

引用格式:武磊,魏子涵,张伟功,等. 一种通用 CNN 加速器的缓冲区地址调度方法[J]. 微电子学与计算机,2020,37(7):82-86. [WU L,WEI Z H,ZHANG W G,et al. A buffer address scheduling method for general CNN accelerator[J]. Microelectronics & Computer,2020,37(7):82-86.]

一种通用 CNN 加速器的缓冲区地址调度方法

武 磊^{1,3},魏子涵²,张伟功^{1,4},王 晶^{1,3},高 岚^{1,3}

(1 首都师范大学 信息工程学院,北京 100048;2 西安电子科技大学 电子工程学院,陕西 西安 710126;

3 电子系统可靠性技术北京市重点实验室,北京 100048;

4 北京成像理论与技术高精尖创新中心,北京 100048)

摘 要:针对由神经网络参数多计算复杂度高导致的数据访问灵活性不足问题,本文提出了一种通用 CNN 加速器的缓冲区地址调度方法.通过在地址控制器中灵活调度片上缓存,卷积层均能由一个或两个特定卷积运算单元实现,降低资源开销;通过输入参数可配置不同大小的图像及卷积核,具有一定的通用性.实验结果表明,该地址调度方法能够适用于不同数目及尺寸的卷积、池化和全连接操作,并且计算性能有所提高.

关键词:硬件加速器;地址调度;卷积神经网络;FPGA

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2020)07—0082—05

A buffer address scheduling method for general CNN accelerator

WU Lei^{1,3},WEI Zi-han²,ZHANG Wei-gong^{1,4},
WANG Jing^{1,3},GAO Lan^{1,3}

(1 College of Information Engineering,Capital Normal University, Beijing 100048, China;

2 College of Electronica and Information Engineering,Xidian University,Xi'an 710126,China;

3 Beijing Key Laboratory of Electronic System Reliability and Prognostics,
Capital Normal University, Beijing 100048, China;

4 Beijing Advanced Innovation Center for Imaging Theory and Technology,
Capital Normal University, Beijing 100048, China)

Abstract: Aiming at the problem of data access conflicts caused by multiple neural network parameters and high computational complexity, this paper proposes a buffer address scheduling method for general CNN accelerator. The on-chip buffer is flexibly scheduled in the address controller, and the convolutional layers can be implemented by one or two specific convolutional operation units to reduce resource consumption; the input parameters can be configured with images and convolution kernels of different sizes, with some universality. The experimental results show that the address scheduling method can be applied to different numbers and sizes of convolution, pooling and full connection operations, and the computing performance is improved.

Key words: hardware accelerator; address scheduling; convolutional neural network; FPGA

收稿日期: 2019—10—20; 修回日期: 2019—11—30

基金项目: 国家自然科学基金(61772350);北京市高水平教师队伍建设计划(CIT&TCD201704082);北京市科技新星计划项目(Z181100006218093);北京未来芯片技术高精尖创新中心科研基金(KYJJ2018008)

1 引言

近年来,深度学习发展迅猛,在图像识别、自然语言处理、视频跟踪等领域取得很多突出成果^[1].卷积神经网络(Convolutional Neutral Network, CNN)是目前应用较广的深度学习算法之一^[2].相较于传统算法, CNN 在计算及存储方面都具有较高的复杂性,通用 CPU 已无法满足计算需求.因此基于 GPU^[3]、ASIC^[4]以及 FPGA^[5]的神经网络硬件加速器设计得到了越来越多的关注. FPGA 凭借其并行性、可配置性等优势,成为目前很多研究人员研究 CNN 加速的硬件平台.

目前,大多数加速器设计^[6-9]是基于特定 CNN 模型的,面对当前迅猛发展的人工智能市场,通用 CNN 加速器的设计成为一种趋势.而在通用 CNN 加速器设计过程中,数据缓冲区地址调度是至关重要的.因此本文在研究了几种典型的 CNN 模型后,针对数据访问通用性不高问题,提出一种通用 CNN 加速器的缓冲区地址调度方法,能适用于不同尺寸大小的图像及卷积核,具有较强的普适性.同时减少中间结果存储,降低访存时间开销.

2 卷积神经网络

2.1 卷积神经网络介绍

卷积神经网络应用广泛,对各种模式识别问题有高度适应性. CNN 一般包括卷积层、池化层和全连接层.输入图像经过多个卷积层和池化层后进入全连接层,最后通过输出层得到最终的分类结果.

卷积层通过卷积运算实现局部特征的提取.卷积层的输入与由权重组成的卷积核进行卷积运算,运算结果加上偏置后得到输出,输出即为卷积局部采样所得特征集合^[10].目前典型的 CNN 模型均选用修正线性单元(Relu)函数作为激活函数.

池化层主要用于特征降维,压缩数据和参数的数量,减小过拟合,同时提高模型的容错性.目前 CNN 池化方式主要有平均池化和最大池化.

全连接层把所有的局部特征结合成全局特征,主要功能是分类^[11].

2.2 典型卷积神经网络模型

几种典型的 CNN 模型具体参数见表 1.由表格可以看出,这些网络模型结构类似,均为卷积、池化然后全连接.但是不同网络模型、不同层其卷积核大小数目不尽相同.

表 1 典型 CNN 模型

网络模型	层数			卷积核大小	池化方式	激活函数
	卷积	池化	全连接			
LeNet5	2	2	3	5×5	最大池化	Relu
AlexNet	5	3	3	11×11,5×5,3×3	最大池化	Relu
VGG16	13	5	3	3×3	最大池化	Relu
GoogleNet (Inception V1)	11	5	1	5×5,3×3,1×1	最大池化 平均池化	Relu
ResNet50	49	2	1	7×7,3×3,1×1	最大池化 平均池化	Relu

3 通用 CNN 地址变换单元

3.1 CNN 加速器整体分析

在只考虑前向传播的情况下, CNN 各层间的运算具有独立性且各层运算具有高度相似性,因此可以通过复用单层运算资源实现完整的 CNN 网络模型计算^[12]. CNN 单层计算单元流水线结构如图 1 所示.图像数据与权重经过预处理后存储在片外主存设备中,对数据进行计算之前,将其装载于片上缓存中.地址控制器 1 产生不同的地址分别从缓存 1 和缓存 2 读取数据与权重到卷积运算单元,卷积运算单元与 Relu 模块处理好的数据由地址控制器 2 控制其存入缓存 3,并在存储后通过控制地址的变

化将数据送入池化模块,地址控制器 3 将池化结果存入缓存 4 中并将其送入下一运算模块.顶层控制器控制整体的地址调度与运算.

3.2 存储方式

CNN 中涉及的操作数一般默认为浮点数,但目前大多使用 16 位或 16 位以下定点数进行处理,本文采用 16 位定点数.

考虑到输入图像数据大多为三通道以及系统整体的并行度,故将存在缓冲区的权重值与特征值均按三通道并行的方式存储.缓存 1 与缓存 2 分别存放权重值和像素值,本文采用图 2 所示的方式进行权重值存储, ki_mn 表示第 i 个卷积核 n 通道的第 m 个权重值.每个地址对应的是一个卷积核三个通

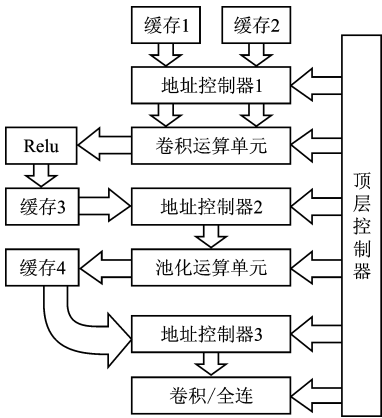


图 1 CNN 单层流水线结构

道同一位置的权重值。这样,在计算过程中,权重值存储缓冲区可以同时输出三个权重值。像素值存储方式与权重值类似,缓存 3 存储卷积运算结果,将同时输出的六个卷积运算结果并行存储。缓存 4 存放池化运算结果,存储方式与缓存 3 类似。

addr	47	31	Width	15	0
0	k0_02	k0_01	k0_00		
1	k0_12	k0_11	k0_10		
2	k0_22	k0_21	k0_20		

图 2 权重值存储方式

3.3 卷积模块的地址调度

在 CNN 中,卷积占整体运算的 90% 以上。特征图尺寸、卷积核尺寸以及卷积核数目的不确定使数据调度过程更复杂。卷积模块的地址调度可分为两个部分,一个是运算前的数据读取,另一个是运算后的结果存储。

在考虑并行计算的情况下,将整个卷积运算划分为若干个卷积运算基础单元,既可同时并行处理多个基础单元,也可通过复用基础单元实现不同输入输出通道的卷积运算。每个卷积运算基础单元中有 6 个 PE 单元,一次可以同时并行处理六个核。每个 PE 单元均由三通道并行输入,三通道运算结果相加后输出,其中包括 18 个实现乘法运算的 DSP 单元以及 24 个加法运算单元。

对应于运算单元,运算前的地址调度可分为三步:(1)循环遍历卷积核,直到当前特征图遍历结束;(2)若进行输入通道复用,则将缓冲区地址设为下一通道的起始地址;(3)若进行输出通道复用,在输入通道复用结束后,将存放特征值的缓冲区地址置为 0,存放权重值的缓冲区地址设为下一组卷积核的起始地址。在卷积核循环遍历阶段,地址调度过程适用

于不同尺寸的输入图像及卷积核,尺寸与复用通道数均可由输入端口赋值改变。

卷积运算的具体过程如图 3 所示。按照卷积运算规则控制缓冲区地址变化从而获得数据。 X_{ij} 表示输入特征图的第 i 行 j 列元素, W_{ji} 表示第 i 个核的第 j 个权重值。卷积核循环遍历一次即得到一个输出结果,然后按照先行后列的顺序依次遍历特征图。 C_{ij}^t 表示第 t 个特征图的第 i 行 j 列元素。若通道复用,则该结果为部分和结果。所有复用过程结束,即所有部分和结果存储完成后,然后控制地址变化进行部分和结果相加并存储。

Input Vectors

$$\begin{pmatrix} X_{00} & X_{01} & X_{02} & X_{10} & X_{11} & X_{12} & X_{20} & X_{21} & X_{22} \end{pmatrix} \cdot W = \begin{bmatrix} C_{00}^0 \\ C_{01}^0 \end{bmatrix}$$
$$\begin{pmatrix} X_{01} & X_{02} & X_{03} & X_{11} & X_{12} & X_{13} & X_{21} & X_{22} & X_{23} \end{pmatrix} \cdot W = \begin{bmatrix} C_{10}^0 \\ C_{11}^0 \end{bmatrix}$$
$$\begin{pmatrix} X_{10} & X_{11} & X_{12} & X_{20} & X_{21} & X_{22} & X_{30} & X_{31} & X_{32} \end{pmatrix} \cdot W = \begin{bmatrix} C_{10}^0 \\ C_{11}^0 \end{bmatrix}$$
$$\begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{21} & X_{22} & X_{23} & X_{31} & X_{32} & X_{33} \end{pmatrix} \cdot W = \begin{bmatrix} C_{11}^0 \\ C_{12}^0 \end{bmatrix}$$

Output Elements

$$W = \begin{bmatrix} W_0^0 & W_1^0 & W_2^0 & W_3^0 & W_4^0 & W_5^0 & W_6^0 & W_7^0 & W_8^0 \end{bmatrix}$$

图 3 卷积具体运算过程

部分和结果相加并存储的具体地址调度过程如图 4 所示。整体过程由两部分构成,一部分是输入通道复用,复用次数 N 为得到一个最终结果需要部分和结果相加的次数。循环次数小于 N 时,地址加卷积运算后特征图尺寸值,将对应位置的部分和结果取出并相加。循环次数达到 N 后,存放最终结果的缓冲区写使能置 1,地址加 1,存储部分和结果的缓冲区地址变为下次部分和结果相加过程的起始地址,直至当前特征图遍历结束。另一部分为输出通道复用,在输入通道复用结束后,进行输出通道复用。输出通道复用次数 M 为卷积核个数与基础单元一次处理卷积核个数的比值。循环过程为输入通道复用过程,当循环次数小于 M 时,地址为下一次输入通道复用的起始地址。循环次数达到 M 后,所有复用过程结束。

3.4 池化模块的地址调度

本文设计的池化模块支持 2×2 以及 3×3 窗口的最大池化。该模块的地址调度过程分为两部分:卷积运算结果读取以及池化运算结果存储。对应于卷积模块,该模块由 6 个独立的池化单元组成。

以 2×2 窗口最大池化为例, X_{nm} 表示第 i 行 j 列的数据。按照最大池化运算规则,对存放卷积最终结果的缓冲区进行地址调度,依次取 X_{00} 、 X_{01} 、 X_{10} 、

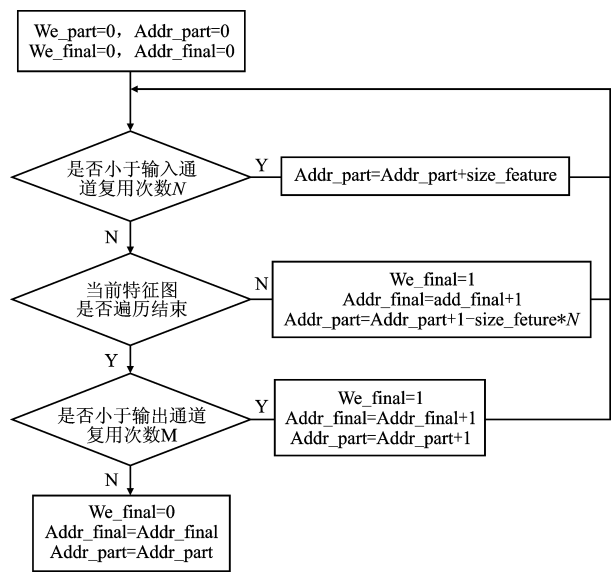


图 4 卷积结果存储过程地址调度

X_{11} ,再取 X_{02} 、 X_{03} 、 X_{12} 、 X_{13} ……以此类推,两行相间取数,直至当前特征图的特征值取尽.始终对当前寄存器组内的四个特征值进行大小比较,输出最大值.每四个周期得到一个结果,得到结果时,将存储池化结果的缓冲区写使能置 1,地址加 1.其余的时刻写使能置 0,地址不变.

3.5 全连接模块的地址调度

全连接层采用串行运算的方式,同样地,整体地址调度过程分为数据读取与结果存储两部分.将存

放上一层运算结果的缓冲区写使能置 0,地址从 0 开始依次加 1 加到本层通道数 T .当地址加到 T 时,得到一个全连接层结果,同时将缓冲区的地址置 0,循环上一过程,循环次数为下一层通道数.

根据全连接运算过程,每 T 周期得到一个结果,得到结果时,缓冲区的写使能置 1,地址加 1.其余的时刻写使能置 0,地址不变.

4 实验结果与分析

本文采用 LeNet5、AlexNet、VGG 三种典型的 CNN 模型对本文的缓冲区地址调度方法进行验证测试,同时对文献[8]所用网络模型进行实现及性能评估.实验使用 Xilinx 公司的 ISE 软件进行硬件仿真分析,器件选取的是 Virtex-7 系列的 xc7vx690t. LeNet5 采用 MNIST 测试集中的图像像素作为卷积神经网络的输入特征图,ImageNet 数据集图像像素用于 AlexNet 与 VGGNet 的输入.实验过程中,验证每一层操作输出的结果,确保实验的准确.

本文对 LeNet5 的卷积层、AlexNet 的前两个卷积层以及 VGGNet 的第一个卷积层进行实验,仿真结果与预期结果一致.具体参数与资源占用情况如表 2 所示.可以看出本文所用缓冲区地址调度方法能够适用于不同神经网络的不同卷积层,在步长、卷积核大小、数目方面具有较强灵活性.

表 2 卷积层实验参数与资源占用

实验序号	1	2	3	4	5
网络模型	LeNet5		AlexNet		VGGNet
卷积核大小	5×5		11×11	5×5	3×3
卷积核个数	6	16	96	256	64
步长	1	1	4	1	1
操作总数	169 344	301 056	209 959 200	877 879 296	163 774 464
执行周期	14 400	4 992	5 904 800	13 154 805	6 021 120
DSP	22	36	22	36	22
BRAM	16	25	324	1 064	1 352

池化操作选用 LeNet5 和 AlexNet 的第一个池化层进行验证. LeNet5 采用 2×2 窗口、步长为 2 的最大池化,AlexNet 池化窗口为 3×3,步长为 2.得到如表 3 所示的资源消耗结果.结果表明本文提出的方法能够满足不同窗口大小的池化操作.

选取 LeNet5 的三个全连接层作为全连接操作的验证实验.其神经元数量分别为 120、84 和 10.实验结果证明本文提出的方法能够适应不同神经元数目的全连接操作.

表 3 池化层实验资源消耗

网络模型	LeNet5	AlexNet
操作总数	2 592	559 872
执行周期	576	104 976
DSP	5	5
BRAM	7	286

由于 DSP 是实现乘加运算的主要方式,因此本文主要以单个 DSP 执行效率来衡量计算性能.单个 DSP 执行效率=操作总数/执行周期数/DSP 数.其

中操作总数的计算方法以文献[13]给出的计算公式为参考.表 4 给出了本文与文献[8]实现 CNN 的性能比较.

表 4 实现 CNN 的性能比较

参数	文献[8]	本文
操作总数	402 154	402 154
执行周期	3 815	19 100
DSP	638	82
BRAM	0	50
计算性能	0.165	0.257

综合以上实验结果,可以看出本文提出的地址调度方法具有较高的灵活性,能够适用于不同网络模型的不同层.同时,采用本文方法的 CNN 加速器在计算性能上有所提升.

5 结束语

针对数据访问普适性的不足问题,本文在研究了几种典型的 CNN 模型后,提出一种通用 CNN 加速器的缓冲区地址调度方法,综合考虑通用性和并行性,通过灵活的地址调度,动态适配不同大小的图片与卷积核.本文对多个网络模型的某些层进行实验,结果表明该方法能够在保证计算性能的前提下,具有较好的灵活性.

参考文献:

[1] RAJARAMAN S, CANDEMIR S, KIM I, et al. Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs[J]. Applied Sciences, 2018, 8 (10):1715.

[2] CHEN Y, LUO T. DaDianNao: A machine-learning supercomputer[C]// The 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO). Cambridge, United Kingdom, IEEE, 2014: 609-622.

[3] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[4] CHEN Y H, KRISHNA T, EMER J S, et al. Eyerriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks[C]// IEEE International Solid State Circuits Conference. San Francis-

co, CA, USA, IEEE, 2016:262-263.

[5] 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.

[6] MA Y, CAO Y, VRUDHULA S, et al. Optimizing the convolution operation to accelerate deep neural networks on FPGA [J]. IEEE Transactions on Very Large Scale Integration Systems, 2018, 26 (7): 1354-1367.

[7] ZHANG J, LI J. Improving the performance of Open CL-based FPGA accelerator for convolutional neural network[C]// Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. Monterey, CA, USA, ACM, 2017:25-34.

[8] ZHOU Y, JIANG J. An FPGA-based accelerator implementation for deep convolutional neural networks [C]// International Conference on Computer Science and Network Technology. Changchun, China, IEEE, 2016:829-832.

[9] ALWANI M, CHEN H, FERDMAN M, et al. Fused-layer CNN accelerators[C]// The 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO). Taipei, Taiwan, IEEE, 2016:1-12.

[10] 卢冶,陈瑶,李涛,等.面向边缘计算的嵌入式 FPGA 卷积神经网络构建方法[J]. 计算机研究与发展, 2018, 55(3):551-562.

[11] 王巍,周凯利,王伊昌,等.卷积神经网络(CNN)算法的 FPGA 并行结构设计[J]. 微电子学与计算机, 2019, 36(4):57-62.

[12] 余子健,马德,严晓浪,等.基于 FPGA 的卷积神经网络加速器[J]. 计算机工程, 2017, 43(1): 109-114.

[13] 方荣强,王晶,姚治成,等.多层神经网络算法的计算特征建模方法[J]. 计算机研究与发展, 2019, 56(6): 1170-1181.

作者简介:

武 磊 女,(1996-),硕士研究生.研究方向为计算机系统结构.

魏子涵 男,(1999-).研究方向为嵌入式系统设计.

张伟功(通讯作者) 男,(1967-),博士,研究员.研究方向为高可靠计算机体系结构及 SoC 设计、高速高可靠总线技术、计算机容错技术究方向. Email:zwg771@cnu. edu. cn.

王 晶 女,(1982-),博士,副研究员.研究方向为计算机体系结构、高性能计算、容错计算、智能芯片设计.

高 岚 女,(1987-),博士,讲师.研究方向为计算机体系结构、GPU 体系结构及编程优化、并行编程模型.