(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111582451 B
(45)授权公告日 2022.09.06

(21)申请号 202010383601.0

(22)申请日 2020.05.08

(65)同一申请的已公布的文献号
申请公布号 CN 111582451 A

(43)申请公布日 2020.08.25

(73)专利权人 中国科学技术大学
地址 230026 安徽省合肥市包河区金寨路96号

(72)发明人 陈松 刘百成 康一

(74)专利代理机构 北京凯特来知识产权代理有限公司 11260
专利代理师 郑立明 韩珂

(51)Int.Cl.
G06N 3/04(2006.01)
G06N 3/063(2006.01)

(56)对比文件
CN 110780923 A,2020.02.11
CN 109784489 A,2019.05.21
CN 106909970 A,2017.06.30
CN 108665063 A,2018.10.16
CN 111008691 A,2020.04.14
CN 108647773 A,2018.10.12
CN 107066239 A,2017.08.18
CN 110782022 A,2020.02.11
US 6148101 A,2000.11.14
王巍 等.卷积神经网络(CNN)算法的FPGA并行结构设计.《微电子学与计算机》.2014,第57-62,66页.
Daniel Gibert 等.A Hierarchical Convolutional Neural Network for Malware Classification.《2019 International Joint Conference on Neural Networks (IJCNN)》.2019,第1-8页.

审查员 张杨

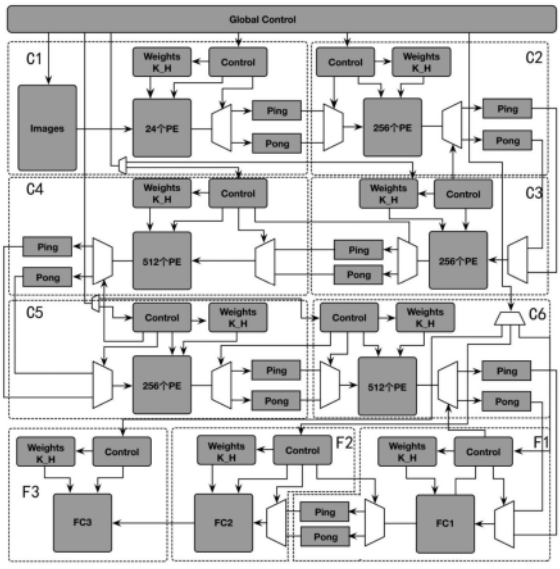
权利要求书2页 说明书6页 附图2页

(54)发明名称

图像识别层间并行流水线型二值化卷积神经网络阵列架构

(57)摘要

本发明公开了一种图像识别层间并行流水线型二值化卷积神经网络阵列架构,包括:依次设置的M1层、M2层、M3层、M4层及M5层五个计算层,并构成层间流水线,其中:M1层、M2层及M3层各自包含两个卷积层的计算,层内各自构成二级流水线,各层末端还有最大值池化层完成池化计算;M4层与M5层各自包含1个与两个全连接层的计算;每一卷积层及每一全连接层内都设有连接全局控制器的控制单元,以及用于存储权重参数和二值编码参数的存储器。该架构可以提高图像识别计算并行度,降低权重存储需求,同时有效避免乘法计算,降低功耗,提高能效。



CN 111582451 B

1.一种图像识别层间并行流水线型二值化卷积神经网络阵列架构,其特征在于,包括:依次设置的M1层、M2层、M3层、M4层及M5层五个计算层,并构成层间流水线,其中:

M1层、M2层及M3层各自包含两个卷积层的计算,层内各自构成二级流水线,各层末端还有最大值池化层完成池化计算;M4层与M5层各自包含1个与两个全连接层的计算;每一卷积层及每一全连接层内都设有连接全局控制器的控制单元,以及用于存储权重参数和二值编码参数的存储器;

其中,所述M1层中的第一个卷积层C1首先进行计算,当部分结果计算完成后,M1层中的第一个卷积层C1和第二个卷积层C2同时进行计算;M1层计算结果完成共需N时钟周期,则M1层在第一个N时钟周期内开展计算,第二个N时钟周期M1层和M2层同时工作,依次类推,第五个N时钟周期M1层、M2层、M3层、M4层及M5层同时工作,从而构成五级流水线。

2.根据权利要求1所述的一种图像识别层间并行流水线型二值化卷积神经网络阵列架构,其特征在于,所述M1层、M2层及M3层的每一卷积层内都设有多个处理单元,输入接控制单元与存储权重参数和二值编码参数的存储器;处理单元包含三个部分,第一个部分是输入缓冲部分,第二个部分为若干卷积计算部分,第三个部分为加法树单元,用来累加第二个部分输出的结果。

3.根据权利要求1或2所述的一种图像识别层间并行流水线型二值化卷积神经网络阵列架构,其特征在于,所述M1层中的第一个卷积层C1的输入是浮点数据,M1层中的第二个卷积层C2,输入的是第一个卷积层C1输出的二值化数据;同样的,M2层及M3层中卷积层的输入均为二值化数据;

所述M1层中的第一个卷积层C1,与M1层中的第二个卷积层C2以及M2层和M3层的卷积层中,处理单元内卷积计算部分的结构不同。

4.根据权利要求3所述的一种图像识别层间并行流水线型二值化卷积神经网络阵列架构,其特征在于,

所述第一个卷积层C1中通过一个选通单元控制输入数据,然后存放在寄存器中,再完成后续累加步骤,公式为:

$$y=\begin{cases} in, & if\ w=1 \\ -in, & if\ w=-1 \end{cases}$$

其中,in是输入值,为浮点数据,w是权重,是二值化数据,y是完成一个像素点卷积的结果。

5.根据权利要求4所述的一种图像识别层间并行流水线型二值化卷积神经网络阵列架构,其特征在于,M1层中的第二个卷积层C2以及M2层及M3层中卷积层输入为二值化数据,则表明输入和权重都是二值化的,使用同或累加运算,再通过下式进行结果转化得到最终的卷积结果:

$y'=2\times y_{xnor}-L_{conv}$

其中,y_{xnor}是同或累加的结果,L_{conv}是卷积核大小,y′为最终的卷积结果。

6.根据权利要求1所述的一种图像识别层间并行流水线型二值化卷积神经网络阵列架构,其特征在于,M1层、M2层及M3层的卷积层以及M4层中的全连接层的末端都设有乒乓缓存单元,用来存储相应层的计算结果。

7.根据权利要求1所述的一种图像识别层间并行流水线型二值化卷积神经网络阵列架

构，其特征在于，M1层、M2层及M3层的每一卷积层、M4层中的全连接层以及M5层的第一个全连接层的输出都进行批归一化操作，公式为：

$$Y=k_i \odot (x \geqslant h_f)$$

上式中，k_i表示k_f的符号，整数则为1，负数则为0，k_f为浮点数，$\odot$是同或运算符。

图像识别层间并行流水线型二值化卷积神经网络阵列架构

技术领域

[0001] 本发明涉及二值化卷积神经网络领域，尤其涉及一种图像识别层间并行流水线型二值化卷积神经网络阵列架构。

背景技术

[0002] 生物学中认为生物的大脑神经元和突触组成网络，可用于产生生物意识，帮助生物产生思考和行动。基于此，研究人工神经网络的科学家从中抽象出数学模型，从信息处理角度对人脑神经元进行抽象，建立简单的数学模型，按照不同的连接方式构成网络。目前，人工神经网络应用广泛，在语音识别领域、图像识别领域、目标检测领域等都有应用。在人工神经网络研究过程中，科学家提出卷积神经网络的概念，它是一类包含深度结构的人工神经网络，由前馈神经网络和负反馈神经网络组成，在识别时只进行前馈神经网络计算，训练时则需要进行负反馈神经网络计算。卷积神经网络研究受视觉细胞研究启发，发现初级视觉皮层中的神经元会响应视觉环境中的简单特征，视觉皮层存在简单细胞和复杂细胞，简单细胞对特定空间位置和偏好方向有强烈反应，通过对简单细胞的输入进行池化可以实现复杂空间上的不变性。由此可知，在卷积神经网络中，基础计算为卷积计算和池化计算。卷积计算是使用特定大小的卷积核来提取某个特定区域内的特征，主要是乘累加的计算过程。池化计算则是下采样的过程，下采样可以去除不重要的特征元素，降低特征图规模，减少计算参数，同时能保留特征图的重要特征，使其不影响后续计算。

[0003] 随着研究的深入，卷积神经网络网络规模逐渐增大，这导致卷积神经网络需要更多存储资源，计算资源消耗也持续增大。因此，研究减少卷积神经网络存储需求和计算需求成了卷积神经网络研究的一个热点。目前，降低卷积神经网络存储需求和计算需求的主流方法有剪枝、奇异值分解、量化、脉冲神经网络等几种方式。剪枝可以在训练时找到相邻两层之间相对不重要的连接并将其权重置0，即相当于剪断连接，因此在计算过程中减少了权重参数存储和计算次数；奇异值分解一般应用在全连接层中，通过奇异值分解的方式可以将两个大规模的矩阵相乘转化为三个较小规模的矩阵相乘，从而也能降低存储需求和计算需求；量化神经网络则是使用较少比特数来表示原浮点数值，一般可用11bit、8bit、5bit、3bit、2bit、1bit等，采用1bit即使用+1和-1两种状态完成计算的网络又叫做二值化卷积神经网络；脉冲神经网络更接近生物神经网络的工作模式，在计算中如果某个突触前神经元的膜电位超过了预设的电压阈值则向后发射一个脉冲，否则对应的突触后神经元因为没有输入脉冲均保持非工作状态，在硬件加速中没有脉冲即无动态功耗，仅存在静态功耗，从而也能降低总功耗。

[0004] 为了达到图片实时性处理效果，科研工作者一般采用GPU、FPGA和ASIC设计加速器。但是，受限于卷积神经网络存储需求和计算需求大，图像识别消耗资源多，很多硬件难以满足存储需求，计算并行度低，无法实现高能效，因此，基于二值化卷积神经网络，设计一种层间并行流水线型阵列架构用于图像识别是非常重要的。

发明内容

[0005] 本发明的目的是提供一种图像识别层间并行流水线型二值化卷积神经网络阵列架构，能够提高图像识别计算并行度，降低权重存储需求，同时有效避免乘法计算，降低功耗，提高能效。

[0006] 本发明的目的是通过以下技术方案实现的：

[0007] 一种图像识别层间并行流水线型二值化卷积神经网络阵列架构，包括：依次设置的M1层、M2层、M3层、M4层及M5层五个计算层，并构成层间流水线，其中：

[0008] M1层、M2层及M3层各自包含两个卷积层的计算，层内各自构成二级流水线，各层末端还有最大值池化层完成池化计算；M4层与M5层各自包含1个与两个全连接层的计算；每一卷积层及每一全连接层内都设有连接全局控制器的控制单元，以及用于存储权重参数和二值编码参数的存储器。

[0009] 由上述本发明提供的技术方案可以看出，图像识别二值化卷积神经网络硬件加速计算可以降低硬件存储需求，避免乘法计算，降低能耗，提高并行度，从而提高识别速度和能效。

附图说明

[0010] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域的普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他附图。

[0011] 图1为本发明实施例提供的一种图像识别层间并行流水线型二值化卷积神经网络阵列架构的示意图；

[0012] 图2为本发明实施例提供的层间并行流水线计算示意图；

[0013] 图3为本发明实施例提供的PE单元的卷积计算部分的第一类C结构示意图；

[0014] 图4为本发明实施例提供的PE单元的卷积计算部分的第二类C结构示意图；

[0015] 图5为本发明实施例提供的二值化乘累加计算转换为同或累加计算示意图；

[0016] 图6为本发明实施例提供的一个3*3大小的卷积核的PE单元示意图。

具体实施方式

[0017] 下面结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明的保护范围。

[0018] 本发明实施例提供一种图像识别层间并行流水线型二值化卷积神经网络阵列架构，其主要包括：依次设置的M1、M2、M3、M4及M5五个计算层，并构成层间流水线，其中：

[0019] M1层、M2层及M3层各自包含两个卷积层的计算，层内各自构成二级流水线，各层末端还有最大值池化层完成池化计算；M4层与M5层各自包含1个与两个全连接层的计算；每一卷积层及每一全连接层内都设有连接全局控制器的控制单元，以及用于存储权重参数和二值编码参数的存储器。

[0020] 如图1所示,M1层包含C1和C2两个卷积层并构成二级流水线,M2层包含C3和C4两个卷积层并构成二级流水线,M3层包含C5和C6两个卷积层并构成二级流水线,M4层包含全连接层F1,M5层包含全连接层F2和F3。

[0021] 如图1所示,整个二值化卷积神经网络阵列架构划分为9块,分别对应6个卷积层(C1～C6)和3个全连接层(层F1～F3);其中,GlobalControl是全局控制器用于实现全局控制。图1中,卷积层C1中Images用于存放待识别的图片,Weights&&K_H用于存放权重参数和二值编码所需要的K_H参数,Control是卷积层C1的控制单元(接收全局控制器下发的控制信号),Ping-PongBuffer是乒乓缓存单元,包含两块相同的存储单元,用于存放卷积层C1的阵列计算结果,其中卷积层C1使用24个PE(处理单元)完成计算,PE的输入接控制单元与存储权重参数和二值编码参数的存储器;卷积层C2中同样包含控制信号、权重和二值编码参数,其中计算阵列使用256个PE,输入由卷积层C1中的Ping-PongBuffer给入,输出同样存放在卷积层C2的Ping-PongBuffer中;卷积层C3包含控制信号、权重和二值编码参数,阵列使用256个PE,输入由卷积层C2中的Ping-PongBuffer给入;卷积层C4包含控制信号、权重和二值编码参数,阵列使用512个PE,输入由卷积层C3中的Ping-PongBuffer给入,输出存放在卷积层C4中的Ping-PongBuffer中;卷积层C5包含控制信号、权重和二值编码参数,阵列使用256个PE,输入从卷积层C4中的Ping-PongBuffer给入,输出存放在卷积层C5中的Ping-PongBuffer中;卷积层C6中包含控制信号、权重和二值编码参数,阵列使用512个PE,输入从卷积层C5中的Ping-PongBuffer给入,输出存放在卷积层C6中的Ping-PongBuffer中;全连接层F1中包含控制信号、权重参数和二值编码参数,阵列计算不采用卷积PE,而使用512个同或计算单元并用加法树完成运算,输入由卷积层C6中的Ping-PongBuffer给入,输出存放在全连接层F1中的Ping-PongBuffer中;全连接层F2中包含控制信号、权重参数和二值编码参数,同样使用86个同或计算单元并用加法树完成计算,输入由全连接层F1中的Ping-PongBuffer给入,输出直接输出到全连接层F3计算;全连接层F3包含控制、权重参数和二值编码参数,输入为全连接层F2的阵列计算结果并逐个完成计算并累加,全连接层F2计算结果不存储在Ping-PongBuffer中。

[0022] 需要说明的是,上述介绍以及图1所示的结构中,所给出的各卷积层内部的PE数目均为举例,并非构成限制;在实际应用中,用户可根据实际情况做配套的调整。

[0023] 本发明实施例所提供的二值化卷积神经网络阵列架构可构成五级流水线,所述M1层中的第一个卷积层C1首先进行计算,当部分结果计算完成后,M1层中的第一个卷积层C1和第二个卷积层C2同时进行计算;M1层计算结果完成共需N时钟周期,则M1层在第一个N时钟周期内开展计算,第二个N时钟周期M1层和M2层同时工作,依次类推,第五个N时钟周期M1层、M2层、M3层、M4层及M5层同时工作,从而构成五级流水线。

[0024] 如图2所示,从上至下有9个层,分别是C1、C2、C3、C4、C5、C6、F1、F2和F3。n1表示C1完成计算所需要的时间,n2表示C2完成计算所需要的时间,n3表示M1完成计算所需要的时间。由于C2计算需要使用C1的计算结果作为输入,所以为了保证计算结果准确并提高计算并行度,在C1完成部分计算时C2开始计算,等到C2完全计算结束时,M2层中的C3和C4依次开始计算,同理有M3中的C5和C6完成计算,M4中只包含F1一个层,M5中包含F2和F3两层,这样设置的原因是流水线结构运行速度取决于计算最慢的那一层。

[0025] 本发明实施例中,第一个卷积层C1输入的数据类型与后续卷积层输入的数据类型

不同。具体来说：第一个卷积层C1的输入是浮点数据，M1层中的第二个卷积层C2，输入的是第一个卷积层C1输出的二值化数据；同样的，M2层及M3层中卷积层的输入均为二值化数据；所述M1层中的第一个卷积层C1，与M1层中的第二个卷积层C2以及M2层和M3层的卷积层中，卷积计算部分的结构不同。

[0026] 如图1所示，第一个卷积层C1输入的是图像images，由于有些输入图片是RGB彩色的，因此无法二值化，因此，第一个卷积层C1中需要使用图3所示的C单元，其通过一个选通单元控制输入数据，然后存放在寄存器中，再完成后续累加步骤，使用这个单元，即免除了乘法计算，公式为：

[0027] $$y=\begin{cases} in, & if\ w=1 \\ -in, & if\ w=-1 \end{cases}$$

[0028] 其中，in是输入值，为浮点数据，是权重，是二值化数据，一般为+1或-1两个态，在硬件上可使用1bit表示数据，y是完成一个像素点卷积的结果，一般一个卷积核有若干个像素点卷积累加，例如3*3大小、5*5大小、7*7大小等，示例性的，可使用3*3大小的卷积核。因为输入图片是浮点数据，所以图3这种C单元计算使用16bit寄存器来存放数据，其中1个符号位，5个整数位，10个小数位。

[0029] 与图3的结构不同，图4适用于卷积层C2～C6的计算。M1层中的第二个卷积层C2以及M2层及M3层中卷积层输入为二值化数据，则表明输入和权重都是二值化的，使用同或累加运算。其中，同或累加运算可以用公式表示如下：

[0030] $$y_{xnor}=\sum_{i=0}^{i=k}\sum_{j=0}^{j=k} in_{i,j}\odot w_{i,j}$$

[0031] 上式中，k表示卷积核的大小，例如3*3的卷积，k＝3，i和j分别表示卷积的特征图像素和权重的位置坐标；in表示输入特征图，w表示权重，⊙是同或运算符，即y_{xnor}是同或累加的运算结果。通过这个公式可以在二值化的卷积神经网络中将乘法运算替换成同或运算。结合图5，如果一个3*3大小的卷积运算，左图使用乘累加运算后结果是-3；右图换成同或累加计算后，则变成了+3。为了保证在硬件电路上计算的一致性，需要通过下式进行结果转化得到最终的卷积结果：

[0032] $y'=2\times y_{xnor}-L_{conv}$

[0033] 其中，y_{xnor}是使用同或累加运算的卷积结果；L_{conv}是卷积核大小，假设使用的是3*3，因为一个神经元对应的不只一个卷积核，则L_{conv}一般为9的倍数；y′为最终的卷积结果。全连接层计算中L_{conv}的大小就是权重矩阵列的大小。前面已经介绍了图5中左图乘累加计算结果为-3，右图中同或累加计算结果为3，将同或累加结果代入上式中求得y′的结果为-3，从而实现了乘累加计算转为同或累加计算后卷积计算结果仍然正确。

[0034] 如果按照这种公式计算，仍然有一次乘法运算，但是，本发明实施例中，M1层、M2层及M3层中的每一卷积层(即C1～C6)、M4层中的全连接层以及M5层的第一个全连接层的输出都进行批归一化操作(BatchNormalization，简写为BN)，作用是在训练的时候加快训练速度，批归一化操作的公式为：

[0035] $$Y=\gamma\times\frac{X-\mu}{\sqrt{\sigma^2+\epsilon}}+\beta$$

[0036] 其中，μ是期望，σ^2是方差，γ 和β批归一化操作的权重及偏置，∈为常数(远小于1的正常数)，加入这个常数的作用是为了防止σ^2等于0；X表示卷积层或者全连接层的输出。

[0037] 结合乘累加转同或累加公式及BN层公式，可推导出公式：

[0038] $Y=k_f\times(X-h_f)$

[0039] $$k_f=\frac{2\times\gamma}{\sqrt{\sigma^2+\epsilon}}$$

[0040] $$h_f=\left(L_{conv}+\mu-bias-\beta\times\frac{\sqrt{\sigma^2+\epsilon}}{\gamma}\right)\times 0.5$$

[0041] 上式中，X是不含加偏置的卷积计算结果，bias是偏置，k_f和h_f是结合卷积计算和BN层计算公式推导两组计算参数，均为浮点数。

[0042] 卷积层C1～C6、全连接层F1～F2都是二值化输出，需要进行二值化编码，再结合上述公式和激活函数，二值化编码表示为：

[0043] $$sign(x)=\begin{cases}+1, if\ x\geq 0\\ -1, if\ x<0\end{cases}$$

[0044] 上式中，sign(x)是符号函数；x是指输入的信息，即批归一化得到的Y。

[0045] 结合激活函数，可以将推导出的函数化简为：

[0046] $Y=k_i\odot(x\geqslant h_f)$

[0047] 上式中，k_i表示k_f的符号，整数则为1，负数则为0，h_f可以在硬件加速前离线处理好再导入加速器中进行计算，可以直接输入到硬件电路中参与计算，完成二值编码因此可以避免乘法运算，同时简化计算过程。

[0048] 如图1所示，卷积层C1～C6的内部各自包含多个PE单元(处理单元)，各处理单元是并行工作模式，图6示例性的给出了3*3卷积的PE单元的结构，主要包括三个部分：第一个部分是输入缓冲部分；第二个部分为若干卷积计算部分；图6中设置了9个如图3所示的C单元，也就意味着图6为卷积层C1的PE单元，对于其他卷积层C2～C6，相应的更换为图4所示的C单元即可；第三个部分为加法树单元，用来累加第二个部分输出的结果，因为本示例中，卷积核大小是3*3，所以输入缓冲需要缓冲3行，同或计算部分的权重寄存单元通过广播方式在计算前完成缓存，加法树当同或计算部分完成计算后开始工作并输出结果。

[0049] 为了更好的说明实施例的计算过程，下面给出该实施例的神经网络一个具体结构，如下表：

[0050]

层	输入	填充	卷积核	输出	权重大小	输出大小
C1	3*32*32	1(-1)	64*3*3*3	64*32*32	1728b	64kb
C2	64*32*32	1(-1)	64*64*3*3	64*32*32	36kb	64kb
MP1	64*32*32	-	-	64*16*16	-	16kb
C3	64*16*16	1(-1)	128*64*3*3	128*16*16	72kb	32kb
C4	128*16*16	1(-1)	128*128*3*3	128*16*16	144kb	32kb
MP2	128*16*16	-	-	128*8*8	-	8kb
C5	128*8*8	1(-1)	256*128*3*3	256*8*8	288kb	16kb
C6	256*8*8	1(-1)	256*256*3*3	256*8*8	576kb	16kb

MP3	256*8*8	-	-	256*4*4	-	4kb
F1	4096	-	4096	1024	4Mb	1kb
F2	1024	-	1024	1024	1Mb	1kb
F3	1024	-	1024	10	10kb	10b

[0051] 表1神经网络结构

[0052] 上表中，C1、C2、C3、C4、C5、C6表示6个卷积层，F1、F2、F3表示3个全连接层，MP1、MP2、MP3表示3个池化层，池化层采用的在2*2区域内保留最大值，即采用最大值池化方式。特别说明，在结合硬件电路架构加速时，在软件训练时已将填充的值从0变成了-1，一般训练时通常都是填充0，修改填充值为-1保证了硬件电路中采用同或替换乘法运算时能够避免+1、0和-1的三值运算。卷积核这一项中第一个数值表示卷积核数目，第二个数值表示卷积核

[0053] 以上所述，仅为本发明较佳的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明披露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应该以权利要求书的保护范围为准。

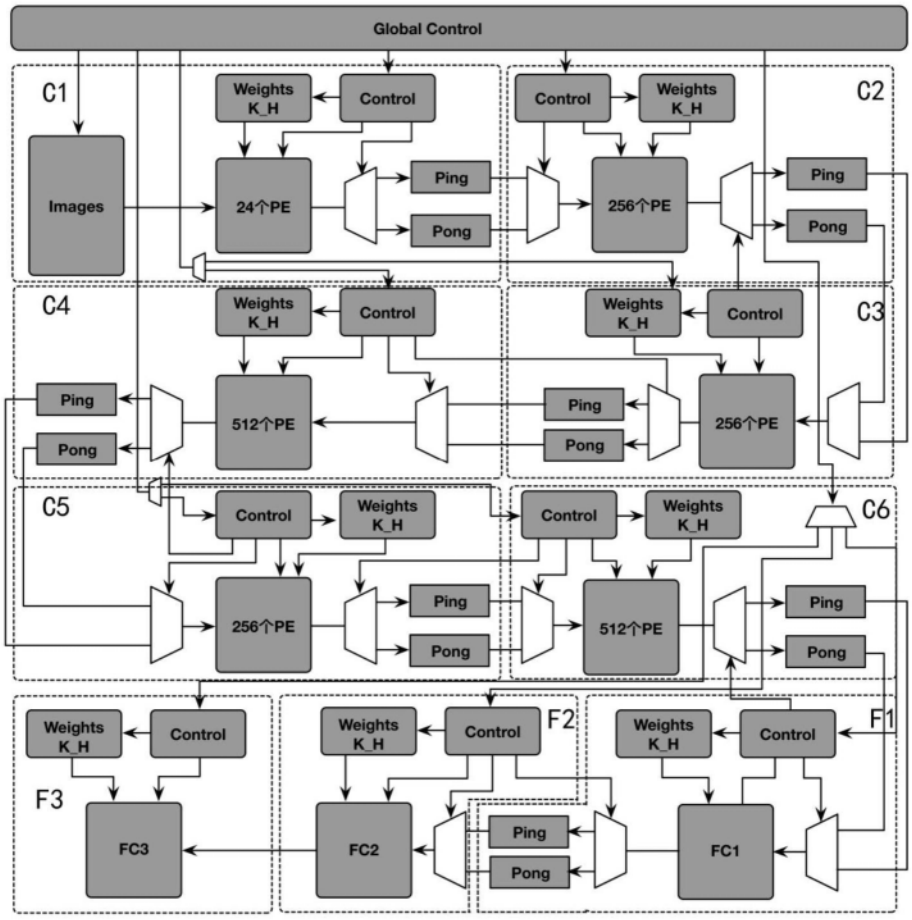

Global Control
C1
Weights K_H
Control
Images
24个PE
Ping
Pong
C2
256个PE
C4
512个PE
C3
C5
C6
F1
F2
F3
FC1
FC2
FC3

图1

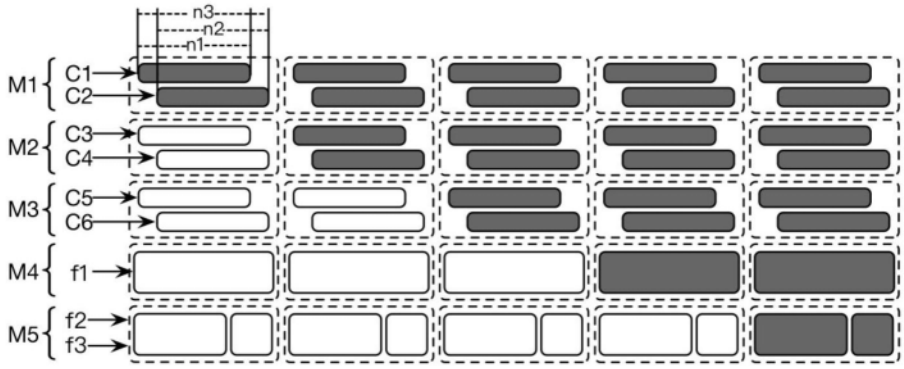

n3
n2
n1
M1
C1
C2
M2
C3
C4
M3
C5
C6
M4
f1
M5
f2
f3

图2

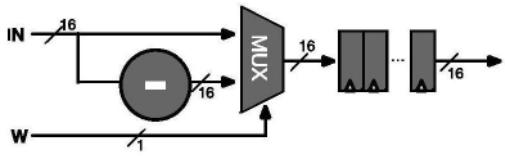


图3

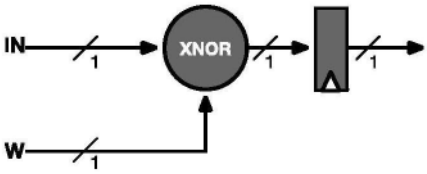


图4

+1	-1	-1		-1	+1	+1		1	0	0		0	1	1
+1	-1	-1	x	-1	+1	-1	= -3	1	0	0	⊙	0	1	0	= 3
-1	+1	+1		-1	-1	+1		0	1	1		0	0	1

图5

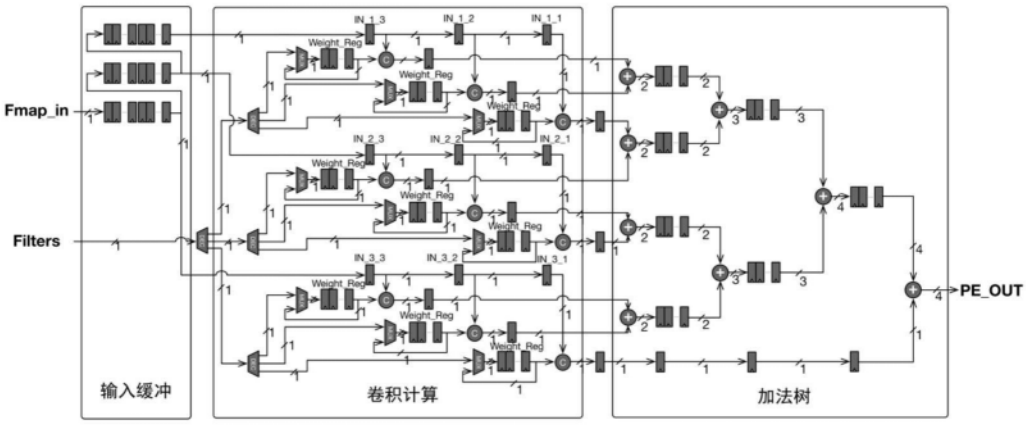


图6