(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 111610963 B
(45)授权公告日 2021.08.17

(21)申请号 202010587029.X

(22)申请日 2020.06.24

(65)同一申请的已公布的文献号
申请公布号 CN 111610963 A

(43)申请公布日 2020.09.01

(73)专利权人 上海西井信息科技有限公司
地址 200050 上海市长宁区江苏路398号503-3室

(72)发明人 谭黎敏 宋捷 桑迟

(74)专利代理机构 上海隆天律师事务所 31282
代理人 潘一诺

(51)Int.Cl.
G06F 7/544(2006.01)
G06N 3/063(2006.01)
G06N 3/04(2006.01)

(56)对比文件
CN 110807522 A,2020.02.18
CN 108241890 A,2018.07.03
CN 111222090 A,2020.06.02
CN 108960414 A,2018.12.07
CN 106844294 A,2017.06.13
CN 110765411 A,2020.02.07
CN 107862378 A,2018.03.30
KR 102038390 B1,2019.10.31
US 2019197083 A1,2019.06.27

审查员 郑艳梅

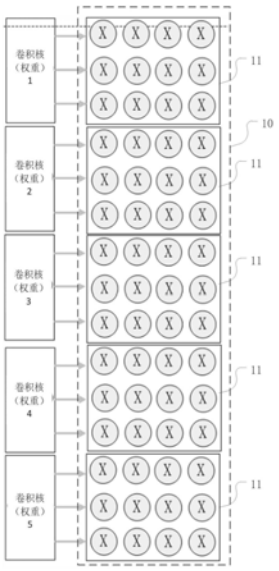
权利要求书2页 说明书7页 附图9页

(54)发明名称

芯片结构及其乘加计算引擎

(57)摘要

本发明提供一种芯片结构及其乘加计算引擎,乘加计算引擎包括:多个卷积引擎,每个所述卷积引擎包括15*M*N个乘法器以及至少设置在部分乘法器之间的加法器,M,N均为大于1的整数,其中,根据所述卷积引擎所应用的不同的卷积核的尺寸,按不同的方式激活所述乘法器之间、所述乘法器与所述加法器之间的连接方式。本发明提供的芯片结构及其乘加计算引擎利用同一套卷积引擎在应用到不同的卷积核的尺寸时,提高乘法器资源的利用率,且根据不同的卷积核的尺寸实现不同的乘法器资源分配,实现数据的动态分布,达到需要的计算方式。



CN 111610963 B

1.一种乘加计算引擎,其特征在于,包括:

多个卷积引擎,每个所述卷积引擎包括15*M*N个乘法器形成的乘法器阵列以及至少设置在部分乘法器之间的加法器,N为大于1的整数,M为大于1的整数,

其中,根据所述卷积引擎所应用的不同的卷积核的尺寸,按不同的方式激活所述乘法器之间、所述乘法器与所述加法器之间的连接方式,

其中,多个所述卷积引擎对输入特征图和卷积核权重进行乘法运算,并在输入特征通道将乘法结果进行累加以获得输出特征图;

每个所述卷积引擎在应用于卷积核时,按卷积核的尺寸划分为多个卷积组,每个所述卷积组的乘法器的行数与所述卷积核的行数一致,每个所述卷积组的乘法器的列数为N,以由所述卷积引擎提供在输出特征第一维度的N倍并行加速,所述卷积引擎还提供在输入特征通道的M倍并行加速,

对于步长S为1、尺寸为P*Q的卷积核,每个所述卷积组包括P*N个乘法器,每列乘法器的相邻两个乘法器之间连接有加法器,每个所述卷积组并行读取(P+N-1)行输入特征图,每个卷积组并行读取P行卷积核权重,其中,所述P行卷积核权重分别输入一行乘法器;所述(P+N-1)行输入特征图中第1至第P行分别输入一行乘法器,第P+1行至第P+N-1行分别输入一列乘法器,P、Q为大于1的整数;

对于步长S为2、尺寸为P*Q的卷积核,每个所述卷积组包括P*N个乘法器,每列乘法器的相邻两个乘法器之间连接有加法器,每个所述卷积组并行读取[S*N+(P-S)]行输入特征图。

2.如权利要求1所述的乘加计算引擎,其特征在于,由所述卷积引擎组成多组级联结构,每组级联结构包括M/2个级联的处理单元,每个处理单元包括第一输入接口至第五输入接口、第一触发器至第五触发器、两个乘法器以及两个加法器、一输出接口、第一时钟信号至第五时钟信号,其中,每个处理单元:

第一输入接口至第四输入接口分别连接至第一触发器及第四触发器;

第一时钟信号至第四时钟信号分别连接至第一触发器及第四触发器;

第一触发器和第二触发器的输出连接至一乘法器;

第三触发器和第四触发器的输出连接至另一乘法器

两个乘法器的输出连接至一加法器,且该加法器的输出连接至另一加法器;

第五输入接口将前一处理单元的输出接口的数据连接至另一加法器;

另一加法器的输出连接至第五触发器;

第五时钟信号连接至第五触发器;

第五触发器的输出连接至该处理单元的输出接口,

所述第一时钟信号至第五时钟信号分别用于开启第一触发器至第五触发器。

3.如权利要求2所述的乘加计算引擎,其特征在于,同一处理单元的第一时钟信号至第五时钟信号相同;在级联中,间隔(N/2-1)个处理单元的前一处理单元的时钟信号比后一处理单元的时钟信号早N/2个时钟周期,N为2的倍数。

4.如权利要求3所述的乘加计算引擎,其特征在于,对于尺寸为P*Q的卷积核,所述卷积引擎组成P组级联结构。

5.如权利要求1至4任一项所述的乘加计算引擎,其特征在于,N为4。

6.如权利要求1至4任一项所述的乘加计算引擎,其特征在于,M为16。

7.一种芯片结构,其特征在于,包括如权利要求1至6任一项所述的乘加计算引擎。

芯片结构及其乘加计算引擎

技术领域

[0001] 本发明涉及卷积神经网络领域,尤其涉及一种芯片结构及其乘加计算引擎。

背景技术

[0002] 卷积神经网络(Convolutional Neural Network,CNN)是一种前馈神经网络,它的人工神经元可以响应一部分覆盖范围内的周围单元,对于大型图像处理有出色表现。它主要包括卷积层(convolutional layer)和池化层(pooling layer)。卷积神经网络已广泛应用于图像分类、物体识别、目标追踪。

[0003] 对于卷积神经网络的处理芯片,如何通过乘法器和加法器的排布、设计从而提高乘法器资源的利用率,且根据不同的卷积核的尺寸实现不同的乘法器资源分配,实现数据的动态分布,达到需要的计算方式,是本领域技术人员亟待解决的技术问题。

发明内容

[0004] 本发明为了克服上述现有技术存在的缺陷,提供一种芯片结构及其乘加计算引擎,以利用同一套卷积引擎在应用到不同的卷积核的尺寸时,提高乘法器资源的利用率,且根据不同的卷积核的尺寸实现不同的乘法器资源分配,实现数据的动态分布,达到需要的计算方式。

[0005] 根据本发明的一个方面,提供一种乘加计算引擎,包括:

[0006] 多个卷积引擎,每个所述卷积引擎包括15*M*N个乘法器以及至少设置在部分乘法器之间的加法器,N为大于1的整数,M为大于1的整数,

[0007] 其中,根据所述卷积引擎所应用的不同的卷积核的尺寸,按不同的方式激活所述乘法器之间、所述乘法器与所述加法器之间的连接方式。

[0008] 在本发明的一些实施例中,多个所述卷积引擎对输入特征图和卷积核权重进行乘法运算,并在输入特征通道将乘法结果进行累加以获得输出特征图。

[0009] 在本发明的一些实施例中,每个所述卷积引擎在应用于卷积核时,按卷积核的尺寸划分为多个卷积组,每个所述卷积组的乘法器的行数与所述卷积核的行数一致,每个所述卷积组的乘法器的列数为N,以由所述卷积引擎提供在输出特征第一维度的N倍并行加速,所述卷积引擎还提供在输入特征通道的M倍并行加速。

[0010] 在本发明的一些实施例中,对于步长S为1、尺寸为P*Q的卷积核,每个所述卷积组包括P*N个乘法器,每列乘法器的相邻两个乘法器之间连接有加法器,每个所述卷积组并行读取(P+N-1)行输入特征图,每个卷积组并行读取P行卷积核权重,其中,所述P行卷积核权重分别输入一行乘法器;所述(P+N-1)行输入特征图中第1至第P行分别输入一行乘法器,第P+1行至第P+N-1行分别输入一列乘法器,P、Q为大于1的整数;

[0011] 对于步长S为2、尺寸为P*Q的卷积核,每个所述卷积组包括P*N个乘法器,每列乘法器的相邻两个乘法器之间连接有加法器,每个所述卷积组并行读取[S*N+(P-S)]行输入特征图。

[0012] 在本发明的一些实施例中，由所述卷积引擎组成多组级联结构，每组级联结构包括M/2个级联的处理单元，每个处理单元包括第一输入接口至第五输入接口、第一触发器至第五触发器、两个乘法器以及两个加法器、一输出接口、第一时钟信号至第五时钟信号，其中，每个处理单元：

[0013] 第一输入接口至第四输入接口分别连接至第一触发器及第四触发器；

[0014] 第一时钟信号至第四时钟信号分别连接至第一触发器及第四触发器；

[0015] 第一触发器和第二触发器的输出连接至一乘法器；

[0016] 第三触发器和第四触发器的输出连接至另一乘法器

[0017] 两个乘法器的输出连接至一加法器，且该加法器的输出连接至另一加法器；

[0018] 第五输入接口将前一处理单元的输出接口的数据连接至另一加法器；

[0019] 另一加法器的输出连接至第五触发器；

[0020] 第五时钟信号连接至第五触发器；

[0021] 第五触发器的输出连接至该处理单元的输出接口，

[0022] 所述第一时钟信号至第五时钟信号分别用于开启第一触发器至第五触发器。

[0023] 在本发明的一些实施例中，同一处理单元的第一时钟信号至第五时钟信号相同；在级联中，间隔(N/2-1)个处理单元的前一处理单元的时钟信号比后一处理单元的时钟信号早N/2个时钟周期，N为2的倍数。

[0024] 在本发明的一些实施例中，对于尺寸为P*Q的卷积核，所述卷积引擎组成P组级联结构。

[0025] 在本发明的一些实施例中，N为4。

[0026] 在本发明的一些实施例中，M为16。

[0027] 根据本发明的又一方面，还提供一种芯片结构，包括如上所述的乘加计算引擎。

[0028] 相比现有技术，本发明的优势在于：

[0029] 利用同一套卷积引擎在应用到不同的卷积核的尺寸时，提高乘法器资源的利用率，且根据不同的卷积核的尺寸实现不同的乘法器资源分配，实现数据的动态分布，达到需要的计算方式。

附图说明

[0030] 通过参照附图详细描述其示例实施方式，本发明的上述和其它特征及优点将变得更加明显。

[0031] 图1示出了根据本发明实施例的卷积引擎应用于尺寸为3*3的卷积核的示意图；

[0032] 图2示出了根据本发明实施例的卷积引擎应用于尺寸为5*5的卷积核的示意图；

[0033] 图3示出了根据本发明实施例的卷积引擎应用于尺寸为7*7的卷积核的示意图；

[0034] 图4示出了根据本发明实施例的应用于尺寸为5*5的卷积核的一卷积组的示意图；

[0035] 图5示出了根据本发明实施例输入特征图进行卷积的示意图；

[0036] 图6示出了根据本发明实施例输入特征图进行卷积后获得输出特征图的示意图；

[0037] 图7示出了根据本发明实施例的卷积引擎的示意图；

[0038] 图8示出了根据本发明实施例的处理单元的示意图；

[0039] 图9示出了根据本发明实施例的级联结构的示意图；

[0040] 图10示出了图9的级联结构中的时钟信号的时序图。

具体实施方式

[0041] 现在将参考附图更全面地描述示例实施方式。然而，示例实施方式能够以多种形式实施，且不应被理解为限于在此阐述的范例；相反，提供这些实施方式使得本公开将更加全面和完整，并将示例实施方式的构思全面地传达给本领域的技术人员。所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施方式中。

[0042] 此外，附图仅为本公开的示意性图解，并非一定是按比例绘制。图中相同的附图标记表示相同或类似的部分，因而将省略对它们的重复描述。附图中所示的一些方框图是功能实体，不一定必须与物理或逻辑上独立的实体相对应。可以采用软件形式来实现这些功能实体，或在一个或多个硬件模块或集成电路中实现这些功能实体，或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0043] 为了解决现有技术的缺陷，本发明提供一种乘加计算引擎。下面将结合图1至图10对本发明提供的乘加计算引擎进行进一步地描述。

[0044] 乘加计算引擎包括多个卷积引擎。每个所述卷积引擎包括15*M*N个乘法器以及至少设置在部分乘法器之间的加法器，N为大于1的整数，M为大于1的整数。其中，根据所述卷积引擎所应用的不同的卷积核的尺寸，按不同的方式激活所述乘法器之间、所述乘法器与所述加法器之间的连接方式。

[0045] 具体而言，卷积引擎的数量可以根据具体算力需求来确定，本发明并非以此为限制。

[0046] 在本发明的一个优选例中，N为4。在该优选例中，每个所述卷积引擎包括15*M*4个乘法器。对于常用的、主流的神经网络模型中，卷积核的尺寸包括7*7、5*5、3*3。对于卷积引擎10在M维度上每个15*4乘法器阵列，当应用到尺寸为3x3卷积核时，可将15行乘法器分为5个卷积组11，每个卷积组11包括3行4列乘法器，每个卷积组11对应一个卷积核计算，5组卷积组11同时可支持5个不同的卷积核并行计算(如图1所示，图1示出了根据本发明实施例的卷积引擎应用于尺寸为3*3的卷积核的示意图)。当对于卷积引擎10在M维度上每个15*4乘法器阵列应用到尺寸为5x5卷积核时，可将15行乘法器分为3组卷积组12，每个卷积组12包括5行4列乘法器，每个卷积组12对应一个卷积核计算，3每个卷积组12同时可支持3个不同的卷积核并行计算(如图2所示，图2示出了根据本发明实施例的卷积引擎应用于尺寸为5*5的卷积核的示意图)。同理，当对于卷积引擎10在M维度上每个15*4乘法器阵列，应用到尺寸为7x7卷积核时，可将15行乘法器分为2组卷积组13，每个卷积组13包括7行4列乘法器(最后一行乘法器不激活)，每个卷积组13对应一个卷积核计算，2每个卷积组13同时可支持2个不同的卷积核并行计算(如图3所示，图3示出了根据本发明实施例的卷积引擎应用于尺寸为7*7的卷积核的示意图)。

[0047] 以上仅仅是示意性地示出和描述本发明提供的卷积引擎在应用至尺寸为7*7、5*5、3*3的卷积核时的每个卷积引擎的分组情况，但本发明并非以此为限制。具体而言，每个所述卷积引擎在应用于卷积核时，可以按卷积核的尺寸划分为多个卷积组，每个所述卷积组的乘法器的行数与所述卷积核的行数一致，每个所述卷积组的乘法器的列数为N。

[0048] 由此，本发明利用同一套卷积引擎提高应用不同尺寸的卷积核时，乘法器的利用

率，且根据不同的卷积核的尺寸实现不同的乘法器资源分配，实现数据的动态分布，达到需要的计算方式。进一步地，本发明中，一个卷积引擎同时支持N行输出特征图的计算输出，即每列乘法器对应一行输出特征图像素值。在前述的优选例中，将N设定为4是基于整体架构的分片计算(tiling)特征及卷积后池化操作便捷性共同考虑的结果。

[0049] 下面参见图5和图6，图5示出了根据本发明实施例输入特征图进行卷积的示意图；图6示出了根据本发明实施例输入特征图进行卷积后获得输出特征图的示意图。多个所述卷积引擎对输入特征图30(尺寸为H*W)和卷积核权重20进行乘法运算，并在输入特征通道将乘法结果进行累加以获得输出特征图40(尺寸为F*E)。进一步，一次传统的卷积运算，需要卷积核在输出特征图40滑动，即形成多个滑动窗口(sliding window)。计算所有的滑动窗口，可以生成出一个完整的输出特征图40。利用如图1至图3所示的卷积引擎，可以并行加速多个卷积核的滑动窗口。卷积核的尺寸为P*Q，卷积引擎的卷积组的行数(N)即为在输出特征图40的E维度(第一维度)上的并行加速。而对于输出特征图40的输出特征通道E维度上可以通过滑动窗口的工作周期来实现。

[0050] 下面参见图4，图4示出了根据本发明实施例的应用于尺寸为5*5的卷积核的一卷积组的示意图。

[0051] 具体而言，对于步长S为1、尺寸为P*Q的卷积核，每个所述卷积组包括P*N个乘法器，每列乘法器的相邻两个乘法器之间连接有加法器，每个所述卷积组并行读取(P+N-1)行输入特征图，每个卷积组并行读取P行卷积核权重，其中，所述P行卷积核权重分别输入一行乘法器；所述(P+N-1)行输入特征图中第1至第P行分别输入一行乘法器，第P+1行至第P+N-1行分别输入一列乘法器，P、Q为大于1的整数。对于步长S为2、尺寸为P*Q的卷积核，每个所述卷积组包括P*N个乘法器，每列乘法器的相邻两个乘法器之间连接有加法器，每个所述卷积组并行读取[S*N+(P-S)]行输入特征图。具体而言，对于3*3的卷积核，每个所述卷积组并行读取(2x4)+(3-2)＝9行输入特征图；对于5*5的卷积核，每个所述卷积组并行读取(2x4)+(5-2)＝11行输入特征图；对于7*7的卷积核，每个所述卷积组并行读取(2x4)+(7-2)＝13行输入特征图。

[0052] 下面以尺寸为5*5的卷积核为例，描述一个卷积组的输入、输出、加法器、乘法器的连接方式。

[0053] 在图4所示的实施例中，卷积组包括5行4列乘法器，每一列乘法器中，相邻两个乘法器之间连接有一加法器。5*5的卷积核的第一行卷积核权重依次输入第一行乘法器；5*5的卷积核的第二行卷积核权重依次输入第二行乘法器；5*5的卷积核的第三行卷积核权重依次输入第三行乘法器；5*5的卷积核的第四行卷积核权重依次输入第四行乘法器；5*5的卷积核的第五行卷积核权重依次输入第五行乘法器。输入特征图的第一行输入第一行第一列的乘法器；输入特征图的第二行输入第二行第一列乘法器后输入第一行第二列乘法器；输入特征图的第三行输入第三行第一列乘法器后，输入第二行第二列乘法器，然后输入第一行第三列乘法器；输入特征图的第四行输入第四行第一列乘法器后，输入第三行第二列乘法器，然后输入第二行第三列乘法器，最后输入第一行第四列乘法器；输入特征图的第五行输入第五行第一列乘法器后，输入第四行第二列乘法器，然后输入第三行第三列乘法器，最后输入第二行第四列乘法器；输入特征图的第六行输入第五行第二列乘法器后，输入第四行第三列乘法器，然后输入第三行第四列乘法器；输入特征图的第七行输入第五行第三

列乘法器后,输入第四行第四列乘法器;输入特征图的第八行输入第五行第四列乘法器。各列乘法器由第五行开始依次通过加法器将乘法结果累加,从而对应四列乘法器获得部分累加值第一行、部分累加值第二行、部分累加值第三行、以及部分累加值第4行。

[0054] 由此,输入特征图可以从8个SRAM(Static Random Access Memory,即静态随机存取存储器)的读接口中并行读出,复用给图4示出的20个乘法器。而相比现有技术中,20个乘法器需要20个SRAM的读接口的方案,本发明节省了SRAM的读接口。对于输出特征图的4行数据,输入特征图实现了最大的片上缓存SRAM读取复用,比如,输入特征图第四行和第五行同时复用于四个乘法器,减少了SRAM的使用,优化了资源和功耗,面积。

[0055] 对于卷积核权重,本实施例使用5个SRAM读接口,同时读出5x5卷积核的5行权重的数据。每行权重数据复用给一行4个乘法器。

[0056] 在每列乘法器中,两个乘法器之间,插入一个加法器,乘法器的输出,按照流水线传递,直到完成所有的累加,到达此引擎的输出位置。

[0057] 图4仅仅是示意性地示出应用至尺寸为5*5的卷积核的卷积组,本发明并非以此为限制。进一步地,在图1至图3所示的实施例中,每列乘法器的相邻两个乘法器之间设置有加法器,从而当应用至7*7、5*5、3*3的卷积核时,可以不激活相邻卷积组之间的加法器,从而实现卷积引擎的复用。

[0058] 下面参见图7,图7示出了根据本发明实施例的卷积引擎的示意图。

[0059] 除了在输出特征通道(输入特征图的高度维度)上的并行加速之外,本发明的卷积引擎还可以提供在输入特征通道的M倍并行加速,M为大于1的整数。,每个卷积组输出的N行部分累加值(psum),不是最终的输出特征图的结果,需要在输入特征通道维度上做累加。考虑到常用的主流卷积神经网络模型,输入特征通道的数量通常以偶数出现,一般是2的n次方的形式。由此,可以利用16个卷积组,用于支持16个不同的输入特征图的通道计算,即在图5和图6中的维度C上加速。

[0060] 在本实施例中,以N为4且M为16,为例,说明输入特征通道的16倍并行加速。如图7,16个通道的数据在卷积组外的加法树中进行累加,最终形成输出特征图的部分累加值。本实施例中,将输入特征图的输入特征通道计算加速并行度定为16,即兼顾了加速的目的,即16倍加速,也考虑了算法模型的普适性,同时需要考虑过多资源导致片上资源紧张,密集布线区域带来的时序问题等。

[0061] 输出特征图部分累加值第一行由16个通道的部分累加值第一行累加获得;输出特征图部分累加值第二行由16个通道的部分累加值第二行累加获得;输出特征图部分累加值第三行由16个通道的部分累加值第三行累加获得;输出特征图部分累加值第四行由16个通道的部分累加值第四行累加获得

[0062] 由此,继承诸如图4中的部分累加值的累加方向,实现了输入特征通道维度16倍并行计算加速和输出特征通道维度4倍的并行计算加速。此外,根据不同硬件算力需求,不同产品的定位,图7中的架构,可考虑并行多份,针对不同输出特征通道维度进一步加速,提升性能。

[0063] 下面为了详细说明图7的结构,对图7示出的16个不同的输入特征图的通道的乘加计算继续细化,仅观察一个输入像素点的16个通道的数据乘加,也就是16个乘法,M/2个加法的硬件加速方案。该硬件加速方案由所述卷积引擎组成多组级联结构,每组级联结构包

括M/2个级联的处理单元。

[0064] 下面参见图8,图8示出了根据本发明实施例的处理单元的示意图。

[0065] 每个处理单元50包括第一输入接口511至第五输入接口515、第一触发器531至第五触发器535、两个乘法器541、542以及两个加法器551、552、一输出接口561、第一时钟信号521至第五时钟信号525。第一触发器531至第五触发器535例如为D触发器。

[0066] 每个处理单元50中,第一输入接口511至第四输入接口514分别连接至第一触发器531及第四触发器534。第一时钟信号521至第四时钟信号524分别连接至第一触发器531及第四触发器534。第一触发器531和第二触发器534的输出连接至一乘法器541。第三触发器533和第四触发器534的输出连接至另一乘法器542。两个乘法器541、542的输出连接至一加法器551,且该加法器551的输出连接至另一加法器552。第五输入接口515将前一处理单元的输出接口的数据连接至另一加法器552。另一加法器552的输出连接至第五触发器535。第五时钟信号525连接至第五触发器535。第五触发器535的输出连接至该处理单元的输出接口561。所述第一时钟信号521至第五时钟信号525分别用于开启第一触发器531至第五触发器535。在本发明的一些实施例中,同一处理单元的第一时钟信号至第五时钟信号相同;在级联中,间隔(N/2-1)个处理单元的前一处理单元的时钟信号比后一处理单元的时钟信号早N/2个时钟周期,N为2的倍数。

[0067] 进一步地,对于尺寸为P*Q的卷积核,所述卷积引擎组成P组级联结构。

[0068] 在如图7所示的实施例中,也就是16个输入特征通道的实施例中,可以使用2^{M-1}个,也就是8个级联的方式,可以完成一个像素点在16个通道内的乘加运算。具体的连接方式如图9所示。图9示出了根据本发明实施例的级联结构的示意图。

[0069] 图9中的级联的处理单元的输出结果,最终输出了一个输入特征图向的一像素点的16通道乘累加结果。以尺寸为3*3的卷积核为例,9个输入像素点的16个输入特征通道的乘累加结果,在这个架构下,需要3套图9中的资源的最终结果再次相加,在6个时钟周期下生成最终结果。这里,参考图1中的第一个场景,卷积核尺寸为3x3,一个输出特征图上的输出像素点,来自一列3个乘法器及其背后16个输入特征通道的所有像素点的乘累加结果,所以每个图1中的乘法器背后都分配了独立的一套图9中的级联阵列。

[0070] 图9中的各时钟信号的时序图可以参见图10。图10示出了图9的级联结构中的时钟信号的时序图。

[0071] 图10中,16输入特征通道的数据每4个分为一组,各自对齐进入相应的处理单元,比如,输入特征通道0-3的数据,最先更新,进入处理单元,两个时钟周期后,输入特征通道4-7的数据再更新,进入处理单元。最后输入特征通道12-15的数据,比最早的输入特征通道0-3晚6个时钟周期进入处理单元。这种控制方式,配合图9中处理单元的连接方式,能实现16个输入特征通道的数据乘加计算,在正确的时刻算出正确的数据,且过程达到流水线操作的目的。

[0072] 假设这里资源紧张,也可以调整通道方向的算力资源,从16通道减为8通道,则仅使用4个处理单元,完成图9中的前8个通道即可。不同算力会匹配不同的算力需求,有的场景计算要求帧率高,比如实时,24帧每秒;有的场景不需要实时,1帧每秒都可以满足需求。

[0073] 根据本发明的又一方面,还提供一种芯片结构,包括如上所述的乘加计算引擎。芯片结构还可以包括高速接口模块、总控模块、输入特征模块、外部储存接口模块、非线性/归

一化/池化计算模块、特征输出模块中的一个或多模块。本发明还可以实现更多的芯片结构的变化方式，在此不予赘述。

[0074] 相比现有技术，本发明的优势在于：

[0075] 利用同一套卷积引擎在应用到不同的卷积核的尺寸时，提高乘法器资源的利用率，且根据不同的卷积核的尺寸实现不同的乘法器资源分配，实现数据的动态分布，达到需要的计算方式。

[0076] 本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本公开的其它实施方案。本申请旨在涵盖本公开的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本公开的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本公开的真正范围和精神由所附的权利要求指出。

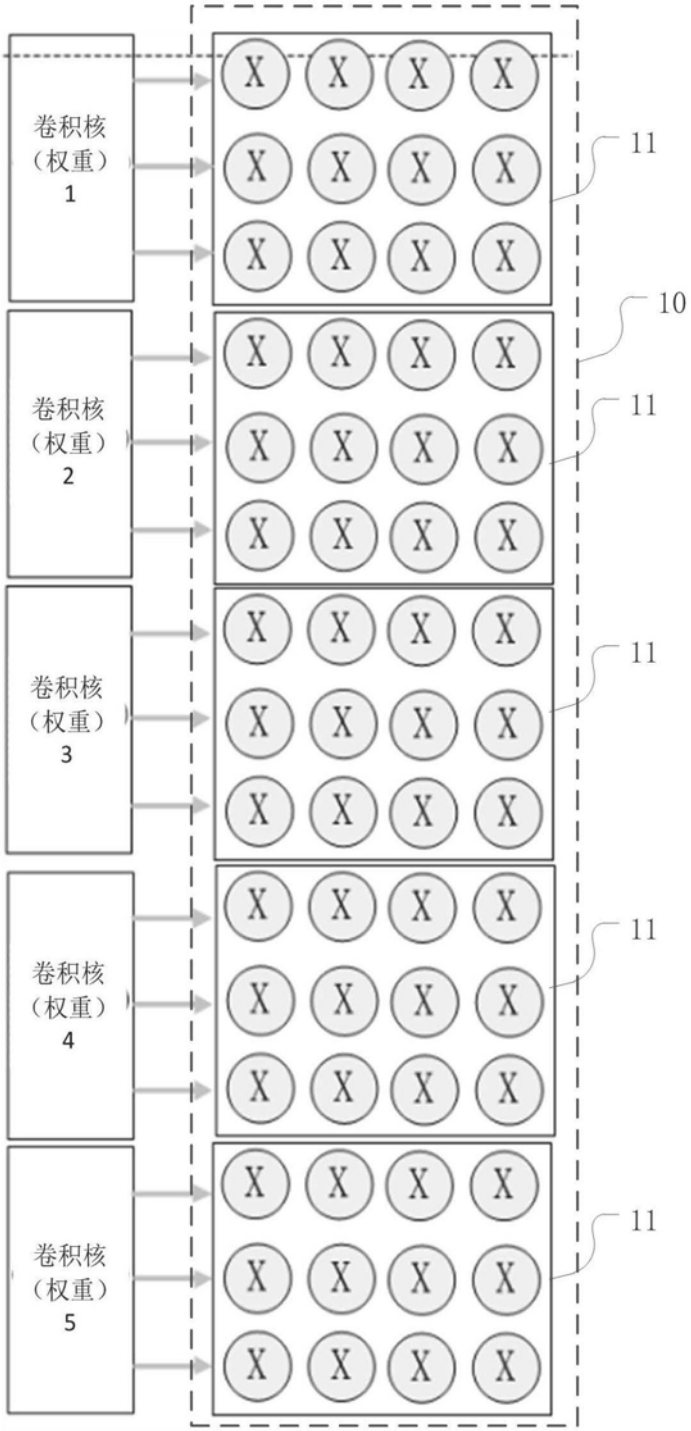


图1

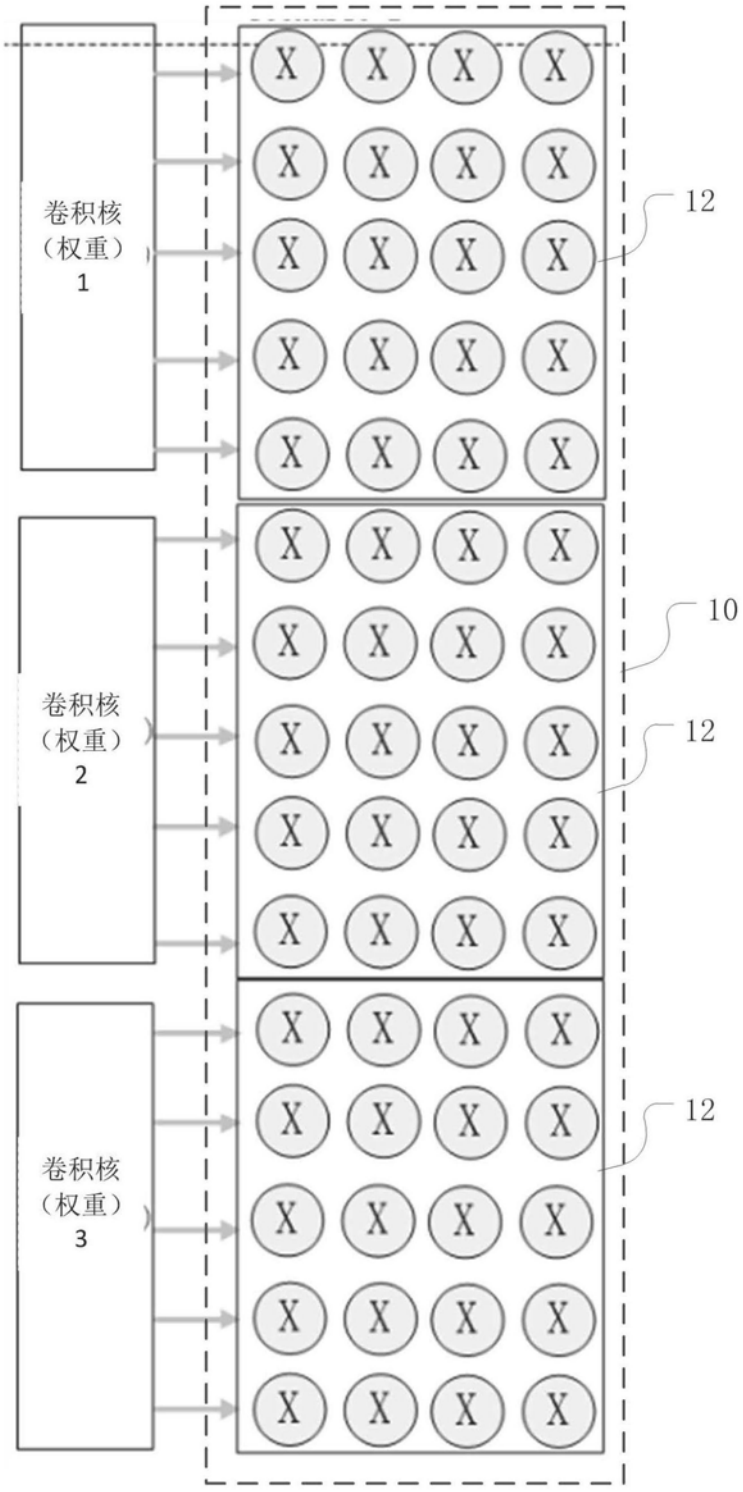


图2

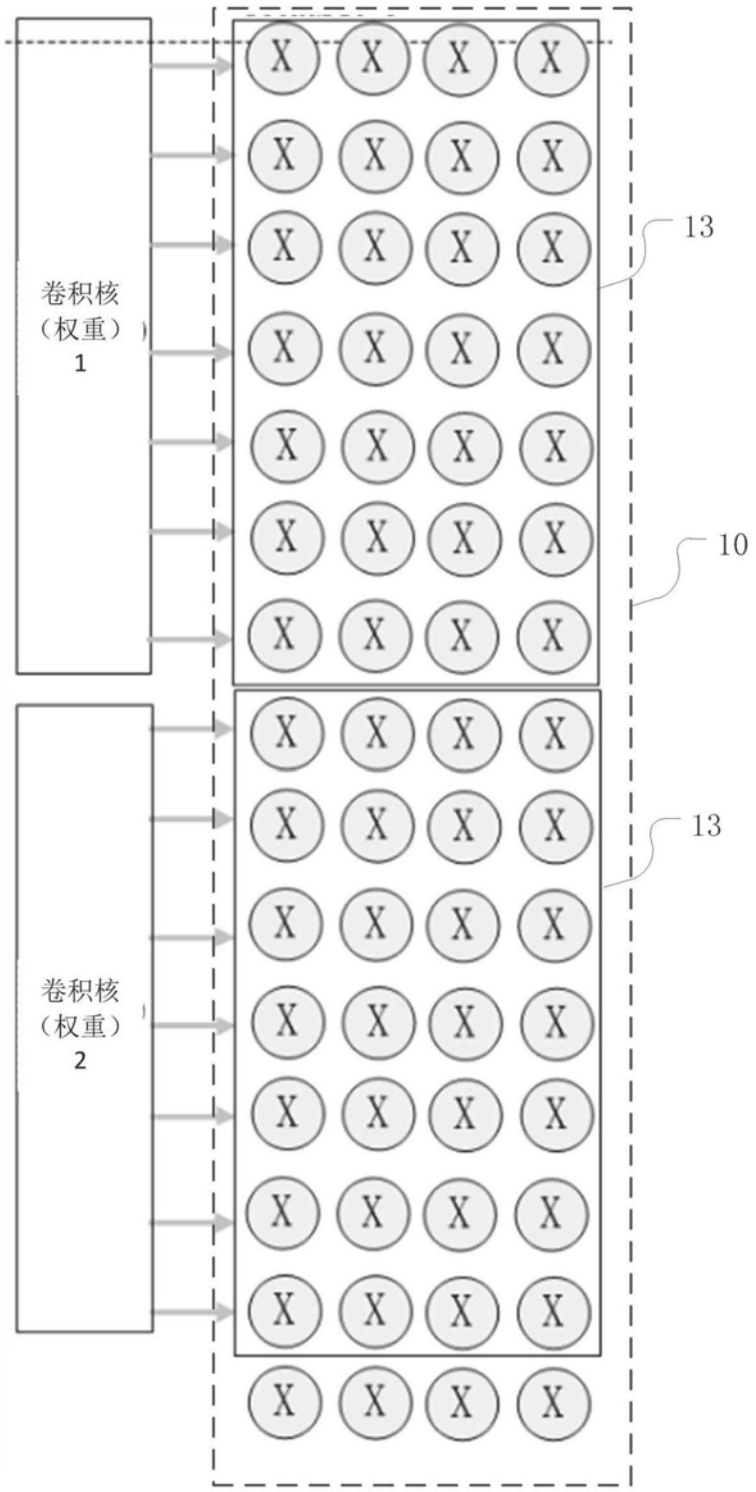


图3

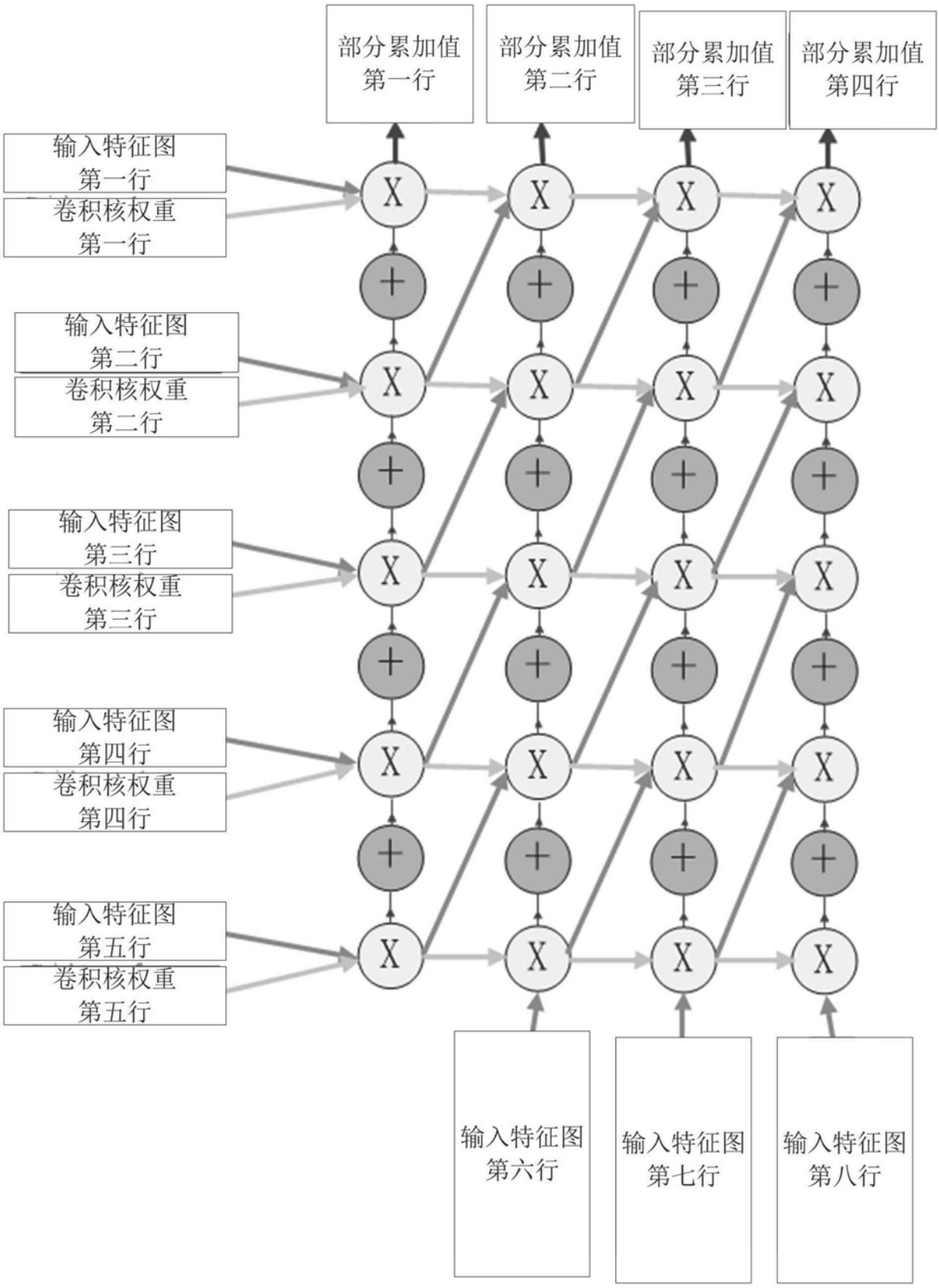

部分累加值
第一行
部分累加值
第二行
部分累加值
第三行
部分累加值
第四行
输入特征图
第一行
卷积核权重
第一行
输入特征图
第二行
卷积核权重
第二行
输入特征图
第三行
卷积核权重
第三行
输入特征图
第四行
卷积核权重
第四行
输入特征图
第五行
卷积核权重
第五行
输入特征图
第六行
输入特征图
第七行
输入特征图
第八行

图4

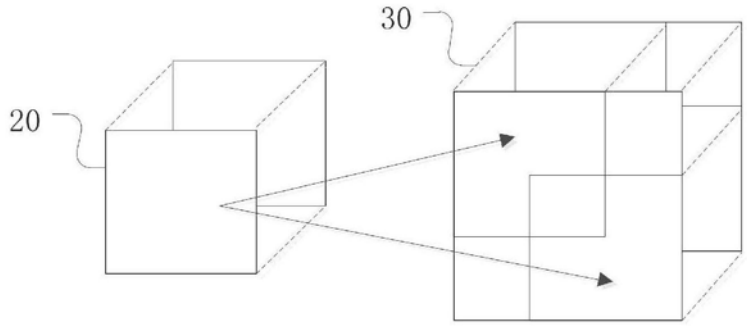


图5

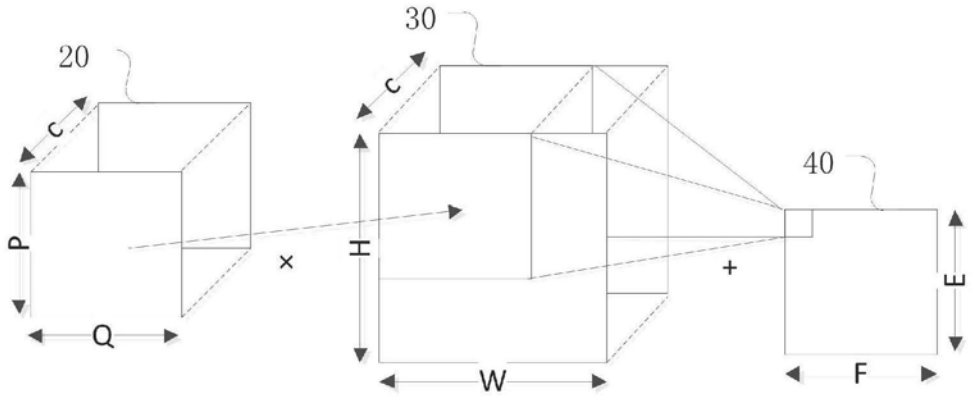


图6

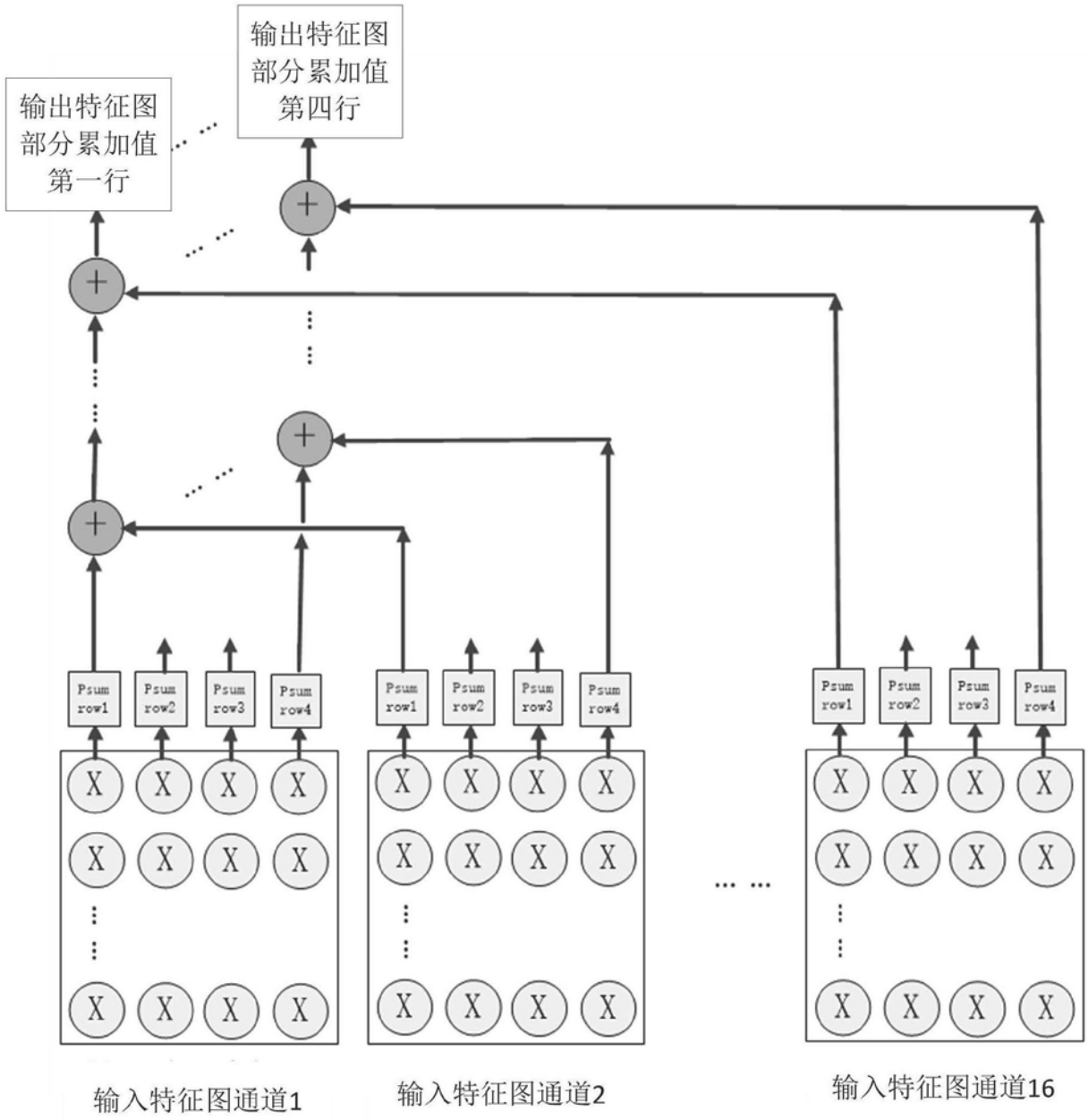

输出特征图
部分累加值
第四行
输出特征图
部分累加值
第一行
Psum row1
Psum row2
Psum row3
Psum row4
输入特征图通道1
输入特征图通道2
输入特征图通道16

图7

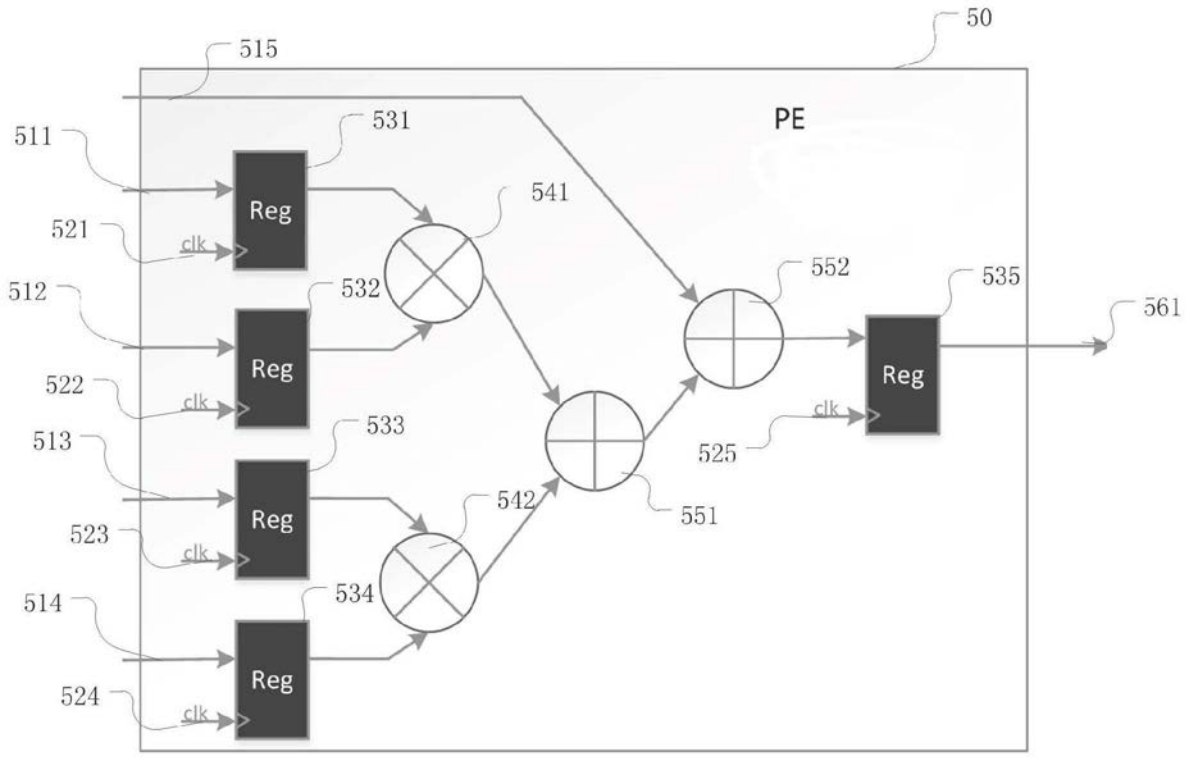

50
515
PE
511
531
Reg
521
clk
541
552
535
561
512
532
Reg
522
clk
533
513
Reg
523
clk
542
525
551
514
534
Reg
524
clk

图8

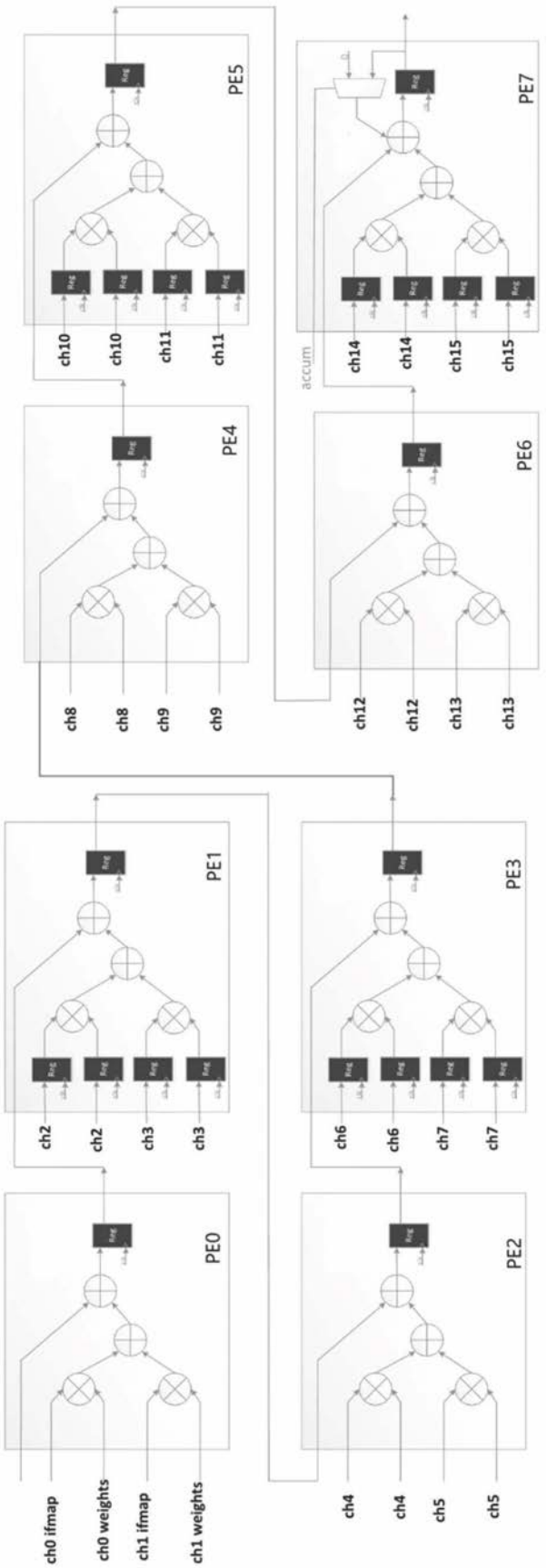

PE5
PE7
PE4
PE6
PE1
PE3
PE0
PE2
accum
ch10
ch10
ch11
ch11
ch14
ch14
ch15
ch15
ch8
ch8
ch9
ch9
ch12
ch12
ch13
ch13
ch2
ch2
ch3
ch3
ch6
ch6
ch7
ch7
ch0 ifmap
ch0 weights
ch1 ifmap
ch1 weights
ch4
ch4
ch5
ch5

图9

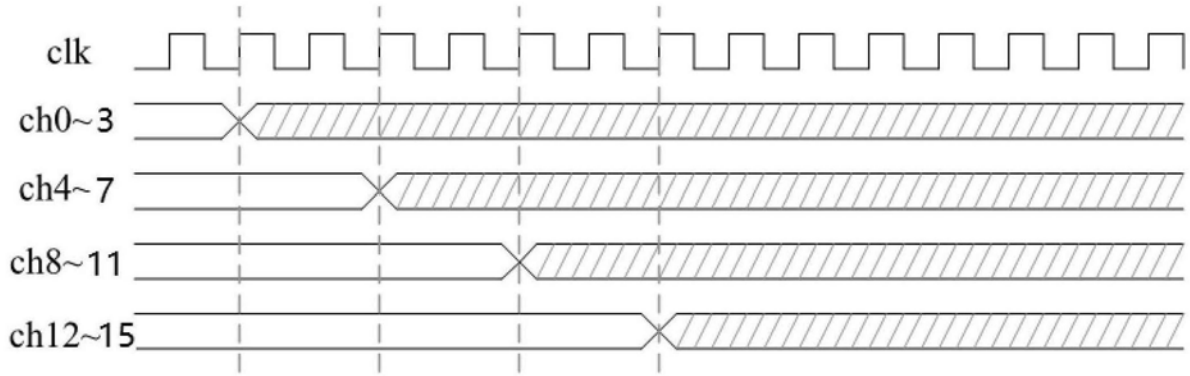


图10