

基于 FPGA 的卷积神经网络并行加速结构设计

刘志成^{1,2}, 祝永新¹, 汪 辉¹, 田 犁¹, 封松林¹

(1 中国科学院 上海高等研究院 CMOS 图像传感器实验室, 上海 201210;

2 中国科学院大学, 北京 100049)

摘 要: 本文根据卷积神经网络特点, 提出了一种基于 FPGA 的流水线并行加速方案, 设计优化了卷积模块电路、激活模块电路以及下采样模块电路, 从而构建了卷积神经网络运算的 FPGA 基本单元. 在网络结构和处理数据相同的情况下, 50 MHz 频率的 FPGA 计算效率为 CPU 的 8 倍、GPU 的近 5 倍, 而功耗则只占 GPU 的 27.8%.

关键词: 卷积神经网络; FPGA; 卷积模块; 激活模块; 下采样模块

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2018)10-0080-05

Parallel Acceleration Design of Convolutional Neural Network Based on FPGA

LIU Zhi-cheng^{1,2}, ZHU Yong-xin¹, WANG Hui¹, TIAN Li¹, FENG Song-lin¹

(1 Shanghai Advanced Research Institute, Chinese Academy of Sciences, CIS Lab, Shanghai 201210, China;

2 University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: According to the characteristics of convolutional neural network, this paper proposes a pipeline parallel acceleration scheme of FPGA. Convolution module circuit, activation module circuit and down-sampling module circuit are designed to construct the FPGA basic unit of convolution neural network operation. With the same network structure and processing data, FPGAs with 50 MHz frequency are 8x and nearly 5x computational efficiency of the CPU and the GPU, while power consuming only 27.8% of the GPU.

Key words: convolutional neural network; FPGA; convolution module; activation module; down sampling module

1 引言

近些年随着硬件计算能力的不断提高, 深度学习在图像分类、目标检测^[1]以及目标跟踪^[2]等研究领域得到了广泛的应用. 其对于图像数据的处理主要采用卷积神经网络^[3](CNN)来实现, 因此该网络结构模型在深度学习中占有重要的地位. CNN以计算密集和资源消耗大著称, 包含了大量独立重复的乘法、加法运算以及下采样和激活等数据处理过程, 目前主要依托 CPU 或者多块 GPU 搭建的运算平台进行开发. 随着应用需求的逐渐多样化^[4-5], 不仅网络复杂度逐渐攀升, 便携性的要求也日益增

高^[6]. 然而 CPU 由于串行执行无法满足 CNN 运算中的并行性要求^[7], 性能难以提升; GPU 则存在着体积大和功耗大的劣势, 无法将 CNN 应用在小尺寸、低功耗的嵌入式平台上. 面对这种情况, 现场可编程门阵列(FPGA)具有大量并且丰富的阵列形式的逻辑和运算单元, 能以较低功耗高度并行地执行 CNN 中的密集运算. FPGA 在尺寸, 功耗以及并行运算方面具有的显著优势, 满足目前对 CNN 运算的要求, 也符合未来的发展趋势. 因此, 研究 FPGA 实现成为 CNN 研究的热点话题^[8].

本文通过对 CNN 算法的详细分析, 将其主体运算层次划分为三个阶段——卷积、非线性激活以

收稿日期: 2018-01-03; **修回日期:** 2018-02-01

基金项目: 国家重点研发计划项目(2017YFA0206104); 上海市科学技术委员会科研计划项目(16511108701); 张江管委会公共服务平台项目(2016-14)

及下采样,设计了与之相对应的FPGA功能模块并进行了系统整合.

2 卷积神经网络模型介绍

卷积神经网络由于不需要对图像进行繁琐的预处理,能自主学习图像特征并且有较强鲁棒性,因此近些年在图像处理领域得到了广泛的应用.其网络结构主要包括卷积层、激活层以及下采样层,下面对三个功能模块依次进行介绍.

2.1 卷积层

卷积层是卷积神经网络中对图像特征的提取操作,也是该网络中计算量最大的部分.其示意图如下图1所示.

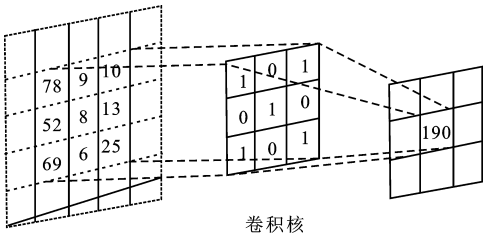


图1 卷积操作示意图

而其具体操作即为卷积层中的每个神经元与上一层的局部感受野相连进行像素的加权求和,其计算公式如下:

X_{ij} = \sum_{i \in M_j} X_i^{l-1} * W_{ij}^l \tag{1}

式中, X_j^l 表示第 l 层中的第 j 个特征图; X_i^{l-1} 为上一层的图像输入; W_{ij}^l 为卷积核的权值. 每当卷积窗口完成一次卷积计算时,都会向右移动一个步长,一般步长取值为 1. 当卷积窗口移动至行末时,会向下移动一行并从左侧行始开始新的一行卷积直到完成整个图像像素的卷积运算. 可以发现其中存在高度并行化的可能,借助FPGA的流水线并行结构设计就可以在一个周期内完成一个卷积窗口内的全部乘法运算,并在最短时间内完成后续累加运算.

2.2 激活层

由卷积层介绍可以看出其乘加运算为线性运算,而线性模型变化性不强,表达力不够,故需要采用非线性函数处理从而加入非线性变化. 因此,激活层的本质就是借助非线性函数增添非线性因素. 如式(2)所示:

X_j^l = f(\sum_{i \in M_j} X_i^{l-1} * W_{ij}^l) \tag{2}

式中, $f(x)$ 为激活函数,而目前激活函数多采用以下三种:

Sigmoid 函数:

y = \frac{1}{1 + e^{-x}} \tag{3}

Tanh 函数:

y = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{4}

ReLU 函数:

y = \max(0, x) \tag{5}

由于 ReLU 函数相比另两个函数具有更快的收敛速度,更简易的实现以及缓解了梯度消失问题的诸多优势,因此被广泛使用. 本文也选择了 ReLU 作为激活函数.

2.3 下采样层

卷积神经网络由于自身特性导致特征向量的数目众多,极易发生过拟合的情况. 从人体视觉系统使用更高层次抽象表示图像特征达到降维目的受到启发,构建下采样层对输入特征图进行压缩以及提取主要特征. 一般有最大值下采样以及均值下采样两种方式,目前常用最大值下采样. 本文也选择最大值下采样方式,即从下采样窗口中挑选最大值进行特征抽象,下采样窗口尺寸一般为 2×2 ,如图2所示.

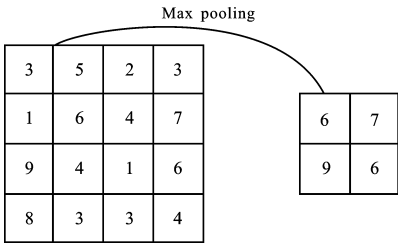


图2 最大值下采样示意图

3 功能模块的FPGA实现

卷积神经网络中存在大量独立重复并且可以并行执行的运算,具体来说网络的一层中从每个输入特征映射图到对应的输出特征映射图之间的计算都是可以并行执行的,因此也更适合设计成FPGA电路的功能模块. 故下面依次介绍卷积、激活、下采样三类最基本运算的电路设计以及模块之间的互联互通.

3.1 卷积模块

卷积运算是卷积神经网络中的核心运算,其包含了大量乘法与累加运算,消耗了大量的计算资源. 故需要利用流水线并行运算的思想,实现每个时钟能产生一个卷积结果的吞吐量.

本文为此提出如下设计思路:即已知卷积核尺

寸为 $K \times K$, 则将乘法运算分为等量的 K 组, 而每组的乘法和累加运算独立完成, 最后再将 K 组乘法累加的结果并列相加即可. 以 3×3 卷积为例, 电路设计即如图 3 所示.

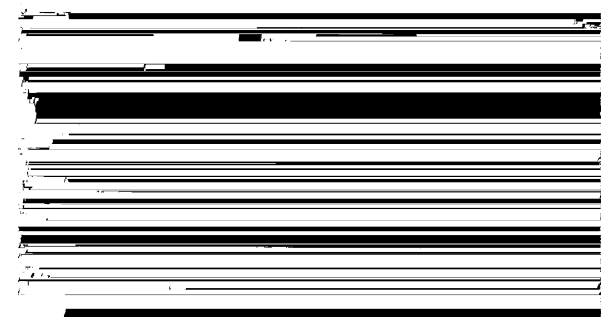


图 3 卷积运算电路示意图

如图 3 所示, 像素数据以数据流的形式依次流过三个移位寄存器; 每个移位寄存器缓存图像一行像素数据且末端连接三个寄存器用来缓存 3×3 卷积窗口中的像素数据; 待数据填满则并行计算 9 个乘法运算, 最后将每行乘法结果分别累加再并行相加即可得到一个卷积窗口的运算结果. 后续每移动一位像素就可以将卷积窗口向右移动一位. 这样的设计相比 CPU 使用了更多计算资源, 同时也大幅提高了计算效率. 即一个周期可以实现 9 次乘法运算和 8 次加法运算, 与 CPU 需要 17 个周期才能完成运算相比, 单个卷积运算就可以加速 17 倍. 可以预知当卷积核越大或流水线越深, 此设计并行加速的效果就越显著.

上述电路要完成预想功能还需要设计有限状态机进行控制, 考虑到与后续下采样模块的交互, 因此状态共分为以下四个模式:

(1) 预填充: 每个移位寄存器缓存原图像一行像素, 则当三个移位寄存器缓存原图像前两行像素外加第三行部分像素使得第一个卷积运算窗口布满像素则完成了预填充阶段;

(2) 奇数行: 由于采用最大值下采样, 而下采样窗口尺寸为 2×2 , 因此寻找窗口中最大值应依序在每一对奇偶行数据之间进行比较选取. 故此模式需指明奇数行状态. 另外, 此模式下需指明要求运算并输出结果, 与后续间隙模式进行区分;

(3) 间隙: 即当卷积窗口移至行末时, 到移至下一行行初的过程中处于间隙期, 会输出无效值. 故此模式下不运算且不输出结果, 而此模式的周期则为卷积核尺寸减 1 的结果. 具体演示如图 4 所示, 灰色区域为卷积窗口.

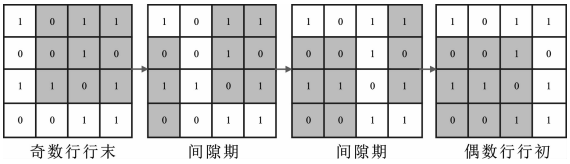


图 4 间隙期演示图

(4) 偶数行: 设计此模式理由同奇数行, 故不再累述. 同样需指明偶数行状态, 并且指明要求计算并输出结果.

需要补充的是, 偶数行模式的下一步又是间隙期. 即奇偶行交替, 中间夹杂着间隙期. 易看出状态的轮换有重复性, 故设计两个计数器: 其一控制状态更替, 即完成一次奇数行-间隙期-偶数行-间隙期后, 计数器恢复奇数行时的初始状态从而实现循环; 其二则一直保持计数直到处理完整幅图像数据表明卷积任务完成.

3.2 激活模块

由于激活函数选择了 ReLU 非线性函数, 故本文的激活模块只需对卷积模块电路输出的结果进行正负值判定即可. 当卷积窗口运算结果为正, 则激活后仍为原值; 运算结果为负, 则激活后为 0.

3.3 下采样模块

本文选择了最大值下采样方式, 即在 2×2 的下采样窗口中选择最大值输出结果, 因此本文提出如下设计思路: 依序在每一对奇偶行数据中分别从行始至行末依序对每一对数据进行比较取较大值, 待比较完毕则将奇数行和偶数行的较大值依序互相比对得到最大值. 由于奇偶行比较操作相同, 故图 5 以奇数行为例展示比较流程图. (注: Data 即为卷积模块经激活后输出的结果; 一行中的数据位置用 POS 变量表示, 每两个比较, 0 代表第一个数据, 1 代表第二个数据, 0 和 1 保持更替; Tmp 为寄存器暂存数据; Data odd[i] 变量为比较得到的较大值, i 始于 0, 上限值取决于卷积窗口滑动一行产生的结果数的一半, 设定为 m .)

同理, 偶数行会产生与之对应的 Data even[i]. 最后再将 Data odd 与 Data even 每个位置的数据进行比较取较大值, 则可得到 2×2 下采样窗口的最大值.

4 实验结果与分析

4.1 功能验证

本次实验中像素数据位数设定为 16 位, 不含符号位; 权值由于存在负数可能, 故数据位数定为 16

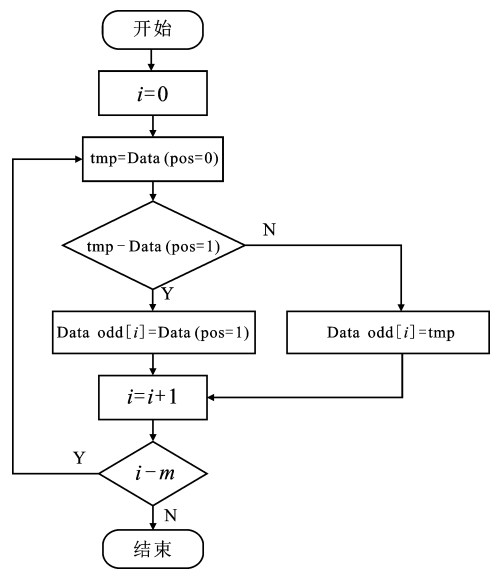


图5 下采样单行内比较操作流程

位,含1位符号位.实验利用Altera公司的Quartus II 11.1平台进行设计开发,FPGA开发板采用DE-4系列的Stratix IV EP4SGX70HF35C2开发板,利用其自带ModelSim-Altera进行功能仿真,而作为对比实验,CPU和GPU的性能测试则是基于Microsoft Visual Studio平台进行.

功能验证实验为了便于观测结果,图像尺寸采用 6×6 ,卷积核尺寸采用 3×3 .如图6所示的图像数据和卷积核数据.

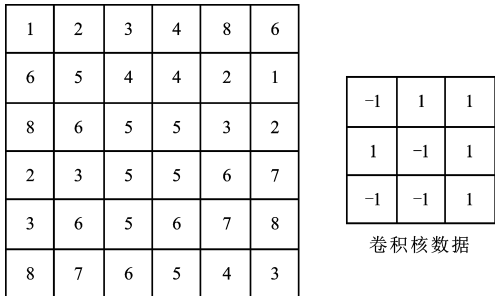


图6 图像数据和卷积核数据

仿真波形如下图7所示.

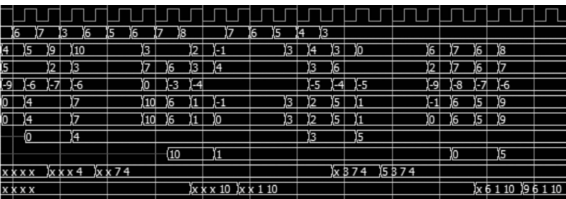


图7 功能仿真波形图

如图7所示,第一行波形为时钟波形50 MHz;第

二行波形为图像像素输入;第三行至第五行波形为卷积模块一个卷积窗口中三行各自乘加的结果,可以看出4个结果为一簇,中间有一段结果保持不变即为间隙期;第六行波形即为卷积窗口三行并行相加的结果;第七行波形即为经过激活后的结果;第八行和第九行分别为奇数行和偶数行的Tmp;第十行和第十一行即为Data odd[i],Data even[i],可以看出依次得到了该行比较的较大值;oRESULT则为最终结果,可以看出整个电路功能正确.

4.2 性能对比实验分析

本次对比实验将功能验证实验中的卷积模块以及下采样模块针对处理图像尺寸稍作修改,对 32×32 的图像利用 3×3 卷积核进行卷积.其主频设置为50 MHz,则三个功能模块构成的基本运算单元在单块FPGA开发板上资源消耗情况如下表1所示.

表1 FPGA资源开销情况

模块	消耗资源	消耗占比/%
LUTS	731/58 080	1
BRAMS	64/6 617 088	<1
REGISTERS	689/58 080	<1
DSPS	9/384	2

由表1看出开发板所拥有的计算资源远大于实验损耗,资源消耗占比不大.同时,使用的逻辑计算单元与寄存器资源较多.当复制该电路实现对多个卷积核或多张图像并行处理时,资源消耗将迅速增高.

表2则显示在CPU、GPU以及FPGA三种平台下相同的算法结构模型对尺寸为 32×32 的相同像素数据经过卷积、激活以及下采样的处理时间.算法耗时包括图像数据的读取传输以及运算处理时间.由表看出,运算效率与CPU相比提高了8倍,与GPU相比也提高了近5倍.

表2 三种平台下运算效率对比

平台	运算量	耗时/ μ s
Xeon E5 2.8GHz CPU	3×3 卷积+激活+下采样	1 758
Nvidia Titan X GPU	3×3 卷积+激活+下采样	1 060
Altera Stratix IV FPGA	3×3 卷积+激活+下采样	217

在电路功耗方面,本文采用Quartus II提供的PowerPlay Power Analyzer来测试能耗情况,并和Nvidia Titan X GPU进行对比,如表3所示.

表 3 GPU 与 FPGA 功耗对比

平台	功耗/mW
Nvidia Titan X GPU	1 687
Altera Stratix IV FPGA	469

可以看出本文实现的 FPGA 电路与 GPU 相比获得了近 5 倍的加速,但功耗只占了 GPU 的 27.8%.

5 结束语

本文依照卷积神经网络运算过程中的高并行度的特点,借助流水线思想在 FPGA 上设计了卷积运算模块电路,并且配合激活电路以及下采样电路从而构建了卷积神经网络并行加速器的基本单元. 根据不同网络只需极少修改然后大量复制该基本单元即可实现对卷积神经网络的硬件加速. 经过 Model-Sim 软件仿真之后,验证其功能正确. 并且在 FPGA 主频为 50 MHz 的情况下,对 32×32 的图像进行 3×3 卷积、激活和下采样的计算效率比 CPU 提高了 8 倍,比 GPU 提高了近 5 倍. 而在功耗方面,本文的 FPGA 方案只有 GPU 的 27.8%. 故利用 FPGA 实现对卷积神经网络的低功耗硬件加速将成为未来的趋势.

参考文献:

[1] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA ,2014.

[2] Li H, Li Y, Porikli F. DeepTrack: learning discriminative feature representations by convolutional neural networks for visual tracking[C]//Proceedings British

Machine Vision Conference. York, UK, 2014.

[3] Lecun Y, Boser B, Denker J S, et al. Backpropagation Applied to Handwritten Zip Code Recognition [J]. Neural Computation, 2014, 1(4): 541-51.

[4] 满凤环,陈秀宏,何佳佳. 一种基于模拟退火算法改进的卷积神经网络 [J]. 微电子学与计算机, 2017, 34 (9): 58-62.

[5] 朱锡祥,刘凤山,张超,等. 基于一维卷积神经网络的车载语音识别研究 [J]. 微电子学与计算机, 2017, 34(11): 21-26.

[6] 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): 2508-2523.

[7] Sankaradas M, Jakkula V, Cadambi S, et al. A massively parallel coprocessor for convolutional neural networks [C]//proceedings of the IEEE International Conference on Application-Specific Systems, Architectures and Processors. Milan, Italy,2009.

[8] 李施豪,应三丛. 基于 FPGA 的卷积神经网络浮点激励函数实现 [J]. 微电子学与计算机, 2017 , 34 (10): 105-109.

作者简介:

刘志成 男,(1993-),硕士研究生. 研究方向为深度学习硬件加速设计.

祝永新 男,(1969-),博士,博士生导师,研究员. 研究方向为计算机体系结构、芯片系统及设计、大数据处理.

汪 辉(通讯作者) 男,(1975-),博士,博士生导师,研究员. 研究方向为 CMOS 传感器及电路. E-mail: wanghui@sari. ac. cn.

田 犁 男,(1985-),博士,副研究员. 研究方向为图像处理与微电子电路.

封松林 男,(1964-),博士后,博士生导师,研究员. 研究方向为基于智慧感知的核心器件、大数据以及云计算.