

引用格式:翟社平,邱程,杨媛媛,等. 基于 FPGA 的卷积神经网络加速器设计与实现[J]. 微电子学与计算机,2019,36(8):83-86. [ZHAI S P, QIU C, YANG Y Y, et al. Design and implementation of convolution neural network accelerator based on FPGA [J]. Microelectroncs & Computer, 2019, 36(8): 83-86.]

基于 FPGA 的卷积神经网络加速器设计与实现

翟社平,邱 程,杨媛媛,李 婧,江婷婷

(西安邮电大学 计算机学院 陕西省网络数据分析与智能处理重点实验室,陕西 西安 710121)

摘 要: FPGA 已广泛用于卷积神经网络的硬件加速器的实现. 本文设计了一种基于 FPGA 的卷积神经网络加速器. 主要利用卷积神经网络中固有的并行性来减少实时嵌入式应用所需带宽与资源使用,并将其在有限资源的 ZC706 开发板上实现,结果显示,在 150 MHz 的工作频率下, FPGA 的峰值运算速度达到 0.54 GOP/s,且功耗很小.

关键词: 卷积神经网络;现场可编程门阵列;加速器;有限资源

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2019)08-0083-04

Design and implementation of convolution neural network accelerator based on FPGA

ZHAI She-ping, QIU Cheng, YANG Yuan-yuan, LI Jing, JIANG Ting-ting

(School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an 710121, China)

Abstract: FPGAs are widely used in the implementation of hardware accelerators for convolutional neural networks. This paper designs a FPGA-based CNN acceleration structure. It mainly utilizes the inherent parallelism in CNN to reduce the bandwidth and resource usage required by real-time embedded applications, and implements it on the ZC706 development board with limited resources. The results show that the FPGA operates at a frequency of 0.54 GOP / s and consumes very little power at 150 MHz.

Key words: convolutional neural network; field-programmable gate array; accelerator; limited resources

1 引言

卷积神经网络(Convolutional Neural Network, CNN)是一种典型的多层神经网络,随着深度学习技术的不断发展,其被广泛应用于机器视觉和语音分析领域^[1]. 传统的卷积神经网络是基于 CPU 执行计算,这样的计算不但缓慢低效,而且难以满足实时性的计算要求. 因此,基于 GPU 的卷积神经网络被广泛应用,文献[2]对基于 GPU 的 CNN 开源

项目进行比较,发现 GPU 普遍存在功耗大、成本高等问题.

FPGA(Field-programmable Gate Array)是一种可自定义编程的硬件电路结构,是一种主流的数字电路设计模式. FPGA 提供的并行运算的计算模式契合卷积神经网络的计算特点,同时, FPGA 可重编程的特点也适合于神经网络多变的网络结构. 因此,基于 FPGA 的 CNN 设计受到了广泛的关注,文献[3]和[4]分别提出一种基于 FPGA 的深度卷积

收稿日期:2018-11-19; 修回日期:2018-12-22

基金项目:工业和信息化部通信软科学项目(2018-R-26);陕西省社会科学基金(2016N008);西安市社会科学规划基金(17X63);西安邮电大学研究生创新基金项目(CXJJ2017062).

神经网络加速器设计,文献[5]设计了一款深度流水线 FPGA 集群实现高效 CNN.

本文首先介绍卷积神经网络 CNN,然后根据 CNN 的特点,通过合适的并行处理与流水线结构在 FPGA 上设计一个 CNN 系统,有效的提高其性能,并与已有的 CPU 与 GPU 实现进行比较.

2 基本概念

2.1 CNN 基本结构

卷积神经网络是深度学习中经典网络之一,其由多个卷积层、池化层和全连接层组成,图 1 是卷积神经网络的经典结构.卷积神经网络可以将待识别的图像直接作为输入,经过多个卷积层、池化层和全连接层,得出识别结果.

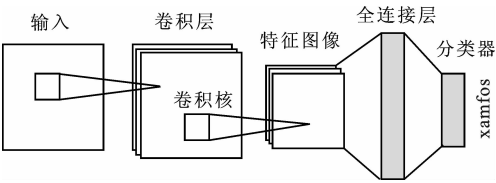


图 1 卷积神经网络结构

(1)卷积层模型

卷积层通过输入 f_j^{in} 与由权重 $w_{i,j}$ 组成的卷积核进行卷积操作,结果经偏置后获得输出,输出 f_i^{out} 即卷积的局部采样所得特征集合.卷积层的模型描述为

f_i^{\text{out}} = \sum_{j=1}^{n_{\text{in}}} f_j^{\text{in}} * w_{i,j} + b_i, \quad 1 \leq i \leq n_{\text{out}} \tag{1}

(2)池化层模型

池化层通常采用最大值采样或均值采样进行池化操作来降低输入矩阵的规模,其操作如式(2)所示.池化操作可有效降低下一层的数据处理量,同时避免特征信息丢失.

f_{i,j}^{\text{out}} = \max_{p \times p} (f_{m,n}^{\text{in}}, f_{m,n+1}^{\text{in}}, \dots, f_{m,n+p-1}^{\text{in}}, f_{m+1,n}^{\text{in}}, f_{m+1,n+1}^{\text{in}}, \dots, f_{m+1,n+p-1}^{\text{in}}, \dots, f_{m+p-1,n}^{\text{in}}, f_{m+p-1,n+1}^{\text{in}}, \dots, f_{m+p-1,n+p-1}^{\text{in}}) \tag{2}

(3)全连接层

全连接层对输入进行线性空间转换,从而得到输出

f^{\text{out}} = \sum_{j=1}^n f_j^{\text{in}} w + b \tag{3}

(4)激活函数

激活函数实现对输入激励的非线性转换,通常在每一层后对输出结果进行处理.常用的激活函数

包括冲击响应(Sigmoid)、非线性(ReLu)、三角函数(Tanh)等.

2.2 AlexNet

AlexNet 是一种典型的卷积神经网络,该模型共有 8 层,由 5 个卷积层和 3 个全连接层组成.在每个卷积层中都含有激活函数 ReLU 和局部相应归一化(LRN)处理,在卷积层 C1、C2、C5 中还有池化处理,具体网络参数见表 1.

表 1 AlexNet 网络参数

层	输入	卷积核	数量	输出
C1	227×227×3	11×11	96	27×27×96
C2	27×27×96	5×5	256	13×13×256
C3	13×13×256	3×3	384	13×13×384
C4	13×13×384	3×3	384	13×13×384
C5	13×13×384	3×3	256	6×6×256

在 AlexNet 当中,卷积层的运算量占了整个网络 90%以上,需要不断的取数据,保存处理结果,因此,可利用 CNN 的并行性来提高执行速度.

3 CNN 加速器设计

3.1 整体结构

基于 FPGA 的卷积神经网络通常需要消耗较多内存,但由于 FPGA 内资源有限,需要不断从外部读取数据,这大大影响网络效率.因此,加速的目标是减少对片外存储器访问次数,使输入图像和卷积核权值仅加载一次并存储在片上存储器中,直到执行完它所执行的所有操作.本文提出如图 2 所示的架构,主控制器为 PS 部分的 ARM 处理器,它通过高速 AXI 总线与 PL 部分的控制器相连接,控制器连接到数据存储器,将输入图像送到每个卷积模块,每个卷积模块具有用来存储权值的系数存储器.卷积模块使用来自数据存储器的输入图像和来自系数存储器的权值来进行卷积运算得到输出结果.

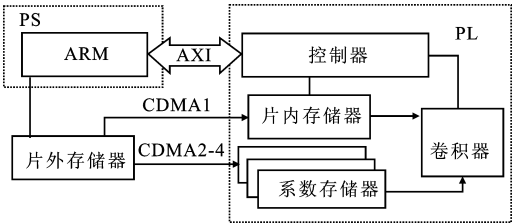


图 2 加速器整体架构

由于该结构适用于不同类型的卷积操作,每个卷积操作具有不同的参数,包括输入图像的大小,卷积核大小和数量,窗口移动的步长等,因此 PL 部分

的控制器需要对运行时的不同卷积核、权值大小和数量进行配置. 这些参数通过设置 ARM 处理器提供的某些寄存器发送到控制器,并由 PL 部分控制器计算必要的地址.

片外存储器是 1G 的 DDR SDRAM 存储器,PL 通过 DDR 存储器控制器提供的四个 AXI 高性能端口对其进行访问. 每个端口都有一个 AXI CDMA,用于在片外和片上存储器之间传输数据时降低 ARM 处理器负载.

3.2 PL 内部设计

在 PL 部分的卷积层设计中,数据和系数存储器的宽度均为 32 位,每行数据存储器用于存储一个数据值,系数存储器每行用于存储两个 16 位权值系数. 由于片上存储器是真正的双口 Block RAM,两行存储器每次都可以访问,因此在每个系数存储器中存储 4 个卷积核,前半部分和后半部分各存储 2 个,在每次运算中,卷积模块可以为 4 个不同的特征映射生成 4 个输出. 如图 3 所示,数据存储器的一个端口在 AXI CDMA 和控制器之间被多路复用,而且每个系数存储器的一个端口在 AXI CDMA 和相应的 MAC 模块之间被多路复用. 一个数据存储器端口用于将输入数据送至由控制器提供地址的系数存储器. 当卷积运算执行完成将结果送至其第二个端口并存储在数据存储器中. 由于所有的卷积模块都连接到这个单一端口,因此它们被一个由控制器控制的多路选择器复用. 输入数据的发送和每个输出数据的结果存储是同时完成,该循环不会被中断.

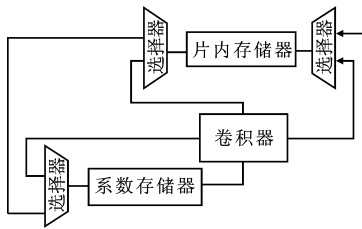


图 3 PL 内部结构

3.3 卷积运算单元

在硬件设计中,可实现不同卷积核或不同卷积层之间卷积运算的并行,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. 因此,要提高运算性能,就要增加数据的利用率和卷积运算并行性. 本文设计一种流水线并行乘加结构,在一

个时钟周期内可完成一个卷积核内的全部的乘加操作. 图 4 是卷积核大小为 3×3 的卷积运算单元,它能够在每个时钟周期内完成 9 次乘法运算,再经过加法器输出最终的输出特征像素. 寄存器 REG 中存放从内部存储器取到的输入图像信息,W 中存放从系数存储器取来的卷积核权值. 当每个卷积运算需要新的卷积核时,W 内的信息才进行修改. 对于需要进行填充操作的卷积神经网络而言,输入特征图的边缘位置需要填充 0. 因此在卷积器中添加了屏蔽器,将输出的边缘位置置 0. 该设计促使流水线不会因为数据填充或特征图的切换而中断,从电路层面保证了卷积计算的并行性.

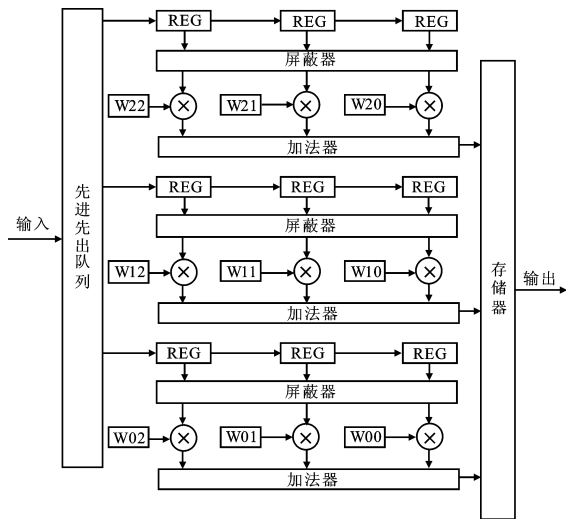


图 4 卷积运算单元

3.4 计算精度

卷积神经网络算法通常使用浮点运算,但在 FPGA 上实现浮点化运算将消耗大量的资源,因此需要选择合适的定点数. 通过多次实验对比发现,输入参数采用 16 位的定点数,权值数据采用 18 位的定点数,可获得较接近浮点数的结果.

4 实验结果及分析

4.1 实验方案

实验采用 Xilinx Zynq-7000 SoC ZC706 嵌入式 FPGA 实验平台,该系统搭载一颗 xc7z045 芯片,片内由 ARM A9 处理器与可重构逻辑部分构成,使用有限的片上资源 545 M 的 BRAM 和 400 个 DSP. 硬件实现环境为 Xilinx Vivado 2017. 对比实验中 CPU 采用 Intel Core i5 2500K 处理器,主频为 3.3 GHz,GPU 采用 NVIDIA GeForce GTX 960. 硬件实现环境为 Xilinx Vivado 2017.

最终设计包含 32 个卷积器,每个卷积器都配有一个 1.8 KB 系数存储器和 1.7 MB 的数据存储器,足以加速 AlexNet 的所有卷积层.由于片上 DSP 资源有限,每个卷积器包含 12 个 DSP 模块,用于在每次运行时生成 4 个输出像素,从而实现操作的最大性能.该设计包含四个通过高性能 AXI 接口连接到片外存储器的 CDMA.

4.2 实验结果

FPGA 资源消耗如表 2,本设计对 FPGA 内部资源进行高效率的使用,BRAM 全部使用,DSP 使用率达到 98%,文献[6]所使用 Virtex 7 开发板硬件资源极其丰富,若将本设计置于该型号 FPGA,将展现更好的性能.

表 2 资源消耗

模块	文献[6]	资源数
BRAM	1 024	545
LUT	28 000	77 442
DSP	2 688	391

将 CNN 在 FPGA 实现的性能与 Intel Core i5 CPU 和 NVIDIA GeForce GTX 960 GPU 做对比,结果见表 3.基于 FPGA 优化设计的卷积神经网络处理单张图片时间远少于 CPU,与 GPU 速度相当;FPGA 功耗大约是 GPU 功耗的二十三分之一,具有极高的能耗优势.

表 3 比较结果

平台	CPU	GPU	FPGA
时间(ms)	42.6	14.1	18.3
功耗(W)	45	235	10

表 4 是不同实现方式的性能对比.文献[7]是使用 ARM MaliT628GPU 优化 CNN 代码得到的结果,比较可知,我们的方法是文献[7]的 27 倍.

表 4 性能对比

设备	OP/S[GOP/S]
文献[7](GPU)	0.02
本文(FPGA)	0.54
性能提升	27×

5 结束语

本文提出一种基于 FPGA 有限资源的卷积神经网络加速器,利用 FPGA 内部资源,提高数据的

读取效率,并通过流水结构和提升卷积运算的并行度提高卷积运算速度,给出了卷积神经网络加速器内部结构,在资源有限的 ZC706 开发板上实现 AlexNet 网络,并与 CPU 和 GPU 进行对比,实验结果显示,嵌入式 FPGA 上的功耗及性能具有较大优势.

参考文献:

[1] 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.

[2] LI X, ZHANG G, HUANG H H, et al. Performance analysis of GPU-based convolutional neural networks [C]//International Conference on Parallel Processing. Philadelphia, USA: IEEE, 2016: 67-76.

[3] ZHANG C, LI P, SUN G, et al. Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks[C]//ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. Monterey, USA: ACM, 2015:161-170.

[4] 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. Monterey, USA: ACM, 2016: 26-35.

[5] ZHANG C, WU D, SUN J, et al. Energy-efficient CNN implementation on a deeply pipelined FPGA cluster[C]//Proceedings of the 2016 International Symposium on Low Power Electronics and Design. San Francisco, USA: ACM, 2016: 326-331.

[6] WANG C, YU Q, GONG L, et al. DLAU: A Scalable Deep Learning Accelerator Unit on FPGA[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2016, 36(3):513-517.

[7] LOKHMOTOV A, FURSIN G. Optimizing convolutional neural networks on embedded platforms with OpenCL[C]//Proceedings of the 4th International Workshop on OpenCL. Vienna, Austria: ACM, 2016: 10-14.

作者简介:

翟社平 男,(1971-),博士,副教授.研究方向为嵌入式系统、语义 Web.

邱程(通讯作者) 男,(1995-),硕士研究生.研究方向为嵌入式系统、深度学习. E-mail: qc_0305@126.com.