使用情况,从表中可以看出,FPGA 中的 DSP 消耗比例较大,其主要原因在于在卷积层的计算过程中需要利用多个 DSP 阵列进行并行计算.

表 2 FPGA 电路板上的资源消耗情况

	资源利用	比例/%
Logic Utilization	164 908	39
Memory Bits	17 166 264	31
DSP Blocks	984	65

该网络在 FPGA 平台上的运行速度为 3 715 fps,功耗仅为 15.43 W,在 Nvidia K40c GPU 上运行的速度为 1 900 fps,表 3 总结了网络在 FPGA 与 GPU 平台上的运行性能,从表 3 中可以看到,在速度方面 FPGA 的计算速度将近为 GPU 的 2 倍.

表 3 与 GPU 的性能比较

平台	计算速度/(f/s)
DE5a-Net	3715
K40c	1900

5 结束语

本文根据卷积神经网络的计算并行性与数据的高度重复性,实现了一种语音识别中的、基于 FPGA 的卷积神经网络加速方案. 通过分析定点位数对准确率的影响,采用 16 位定点数的数据精度. 通过卷积层内复用、全连接层共享控制模块、多个输入共同计算等方式,设计并优化了卷积层与全连接层的硬件映射方式,通过 PCIe 实现 FPGA 与 PC 的交互,在 150MHz 的频率下计算速度达到 3 715 fps,在网络结构相同的情况下,对 GPU 和 FPGA 进行了计算速度的比较,达到 GPU 速度的两倍.

(下转第 46 页)

tion, and computer applications[J]. Analog Dialogue, 2005, 39(10): 1-4.

[3] El Gharniti O, Kerherve E, Begueret J B. Modeling and characterization of on-chip transformers for silicon RFIC[J]. IEEE Transactions on Microwave Theory and Techniques, 2007, 55(4): 607-615.

[4] Cheng Ran, Chen Bo, Luo Danting, et al. Direct extraction method of equivalent circuit parameters for stacked transformer[J]. Journal Of Infrared And Millimeter Waves, 2016, 35(2): 172-176.

[5] 张风体. 基于片上变压器的数字隔离器的分析与设计[D]. 成都:电子科技大学, 2014.

[6] 冯勇, 刘诗斌, 李晓骏. 一种新型 CMOS 施密特触发器[J]. 计算机测量与控制, 2008, 16(11):1692-1694.

[7] 张风体. 一种基于片上变压器的数字隔离器编解码方案[J]. 微电子学与计算机, 2014, 31(1):164-167.

[8] Chen B, Wynne J, Kliger R. High speed digital isola-

tors using microscale on-chip transformers[J]. Elektronik magazine, 2003.

作者简介:

彭 锐 男, (1992-), 硕士研究生. 研究方向为功率驱动电路设计. E-mail: pengrui@ime. ac. cn.

蔡小五 男, (1979-), 博士, 研究员, 博士生导师. 研究方向为智能功率集成电路设计.

刘海南 男, (1979-), 硕士, 副研究员. 研究方向为集成电路的高可靠性设计.

罗家俊 男, (1973-), 博士, 研究员, 博士生导师. 研究方向为可靠性集成电路与器件设计技术.

赵海涛 男, (1983-), 硕士, 助理研究员. 研究方向为大容量存储器设计、智能功率集成电路设计等.

吴秀龙 男, (1979-), 博士, 教授, 博士生导师. 研究方向为超大规模集成电路设计.

(上接第 41 页)

参考文献:

[1] Cox R V. Speech and language processing for next-millennium communications services[J]. the IEEE, 2002, 88(8): 1314-1337.

[2] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 7553(521): 436-444.

[3] 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.

[4] 李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述[J]. 计算机应用, 2016, 36(9): 2508-2515.

[5] Qian Y. Very deep convolutional neural networks for noise robust speech recognition [J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2016, 24(12): 2263-2276.

[6] Qiu J. Going deeper with embedded FPGA platform for convolutional neural network[C]//Acm/sigda International Symposium on Field-Programmable Gate Arrays [s. l.]. 2016 :26-35.

[7] Gokhale V. A 240 G-ops/s mobile coprocessor for

deep neural networks[C]//Computer Vision and Pattern Recognition Workshops. USA Columbus, Ohio, 2014.

[8] Chen, Y H. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks [J]. IEEE Journal of Solid-State Circuits, 2016, 52 (1):262-263.

[9] Ma Y. Scalable and modularized RTL compilation of convolutional neural networks onto FPGA [C]//International Conference on Field Programmable Logic and Applications. South Africa, 2016:1-8.

作者简介:

屈家丽 女, (1993-), 硕士研究生. 研究方向为数字电路. E-mail: qujiali@sjtu. edu. cn.

贺光辉 男, (1980-), 博士, 副研究员. 研究方向为通信系统、VLSI、硬件加速等.

王国兴 男, (1975-), 博士, 副教授. 研究方向为生物医疗电子、集成电路设计.