DOI:10.19344/j.cnki.issn1671-5276.2022.03.035

一种基于 FPGA 的高性能 MobileNet 加速器

周冠宇

(南京航空航天大学,江苏 南京 210016)

摘　要:针对轻量级卷积神经网络模型 MobileNet 现有 FPGA 实现版本中深度卷积与点卷积间计算等待时间过长、处理器与 FPGA 频繁通信导致计算效率较低、资源利用率不高等问题,提出一种针对 MobileNet 的 FPGA 优化设计,有效地提高了系统的实时性能和硬件加速单元的资源利用率。

关键词:现场可编程门阵列;MobileNet;并行计算;加速

中图分类号:TP183　　**文献标志码**:B　　**文章编号**:1671-5276(2022)03-0145-04

A High-performance MobileNet Accelerator Based on FPGA

ZHOU Guanyu

(Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: In view of the lengthy calculation waiting time between deep convolution and pointwise convolution of the existing lightweight convolutional neural network model MobileNet's existing FPGA implementation version and low computational efficiency and poor resource utilization caused by the frequent communication between the processor and the FPGA, an optimized FPGA design for MobileNet is proposed to improve the real-time performance of the system and the resource utilization of the hardware acceleration unit.

Keywords: field programmable gate array; MobileNet; parallelism computing; acceleration

0　引言

目前,深度学习在图像分类、目标检测、机器翻译、语音识别以及其他相关领域都取得了很多成果。卷积神经网络(convolutional nerual network)作为深度学习最具代表性的算法之一,近几年来也取得了极大的发展,一些新的优化算法和改良的卷积神经网络模型结构被不断提出。

MobileNet 是谷歌在 2017 年发布的一种轻量级卷积神经网络模型[1],与传统的卷积神经网络模型相比,它在保证精度损失不大的前提下,弥补了传统卷积神经网络模型过于巨大导致的无法应用于嵌入式设备的不足。目前基于 MobileNet 的现场可编程门陈列(field programmable gate array, FPGA)实现[2-5]主要集中在通过增加资源的使用量来提高分类速度,很少考虑硬件结构上的优化,对计算资源的浪费情况较为严重。针对上述不足,本文设计了一个高性能的 FPGA 硬件架构和加速单元,并与现有的 MobileNet 模型进行了性能及结果分析。

1　方法论述

MobileNet 模型的关键点在于使用深度可分离卷积代替了普通的卷积结构。得益于深度可分离卷积的特点,MobileNet 在精度损失不大的情况下,大大减少了整个网络中的参数量和计算量。

深度可分离卷积结构如图 1 所示。它可以被分为两个更小的操作:深度卷积和可分离卷积。深度卷积的输入特征通道数和输出特征通道数数目相等,各个输入通道相互独立。可分离卷积在计算方式上其实就是普通的卷积,只不过采用的是 1×1 的卷积核,因此它也被称为点卷积。

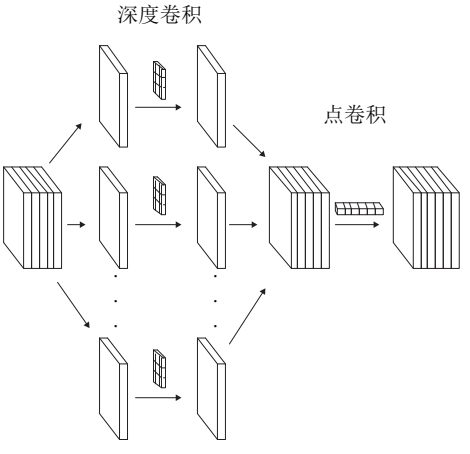

图 1　深度可分离结构

深度卷积的表达公式如式(1)所示。

$$G_{k,l,m} = \sum_{i,j} K_{i,j,m} \cdot F_{k+i-1,l+j-1,m} \tag{1}$$

式中:K 是深度卷积核权重;下标 i,j 表示作用在通道 m 上的卷积核像素点的位置;G 代表特征图输出;k 和 l 代表输出特征图的大小。

作者简介:周冠宇(1996—),男,山东枣庄人,硕士研究生,研究方向为智能仪器及自动化技术。

点卷积的表达公式如式(2)所示。

$$Y_{i,N} = \sum_{i=1}^{M} x_i \cdot W_{i,N} \tag{2}$$

式中:$Y_{i,N}$代表第 N 个输出通道第 i 个位置的点卷积值;x_i 代表输入第 i 个通道输入特征图上的像素点;$W_{i,N}$代表输出第 N 个通道对应于第 i 个输入通道的权重。

MobileNet 的基本结构如图 2 所示。标准卷积、深度卷积以及点卷积后都会紧跟批量归一化(batch normalization, BN)以及激活函数 ReLU 层。

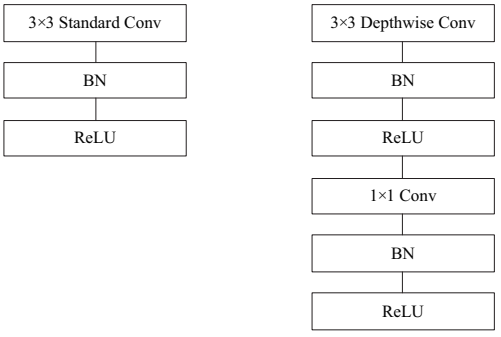

图 2　MobileNet 基本结构

BN 主要有加快卷积神经网络的训练和收敛速度、防止过拟合、控制梯度爆炸 3 个作用。它将所有样本上每个特征的值归一化为平均值为 0 且方差为 1 的数据。这使卷积值落入非线性函数有效值区域的中心,从而避免了梯度消失。在网络训练后完成,批量归一化层可由式(3)表示。其中 γ 与 β 值为常数。

$$y=\gamma x+\beta \tag{3}$$

激活函数给卷积神经网络内部带来了非线性,ReLU 函数如式(4)所示。

$$\mathrm{ReLU}(x)=\max(0,x) \tag{4}$$

池化层也被称为下采样层,其具体操作与卷积层的操作基本相同,只不过池化层的运算操作为只取对应窗口内的最大值、平均值等(最大池化、平均池化),即矩阵之间的运算规律不一样,并且不经过反向传播的修改,平均池化过程可用式(5)表示。

$$\mathrm{Pool}_n^{(l)}(i,j)=\mathrm{mean}(x_n^{(l)}(i,j)) \tag{5}$$

Softmax 函数在深度学习中负责将多个神经元的输出,映射到(0,1)区间内,输出的数值可以看成概率,从而来进行图像分类。函数公式如式(6)所示。

$$Y(x_i) = \frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}} \tag{6}$$

2　硬件加速

2.1　系统架构

本文设计的卷积神经网络加速系统采用“ARM+FPGA”的架构(图 3),ARM 侧除了初始配置以及最后对结果的处理外,没有参与到系统中间的运算以及数据传输中,这显著降低了系统运行过程中 ARM 与 FPGA 通信所需的时间,提供了更高的计算效率。

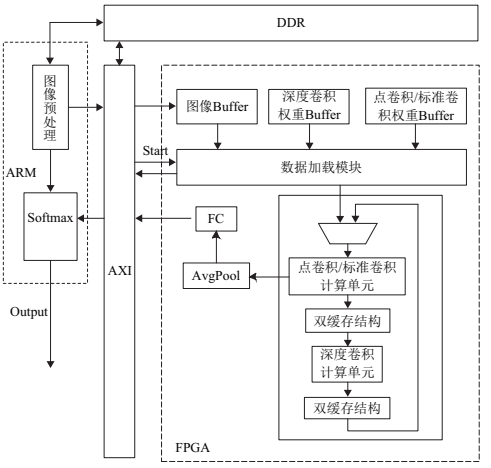

图 3　系统架构

卷积运算模块是整个硬件结构的核心,其内部由深度卷积运算单元和点卷积运算单元组成。由于 FPGA 内部资源有限,不能将整个 MobileNet 网络模型以平铺式的架构在 FPGA 上实现,而 MobileNet 各层之间的计算方式存在极高的相似性。因此采用分时复用单层计算资源的方式来实现整个网络。该模块中的深度卷积运算单元和点卷积运算单元的图像缓冲区域均为双缓冲结构:一个是工作缓冲,负责存储上一层网络的输出特征图;另一个是结果缓冲,负责存储本层网络的中间结果。

2.2　卷积运算单元

1)并行化深度卷积运算单元设计

每一个深度卷积计算单元 DWC Unit 负责一个输入通道的计算,其结构如图 4 所示。深度卷积的卷积窗口为 3×3,卷积窗口内各个像素点在进行乘累加时相互独立,无依赖关系。因此本文在空间上对该卷积过程展开,进行卷积核内部的并行处理。一个 DWC Unit 内包括 9 个乘法单元,mult 用来执行乘法操作,mult 间通过加法树连接,一个周期就可以计算得出一个像素点,之后计算出来的像素点经过 BN 层以及 ReLU 激活函数被送入双缓冲结构中进行暂存。多个 DWC Unit 同时进行多个输入通道的并行计算也可以加快网络层计算的吞吐量,本文设置 4 个 DWC Unit 同时计算,不同 DWC Unit 计算得出的结果被送至不同的片上缓存区域。

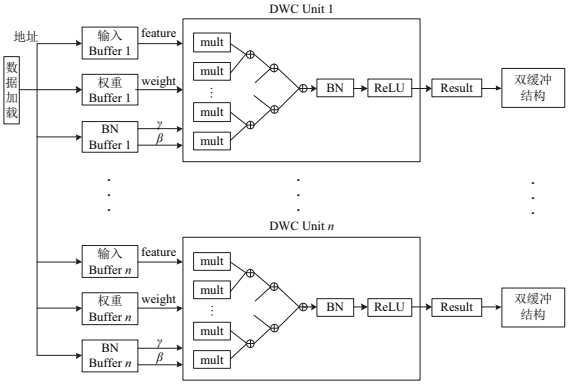

图 4　并行化深度卷积计算单元结构

2)并行化点卷积运算单元设计

从第二节的分析中可以看出,标准卷积层与点卷积层的结构类似,只是卷积核窗口大小不同。而整个 MobileNet 网络结构中只有第一层为标准卷积结构,如果单独设计标准卷积结构,那么在第一层计算完毕后,用于进行标准卷积的计算单元会进入空闲状态,造成较大的资源浪费。因此,本文设计了 PWC Unit 计算单元,使其能够实现标准卷积以及点卷积两种网络层的运算,节省了资源,同时优化了数据流向,更有利于流水线的形成。PWC Unit 计算单元结构如图 5 所示,每一个 PWC Unit 中只包含一个 mult 用来进行 3×3 或者 1×1 的卷积运算,不再考虑卷积核内部的并行处理。卷积结果经过 BN 和 ReLU 后经过加法树后送入数据分散模块。由于深度卷积层与点卷积层之间存在数据依赖性,因此深度卷积与点卷积计算单元的个数需要保持一致。而点卷积的结果计算后需要经过数据分散模块分散存储在不同的片上缓存区域中才能满足深度卷积计算单元的计算要求。

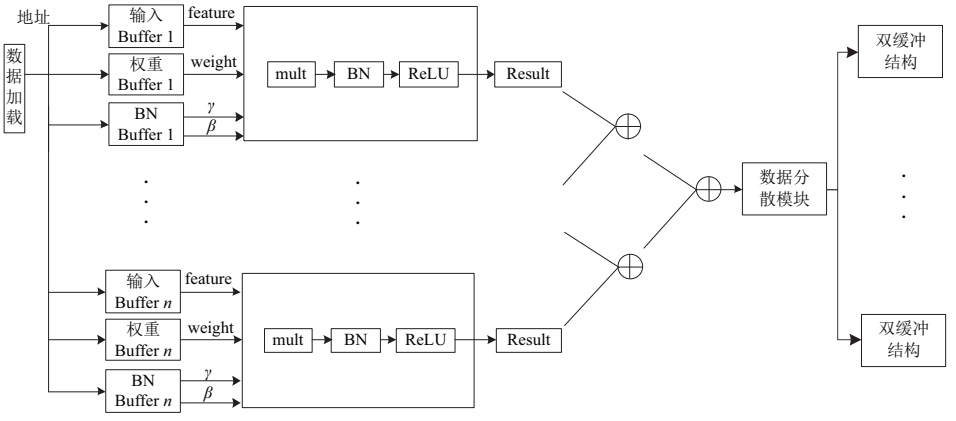

图 5　并行化点卷积计算单元结构

2.3　时序策略设计

本文时序策略分为两块,整个方案采用分时复用卷积计算模块的方式实现,网络模型中的各卷积层都通过一个卷积计算模块进行实现。而在深度卷积与点卷积计算单元之间,采用层间流水的方式进行两者的并行运算。当 ARM 侧向 FPGA 侧发出启动信号后,标准卷积开始进行计算。当标准卷积计算得出深度卷积需要的第一个卷积窗口时,深度卷积开始启动并与标准卷积同时进行运算。由于点卷积和标准卷积共用一个计算模块,因此当标准卷积层计算完毕,解除对计算模块的占用后,深度卷积和点卷积才会同时进行运算,在标准卷积计算完成之前,深度卷积的结果会暂存在片上缓存中。缓存区域均采用双缓存结构,当第 n 层深度卷积正在进行运算时,由于本文采用了层间并行运算的设计方式,此时点卷积计算单元正在进行 $n-1$ 层的运算,它的输入是 $n-1$ 层深度卷积的计算结果,因此深度卷积计算单元得到的结果需要被存放在结果缓冲中,点卷积计算单元需要的输入特征图则是从工作缓冲中读取,这种结构保证了当前层正在参与计算的数据结果不会被计算出来的输出特征图覆盖,同时减少了对外部存储的访问,提高了整个网络的运行速度。

3　实验结果

3.1　实验平台

本文采用的 PYNQ 板卡作为加速器验证平台,内嵌 Xilinx XC7Z020 FPGA 芯片。整个硬件系统架构为 ARM+FPGA 的异构架构。硬件开发使用的是某公司提供的 Vivado 2017.4 开发环境,软件开发环境为 Xilinx SDK 2017.4。

3.2　实验结果

在 100 MHz 工作频率下,本文设计的加速器资源使用情况如表 1 所示。通过分析可以发现 DSP48E 和 BRAM 是使用最多的两种资源,这主要是因为模型中存在大量的乘法运算以及中间结果被全部缓存在 BRAM。

表 1　资源使用数量

资源	资源总数	使用数	利用率/%
LUT	53 200	3 137	5.80
FF	106 400	6 651	6.25
DSP48E	220	132	60.00
BRAM	140	106	75.71

与其他文献的工作对比如表 2 所示。

表 2　本设计与现有成果对比

项目	文献[2]	文献[3]	本文
硬件平台	XCZU2EG	Stratix-V	Pynq-z2
网络结构	MobileNet v1	MobileNet v1	MobileNet v1
位宽	8bit fixed	8bit fixed	16bit fixed
DSP 使用数量	212	328	132
LUT	31 198	43 792	7 312
FF	46 809	41 507	14 651
BRAM	145	138	106
FPS	205.3	231.7	135.2

对表 2 分析可知,与文献[2]和文献[3]相比,本设计使用的资源较少,同时实现了较高的分类速度。

4　结语

本文针对 MobileNet 的网络结构,提出了深度流水化的加速器优化方案。设计了一种时序控制策略,使得深度卷积层和点卷积层能够同时计算来减少两者之间的计算等待时间;设计了一种系统架构来有效地减少处理器与 FPGA 的通信次数;设计了能够同时支持标准卷积与点卷积计算的计算单元,节省了资源。本文提出的加速器在工作频率为 100 MHz 的情况下,FPS 可以达到 135.2。

参考文献:

[1] HOWARD A G,ZHU M L,CHEN B,et al. MobileNets:efficient convolutional neural networks for mobile vision applications [EB/OL]. (2019-4-17)[2021-03-01] https://arxiv.org/abs/1704.04861.

[2] WU D,ZHANG Y,JIA X J,et al. A high-performance CNN processor based on FPGA for MobileNets [C]//2019 29th International Conference on Field Programmable Logic and Applications (FPL). Barcelona,Spain:IEEE,2019:136-143.

[3] SU J,FARAONE J,LIU J Y,et al. Redundancy-reduced MobileNet acceleration on reconfigurable logic for ImageNet classification [M]//Applied Reconfigurable Computing. Architectures, Tools, and Applications. Cham: Springer International Publishing,2018:16-28.

[4] LIAO J W,CAI L W,XU Y,et al. Design of accelerator for MobileNet convolutional neural network based on FPGA[C]//2019 IEEE 4th Advanced Information Technology,Electronic and Automation Control Conference (IAEAC). Chengdu, China: IEEE,2019:1392-1396.

[5] CHEN H Y,SU C Y. An enhanced hybrid MobileNet[C]//2018 9th International Conference on Awareness Science and Technology (iCAST). Fukuoka,Japan:IEEE,2018:308-312.

收稿日期:2021-03-10

(上接第 137 页)

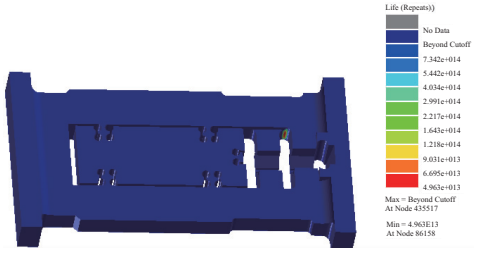

图 13　优化后疲劳分析图

由图 13 可知,优化后的柔性定位平台的最大等效应力为 4.11×10^{7}Pa,最小疲劳寿命为 4.96×10^{13},与优化前模型相比,等效应力和质量分别减少 2.4%和 6.3%;寿命提升了 27.2%,详细数据对比如表 3 所示。

表 3　优化前后数据对比

项目	质量/kg	等效应力/Pa	寿命(最小)
优化前	2.07	4.21×10^{7}	3.90×10^{13}
优化后	1.94	4.11×10^{7}	4.96×10^{13}

5　结语

柔性定位平台是超高加速度宏微运动平台的关键部件,提升其疲劳寿命对宏微运动平台的研究有很大意义。本文通过对柔性定位平台的分析与优化,大大减轻了柔性定位平台的质量,增加了其工作寿命。

1) 在 SolidWorks 中搭建三维模型,并导入到 Workbench 中进行静力学分析,实现 SolidWorks 和 Workbench 的联合仿真。

2)在 Workbench 中进行设置相关参数,把结果文件导入 nCode 中,在 nCode 中实现疲劳分析。

3)对初始模型进行拓扑优化,并根据拓扑优化结果对原始模型加以修整。修整后的模型质量大幅下降,寿命大幅上升。

参考文献:

[1] 马兵,张璐凡,吕彭民,等. 超高加速宏微运动平台研发设计动态[J]. 机械设计,2020,37(S1):19-23.

[2] XIE Y L,LI Y M,CHEUNG C F,et al. Design and analysis of a novel compact *XYZ* parallel precision positioning stage [J]. Microsystem Technologies,2021,27(5):1925-1932.

[3] 马兵,张璐凡,聂福全,等. 超高加速宏微运动平台振动能量分析[J]. 机械设计,2020,37(11):27-32.

[4] ZHANG L F,LI X L,FANG J W,et al. Vibration isolation of extended ultra-high acceleration macro-micro motion platform considering floating stator stage [J]. International Journal of Precision Engineering and Manufacturing, 2019, 20(8): 1265-1287.

[5] 张揽宇,高健. 高速大行程宏微复合运动平台的振动抑制与精密定位方法研究[J]. 机械工程学报,2020,56(11):131.

[6] 宋泽坤. 宏微复合精密运动平台的建模与控制方法研究[D]. 哈尔滨:哈尔滨工业大学,2020.

[7] 张金迪,高健,钟耿君,等. 新型三自由度宏微运动平台设计与仿真分析[J]. 现代制造工程,2019(8):125-129.

[8] 朱盼盼. 柴油机曲轴的疲劳分析与优化设计[D]. 长春:吉林大学,2016.

[9] SHIN W, CHANG K H, MUZAFFER S. Fatigue analysis of cruciform welded joint with weld penetration defects [J]. Engineering Failure Analysis,2021,120:105111.

[10] 白永明,邱恩举,王宏建. 基于 ANSYS 的连接件随机振动疲劳寿命分析及优化设计[J]. 兵器装备工程学报,2019,40(11):178-182.

[11] 孙新东,刘广璞. 十字轴万向节的拓扑优化和疲劳分析[J]. 机械工程师,2019(2):46-49.

[12] 卢剑,钟自锋. 发电机支架振动疲劳分析及其优化设计[J]. 机械强度,2019,41(5):1244-1248.

收稿日期:2021-02-22