(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 108182471 B
(45)授权公告日 2022.02.15

(21)申请号 201810068051.6

(22)申请日 2018.01.24

(65)同一申请的已公布的文献号
申请公布号 CN 108182471 A

(43)申请公布日 2018.06.19

(73)专利权人 上海岳芯电子科技有限公司
地址 201821 上海市嘉定区福海路1011号3幢B区1538室

(72)发明人 梁晓峣 伍骏

(74)专利代理机构 上海国智知识产权代理事务所(普通合伙) 31274
代理人 潘建玲

(51)Int.Cl.
G06N 3/04(2006.01)
G06N 5/04(2006.01)
G06V 10/94(2022.01)
G06V 10/82(2022.01)

(56)对比文件
CN 107533667 A,2018.01.02
WO 2016186811 A1,2016.11.24
CN 105631854 A,2016.06.01
CN 106951395 A,2017.07.14
CN 106250103 A,2016.12.21
Xuechao Wei et al.Automated systolic array architecture synthesis for high throughput CNN inference on FPGAs.《2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)》.2017,第1-6页.

审查员 田松雪

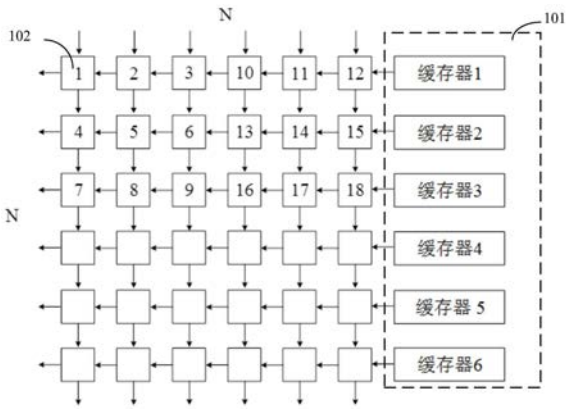
权利要求书1页 说明书7页 附图4页

(54)发明名称

一种卷积神经网络推理加速器及方法

(57)摘要

本发明公开了一种卷积神经网络推理加速器及方法,所述加速器包括:输入图像缓存器模块,包括N个缓存器,用于载入输入图像数据,每个缓存器存储图像对应一行的数据;N*N个运算单元,连接所述输入图像缓存器模块,用于进行卷积运算,所述N*N个运算单元支持图像数据在相邻运算单元间传递的脉动形式,其连接缓存器的运算单元从缓存器中读取图像数据,剩余的运算单元从邻近的运算单元读取图像数据,本发明针对卷积神经网络带来的数据可复用性设计双向脉动阵列,提高了数据的加载效率,从而加速了卷积神经网络。



CN 108182471 B

1.一种卷积神经网络推理加速器,包括:

输入图像缓存器模块,包括N个缓存器,用于载入输入图像数据,每个缓存器存储图像对应一行的数据,所述输入图像缓存器模块每隔k行放置一个额外的缓存器,用于在卷积运算换行前,缓存图像下一行的数据,以提高数据加载效率;

N*N个运算单元,连接所述输入图像缓存器模块,用于进行卷积运算,于进行卷积运算时,所述N*N个运算单元支持图像数据在相邻运算单元间传递的脉动形式,其连接缓存器的运算单元从缓存器中读取图像数据,其他的运算单元从邻近的运算单元读取图像数据,在卷积运算中,各卷积核对图像的每k行数据进行卷积,而非图像的一行数据,其中,k为卷积核大小,利用图像每k行数据开始做卷积运算时填充运算单元的时间,于所述额外的缓存器载入图像k+1行的数据,并将权值通过运算单元的寄存器向下滑动到下方邻近的寄存器中,换行前的k-1行图像数据保留在原缓存器复用,从而实现高效的卷积换行操作;

多级流水线加法器,包括多个加法器,用于对每列的运算单元的输出结果进行累加;

多路选择器,用于根据控制信号,根据卷积核的大小选择特定的列进行相加。

2.如权利要求1所述的一种卷积神经网络推理加速器,其特征在于:于进行卷积运算时,针对卷积运算换行,所述N*N个运算单元中的权值寄存器支持上下移动原有的权值数据,并接收其他寄存器传输过来的权值数据。

3.如权利要求1所述的一种卷积神经网络推理加速器,其特征在于:将同一通道的不同卷积核横向存放在运算单元的权值寄存器中,将对应图像的不同通道的同位置的卷积核纵向存放在运算单元的权值寄存器。

4.一种卷积神经网络推理加速方法,包括如下步骤:

步骤S1,将输入图像数据载入输入图像缓存器模块的多个输入图像缓存器中,同时向运算单元中的权值寄存器载入权值数据;

步骤S2,对N*N个运算单元进行卷积运算,于进行卷积运算时,所述N*N个运算单元支持图像数据在相邻运算单元间传递的脉动形式,其连接缓存器的运算单元从缓存器中读取图像数据,其他的运算单元从邻近的运算单元读取图像数据,在卷积运算中,各卷积核对图像的每k行数据进行卷积,而非图像的一行数据,其中,k为卷积核大小,利用图像每k行数据开始做卷积运算时填充运算单元的时间,于所述额外的缓存器载入图像k+1行的数据,并将权值通过运算单元的寄存器向下滑动到下方邻近的寄存器中,换行前的k-1行图像数据保留在原缓存器复用,从而实现高效的卷积换行操作;

步骤S3,当运算单元完成一次乘法后,对每一列的运算单元的输出值进行多级流水线形式累加运算;

步骤S4,在进行了多级流水线形式累加运算后,根据卷积核的大小,选取特定列的累加和进一步加法运算,得到N/k个输出结果,所述卷积核大小为k*k。

一种卷积神经网络推理加速器及方法

技术领域

[0001] 本发明涉及针对卷积神经网络的专用加速架构,特别是涉及一种用于卷积神经网络的推理阶段,加速其推理运算速度的基于双向脉动与多级流水线的卷积神经网络推理加速器及方法。

背景技术

[0002] 卷积神经网络是一种前馈神经网络,常应用于图像识别,一般包括卷积层、池化层和全连接层。卷积层的卷积操作是,卷积核中的每一个权值与其对应的输入数据点对点相乘,然后将点乘结果累加,得到输出的一个数据,之后,根据卷积层的步长设定,滑动卷积核,重复上述操作。

[0003] 目前,针对神经网络的加速架构很多,包括通用处理器做神经网络加速,专用ASIC加速架构,以及利用新型材料对神经网络进行加速。

[0004] Nvidia公司提出了一种基于GPU通用架构的加速器,它既支持GPU的传统运算,又加速了神经网络的计算,这一架构的优势在于可以保留原有的通用计算框架,支持cuda语言编程,对习惯于cuda编程的程序员来说,该架构易于上手,但是缺点在于为了支持通用计算,无法灵活地根据神经网络运算的特点改变原本的GPU架构,另外,这种设计为了灵活性,能耗是无可避免的,因此该架构对神经网络的加速不是最优化的。

[0005] 专用ASIC加速架构多种多样,Xie Y等人提出“an instruction set architecture for neural networks”(International Symposium on Computer Architecture.IEEE Press,2016:393-405),考虑到机器学习的算法是具有专用性的,一种算法对某个数据集效果特别好,换了一个数据集后准确率可能直线下降,而投入市场的芯片,面向的应用多种多样,不可能用一种机器学习的算法就能完全解决,因此,为了能支持多种机器学习的算法,该设计分析了各种神经网络、机器学习算法的运算特点,比如,矩阵乘向量的运算、向量乘标量的运算都会出现在各类神经网络中,将运算细化到矩阵、向量这一层级,设计了一套通用于各类算法的指令集。但是该设计的指令粒度太细,导致流水线过长,在执行过程中,更可能出现阻塞,因此,该设计的架构的运算性能并不是很好。Chen Y,Luo等于“A Machine-Learning Supercomputer”(Ieee/acm International Symposium on Microarchitecture.IEEE,2014:609-622.)中提出了一种针对神经网络的芯片,该芯片将神经网络的全部权值存储在片上,通过调度使得所需的权值可以快速被找到,解决处理器常见的数据加载的瓶颈问题。但是随着神经网络的发展,网络规模越来越大,权值信息越来越多,如果要存储所有的权值信息,那么耗费的硬件资源将不可想象,因此该架构对于存储方面,过于大方,不符合实际应用的需求。Du Z等人于“shifting vision processing closer to the sensor”(International Symposium on Computer Architecture.ACM,2015:92-104.)中提出利用卷积神经网络的权值共享的特性,将权值整体载入静态随机存储器中,减少了访问动态随机存储器带来的内存开销,但是由于大型网络的权值太多,而静态随机存储器的容量很小,因此该设计只能应用于非常小的网络中,应用面不广。

[0006] 利用新型材料对神经网络进行加速的工作也有很多，Shafiee A等人于“AConvolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars”(International Symposium on Computer Architecture.IEEE Press,2016:14-26)利用了新型材料忆阻器可用于存储又可用于计算矩阵乘加的特性，实现了神经网络的前向传播过程。Song L等人于“A Pipelined ReRAM-Based Accelerator for Deep Learning”(IEEE International Symposium on High PERFORMANCE Computer Architecture.IEEE,2017:541-552.)同样利用了忆阻器的特性，实现了卷积神经网络的前向传播与反向传播，为后人提供了一种加速器设计的新思路，但是利用新型材料做加速器设计的工作都有一个问题，那就是新型材料由于还未投入市场，其真实性能无法考量，暂时无法应用在实际的开发中。

发明内容

[0007] 为克服上述现有技术存在的不足，本发明之目的在于提供一种卷积神经网络推理加速器及方法，以针对卷积神经网络带来的数据可复用性，设计双向脉动阵列，提高数据的加载效率，从而加速卷积神经网络。

[0008] 为达上述及其它目的，本发明提出一种卷积神经网络推理加速器，包括：

[0009] 输入图像缓存器模块，包括N个缓存器，用于载入输入图像数据，每个缓存器存储图像对应一行的数据；

[0010] N*N个运算单元，连接所述输入图像缓存器模块，用于进行卷积运算，于进行卷积运算时，所述N*N个运算单元支持图像数据在相邻运算单元间传递的脉动形式，其连接缓存器的运算单元从缓存器中读取图像数据，其他的运算单元从邻近的运算单元读取图像数据。

[0011] 优选地，于进行卷积运算时，针对卷积运算换行，所述N*N个运算单元中的权值寄存器支持上下移动原有的权值数据，并接收其他寄存器传输过来的权值数据。

[0012] 优选地，所述输入图像缓存器模块每隔k行放置一个额外的缓存器，用于在卷积运算换行前，缓存图像下一行的数据，以提高数据加载效率，并结合支持权值上下滑动的脉动阵列，可以实现高效的卷积换行操作。

[0013] 优选地，将同一通道的不同卷积核横向存放在运算单元的权值寄存器中，将对应图像的不同通道的同位置的卷积核纵向存放在运算单元的权值寄存器。

[0014] 优选地，所述加速器还包括多级流水线加法器，包括多个加法器，用于对每列的运算单元的输出结果进行累加。

[0015] 优选地，所述加速器还包括多路选择器，用于根据控制信号，选择特定的列进行加法运算。

[0016] 为达到上述目的，本发明还提供一种卷积神经网络推理加速方法，包括如下步骤：

[0017] 步骤S1，将输入图像数据载入输入图像缓存器模块的多个输入图像缓存器中，同时向运算单元中的权值寄存器载入权值数据；

[0018] 步骤S2，对N*N个运算单元进行卷积运算，于进行卷积运算时，所述N*N个运算单元支持图像数据在相邻运算单元间传递的脉动形式，其连接缓存器的运算单元从缓存器中读取图像数据，其他的运算单元从邻近的运算单元读取图像数据。

[0019] 优选地，所述方法还包括：

[0020] 当运算单元完成一次乘法后，对每一列的运算单元的输出值进行多级流水线形式累加运算。

[0021] 优选地，所述方法还包括：

[0022] 在进行了多级流水线形式累加运算后，根据卷积核的大小，选取特定列的累加和进一步加法运算，得到N/k个输出结果，所述卷积核大小为k*k。

[0023] 优选地，所述方法还包括：

[0024] 当完成了图像N/(k+1)个通道k行的卷积运算后，需要进行图像的换行操作，针对卷积运算换行，所述N*N个运算单元中的权值寄存器支持上下移动原有的权值数据，并接收其他寄存器传输过来的权值数据。

[0025] 现有技术相比，本发明一种卷积神经网络推理加速器针对卷积神经网络带来的数据可复用性，设计了双向脉动阵列，以提高数据的加载效率，从而加速卷积神经网络，同时，本发明还设计了多级流水线加法器结构进行卷积加法，提高了加法效率。

附图说明

[0026] 图1为本发明一种卷积神经网络推理加速器之一实施例的架构示意图；

[0027] 图2a为本发明具体实施例中多卷积核对多通道图像进行卷积运算的示意图；

[0028] 图2b为不同周期卷积核与图像进行卷积的示意图；

[0029] 图3为本发明具体实施例中支持多级流水线的加法器示意图；

[0030] 图4为本发明具体实施例中支持多种卷积核运算的多路选择器示意图；

[0031] 图5a为本发明具体实施例中完成数据初始化载入后图像数据及权值数据的排布情况示意图；

[0032] 图5b为本发明具体实施例中支持输入数据向左滑动示意图；

[0033] 图5c为本发明具体实施例中支持权值向下滑动示意图；

[0034] 图6为本发明一种卷积神经网络推理加速方法的步骤流程图。

具体实施方式

[0035] 以下通过特定的具体实例并结合附图说明本发明的实施方式，本领域技术人员可由本说明书所揭示的内容轻易地了解本发明的其它优点与功效。本发明亦可通过其它不同的具体实例加以施行或应用，本说明书中的各项细节亦可基于不同观点与应用，在不背离本发明的精神下进行各种修饰与变更。

[0036] 图1为本发明一种卷积神经网络推理加速器之一实施例的架构示意图。如图1所示，本发明一种卷积神经网络推理加速器，基于双向脉动与多级流水线，包括：

[0037] 输入图像缓存器模块101，包括N个缓存器，用于载入输入图像数据，每个缓存器存储图像一行的数据。

[0038] 由于在动态随机存储器中，图像数据是按行顺序存储的，这意味着从动态随机存储器读取图像数据会一行行读入图像数据，但是在卷积运算中，卷积核会先对图像的前k行的k列数据(k为卷积核大小)进行卷积，而非图像的一行数据，因此本发明设计了输入图像缓存器模块，每一个缓存器存储图像一行的数据，以支持正确有效的卷积运算。

[0039] 在本发明具体实施例中，假设运算单元共有N*N个，卷积核大小为k*k，则缓存器共有N个，可载入N/(k+1)个通道的图像数据。具体来说，将第一通道的前k行数据载入输入图像缓存器1,2,3…k中，将第二通道的前k行载入缓存器k+2,k+3,…2k+1中，依次类推。

[0040] 较佳地，每k个缓存器预留一个额外的缓存器来存储图像第k+1行的数据，以提前缓存下一行的新的图像数据。在本发明中，卷积运算在换行时，图像数据会与换行前的图像数据有(k-1)行的数据可以复用，为了利用这一特点，因此在本发明中，每隔k行放置一个额外的缓存器，用于提前缓存下一行的新的图像数据，而可复用的(k-1)行图像数据保留在原本的缓存器中。

[0041] N*N个运算单元102，用于进行卷积运算，N*N个运算单元102针对卷积神经网络的滑动计算，支持横向相邻运算单元复用图像数据。具体地说，每个运算单元中有两个寄存器单元，分别用于存储输入图像数据和权值，以及一个乘法器，对两个寄存器中的数据进行乘法操作，其中存储输入图像数据的寄存器支持横向相邻运算单元复用图像数据的操作，权值寄存器支持上下移动权值数据的操作。

[0042] 一般来说，不同的卷积核在对同一个输入图像数据进行卷积时，最直接的方法是，其所对应的运算单元都从输入图像缓存器模块中读取数据，这样做的缺点在于从缓存器中读取了重复的图像数据，增加了访问带宽，容易造成数据读取冲突。而根据图2a所示，不同的卷积核在对同一个输入图像数据进行卷积时，灰色矩形块表示的输入图像数据是可复用的，根据图2b所示，在周期1和周期2，卷积核在一张输入图像上滑动时，灰色矩形块覆盖的输入图像数据，是两次操作中可复用的数据，这都是卷积运算带来的数据可复用性，基于此，本发明设计了一种支持相邻运算单元复用图像数据的脉动形式，能支持连接缓存器的运算单元从缓存器中读取图像数据，而剩余的运算单元从邻近的运算单元读取图像数据，以此来避免从缓存器中读取大量可复用数据，同时又能实现多个卷积核同时对一张图像进行卷积运算的功能，极大地减少了访问带宽，减少数据访问冲突，提高了架构的运算性能。

[0043] 较佳地，N*N个运算单元102针对卷积运算换行，支持权值上下滑动。具体地，卷积核会从图像的前k(k为卷积核大小)行开始滑动，滑到前k行的末尾后，卷积核会向下滑动一行，接着，重复上述操作。换行意味着输入图像数据与之前载入缓存器中的数据不一样了，最直接的办法是，保留运算单元的权值寄存器中的权值数据，将图像数据缓存器中的数据擦除后，重新写入新的k行的输入图像数据，但是实际上，换行后的图像数据与换行前的图像数据有(k-1)行图像数据是可复用，若可以避免写重复数据，则可以节省载入数据的时间，因此，本发明设计了支持权值上下滑动的脉动形式，为此每隔k个缓存器增加了一个空缓存器，用于存储换行后新的一行图像数据，可复用的多行图像数据保留在原本的图像数据缓存器中，同时，令运算单元中的权值寄存器支持上下移动原有的数据，并接收其他寄存器传输过来的权值数据，通过这样的改进，可以避免将图像数据缓存器中的图像数据擦除后重新写入新的k行的图像数据，节约了大量的数据加载时间。

[0044] 优选地，本发明之卷积神经网络推理加速器还包括：

[0045] 多级流水线加法器，包括多个加法器，用于对每列的运算单元的输出结果进行累加。由卷积运算规则可知，每列的运算单元的输出结果要进行累加运算，为提高加法效率，本发明设计了多级流水线加法器，将累加操作变为流水线形式，可有效的提高加法效率。

[0046] 多路选择器，用于根据控制信号，选择特定的列进行加法运算。同一列运算单元的

乘积结果累加后，需要根据卷积核大小进行进一步累加，即根据卷积核的大小决定特定的列进行加法，得到一个输出结果，因此本发明设计了多路选择器，根据控制信号，选择特定的列进行加法运算，即可支持多种卷积核的卷积运算。

[0047] 以下将配合一具体实施例来进一步说明本发明：如图1所示，所述加速器其包括矩形框表示的N*N个运算单元，包括N个缓冲器的输入图像缓存器模块，每个运算单元中有两个寄存器单元，分别用于存储输入图像数据和权值，以及一个乘法器，对两个寄存器中的数据进行乘法操作，由卷积运算规则可知，每列的运算单元的输出结果要进行累加运算，如图3所示，在本发明具体实施例中，采用了多级流水线加法器对每列的运算单元的输出结果进行累加。

[0048] 每列完成累加运算后，特定列还需要进行进一步加法运算，如图4所示，采用多路选择器，根据控制信号，选择特定的列进行加法运算，即可支持多种卷积核的卷积运算。

[0049] 具体地，首先，将输入图像数据载入输入图像缓存器模块的多个输入图像缓存器中，假设运算单元共有N*N个，卷积核大小为k*k，则缓存器共有N个，可载入N/(k+1)个通道的图像数据。具体来说，第一通道的前k行数据载入输入图像缓存器1,2,3…k中，将第二通道的前k行载入缓存器k+2,k+3,…2k+1中，依次类推，较佳地，每k个缓存器会预留一个额外的缓存器来存储图像第k+1行的数据。

[0050] 在载入图像数据时，同时向运算单元中的权值寄存器载入权值数据。在本发明具体实施例中，卷积核排布规则如图5a所示，将同一通道的不同卷积核横向存放在运算单元的权值寄存器中，将对应图像的不同通道的同位置的卷积核纵向存放在运算单元的权值寄存器中。从图5a可以看出，若横向的运算单元共有N个，则可以放置N/k个不同的卷积核，如果该卷积层的卷积核个数大于N/k个，则需要载入新的卷积核，保留原来的图像数据，重复一遍卷积运算；若纵向的运算单元共有N个，则可以放置N/(k+1)个通道的卷积核，同理，如果该图像的通道数大于N/(k+1)个，则需要载入剩余通道对应的卷积核和输入图像数据，重复一遍卷积运算。

[0051] 之后，开始卷积运算，卷积运算的滑动操作由图像数据向左滑动的方式得以实现，如图5b所示，当运算单元做完一次乘法运算后，会发出控制指令，令图像数据在相邻运算单元可相互传输，图中的输入图像缓存器会将一个图像数据向左边相连的运算单元中传输，接收到图像数据的运算单元会向其左侧邻近的运算单元传输可复用的图像数据，1个周期后，只有最右侧的一列运算单元接收了缓存器中的新的图像数据，其他运算单元均在复用其右侧运算单元点乘过的图像数据，当所有运算单元都得到新的图像数据后，运算单元即可统一做乘法运算，在图中，即第一行最右侧的运算单元在某周期完成了1*9的点乘后，会接收输入图像缓存器传来的10，而1则会传输到其左侧的运算单元中，替代7的位置，而7会被传输到其左侧的单元中，依次类推。需要注意的是，由于缓存器中的图像数据只会向最右侧的运算单元传输数据，因此，图像每k行数据开始做卷积运算时，都有一小段填充运算单元的时间。在该过程中，额外的缓存器开始载入图像下一行的数据。

[0052] 根据卷积运算的规则，当运算单元完成一次乘法后，每一列的运算单元的输出值在进行了图3的多级流水线形式累加运算后，会根据卷积核的大小，选取特定列的累加和进一步加法运算，如图4所示。经过图3和图4的运算后，即可得到N/k个输出结果。

[0053] 当完成了图像N/(k+1)个通道k行的卷积运算后，需要进行图像的换行操作，输入

图像缓存器k+1,2k+2…在之前的卷积运算时已载入图像的N/(k+1)个通道的第k+1行数据，此时，输入图像缓存器1,k+2…中的图像数据将会被清空，因为它们不会再被复用，而缓存器2,3,…,k,k+3,k+4,…,k+5,…中的图像数据将被保留。此时，为了匹配新的图像数据，权值数据要整体移动，除了最下方一行的权值数据将移动到最上方的运算单元外，其他运算单元中的权值数据均会移动到其下方相连的运算单元。如图5c所示，第一排的运算单元中的权值数据2,4,3,2,9将会替代第二排运算单元中的权值数据，而第二排的运算单元中的权值数据3,7,3,5,2将会替代第三排运算单元中的权值数据，依次类推，另外，第1行的图像数据被清空，第二行的图像数据48,39,92,38,47,33,61,81以及第三行的图像数据82,29,30,98,67,78,91,73将被保留，第四行的图像数据89,90,29,39,42,21,35已完成载入。

[0054] 当图像完成换行操作后，运算单元重复如图5b所描述的滑动计算卷积和预填充下一行图像数据的操作。

[0055] 图6为本发明一种卷积神经网络推理加速方法的步骤流程图。如图6所示，本发明一种卷积神经网络推理加速方法，包括如下步骤：

[0056] 步骤S1，将输入图像数据载入输入图像缓存器模块的多个输入图像缓存器中，同时向运算单元中的权值寄存器载入权值数据。假设运算单元共有N*N个，卷积核大小为k*k，则输入图像缓存器共有N个，可载入N/(k+1)个通道的图像数据。具体来说，第一通道的前k行数据载入输入图像缓存器1,2,3…k中，将第二通道的前k行载入缓存器k+2,k+3,…2k+1中，依次类推，较佳地，每k个缓存器会预留一个额外的缓存器来存储图像第k+1行的数据。

[0057] 步骤S2，进行卷积运算，针对卷积运算的滑动操作，支持横向相邻运算单元复用图像数据。在本发明具体实施例中，卷积运算的滑动操作由图像数据向左滑动的方式得以实现，即当运算单元做完一次乘法运算后，会发出控制指令，令图像数据在相邻运算单元可相互传输，输入图像缓存器会将一个图像数据向左边相连的运算单元中传输，接收到图像数据的运算单元会向其左侧邻近的运算单元传输可复用的图像数据，1个周期后，只有最右侧的一列运算单元接收了缓存器中的新的图像数据，其他运算单元均在复用其右侧运算单元点乘过的图像数据，当所有运算单元都得到新的图像数据后，运算单元即可统一做乘法运算。需要注意的是，由于缓存器中的图像数据只会向最右侧的运算单元传输数据，因此，图像每k行数据开始做卷积运算时，都有一小段填充运算单元的时间。在该过程中，额外的缓存器开始载入图像下一行的数据。

[0058] 优选地，本发明之卷积神经网络推理加速方法还包括：

[0059] 当运算单元完成一次乘法后，每一列的运算单元的输出值会进行多级流水线形式累加运算。

[0060] 优选地，本发明之卷积神经网络推理加速方法还包括：

[0061] 在进行了多级流水线形式累加运算后，根据卷积核的大小，选取特定列的累加和进一步加法运算，即可得到N/k个输出结果。

[0062] 优选地，当完成了图像N/(k+1)个通道k行的卷积运算后，需要进行图像的换行操作，针对卷积运算换行，本发明支持权值上下滑动。具体地，输入图像缓存器k+1,2k+2…在之前的卷积运算时已载入图像的N/(k+1)个通道的第k+1行数据，此时，输入图像缓存器1,k+2…中的图像数据将会被清空，因为它们不会再被复用，而缓存器2,3,…,k,k+3,k+4,…,k+5,…中的图像数据将被保留，此时，为了匹配新的图像数据，权值数据要整体移动，除了最

下方一行的权值数据将移动到最上方的运算单元外，其他运算单元中的权值数据均会移动到其下方相连的运算单元。

[0063] 当图像完成换行操作后，返回步骤S2重复滑动计算卷积和预填充下一行图像数据的操作。

[0064] 综上所述，本发明一种卷积神经网络推理加速器针对卷积神经网络带来的数据可复用性，设计了双向脉动阵列，以提高数据的加载效率，从而加速卷积神经网络，同时，本发明还设计了多级流水线加法器结构进行卷积加法，提高了加法效率。

[0065] 与现有技术相比，本发明具有如下优点：

[0066] (1)本发明提出了双向脉动阵列的设计，其中支持权值上下滑动的设计，充分利用了图像换行时数据的可复用性，权值配合图像数据进行滚动，用最少的开销，即可实现原本需要载入很多重复数据的换行操作。

[0067] (2)本发明设计了专用的输入图像缓存器，以支持正确的卷积操作，并配备额外的图像缓存器，配合权值上下滑动的脉动阵列，共同支持图像换行，虽然有少量的硬件开销，但避免了向缓存器写入大量重复的数据，减少了数据载入时间。

[0068] (3)本发明中的多级流水线加法器和多路选择器均是为了支持多卷积核、多通道并行运算设计的，用少量的硬件资源，即可实现最大程度的并行，使本发明的架构运算性能达到最优。

[0069] 上述实施例仅例示性说明本发明的原理及其功效，而非用于限制本发明。任何本领域技术人员均可在不违背本发明的精神及范畴下，对上述实施例进行修饰与改变。因此，本发明的权利保护范围，应如权利要求书所列。

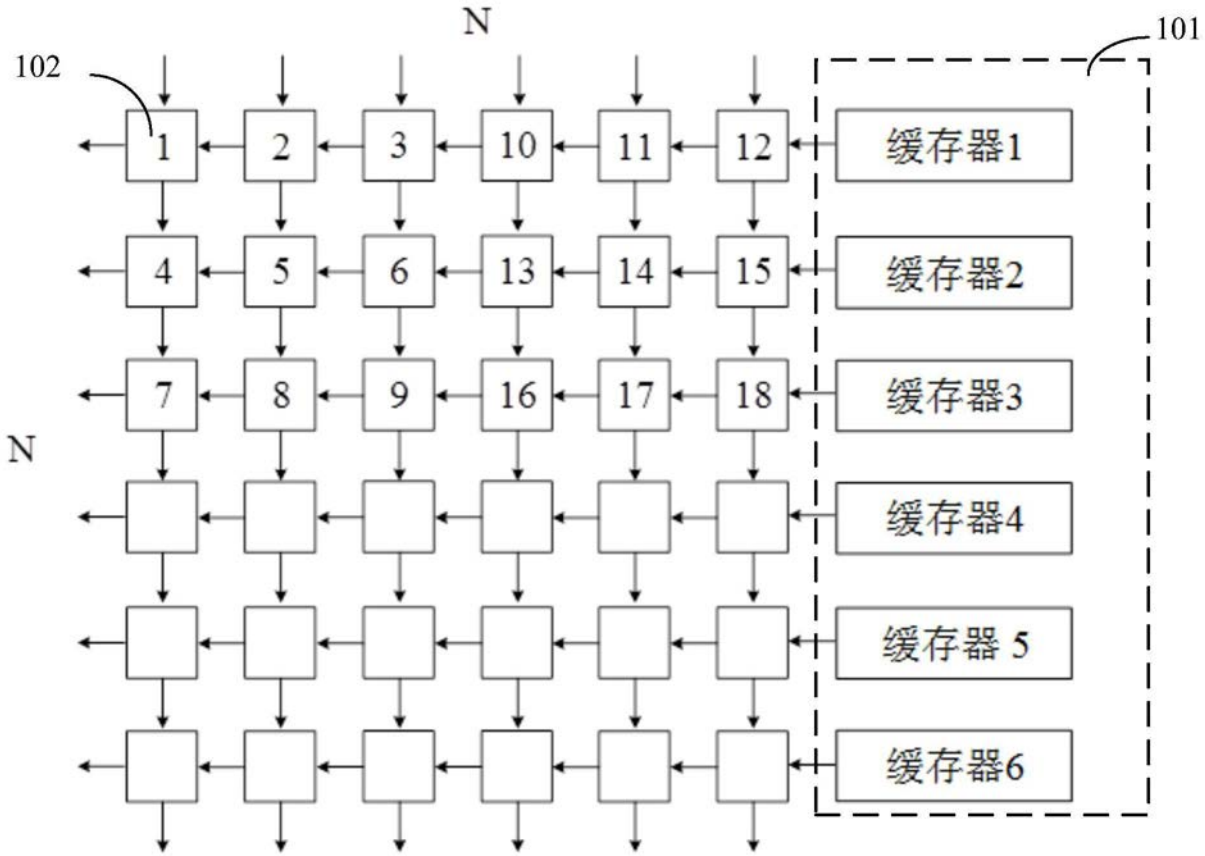

N
101
102
1
2
3
10
11
12
缓存器1
4
5
6
13
14
15
缓存器2
7
8
9
16
17
18
缓存器3
N
缓存器4
缓存器 5
缓存器6

图1

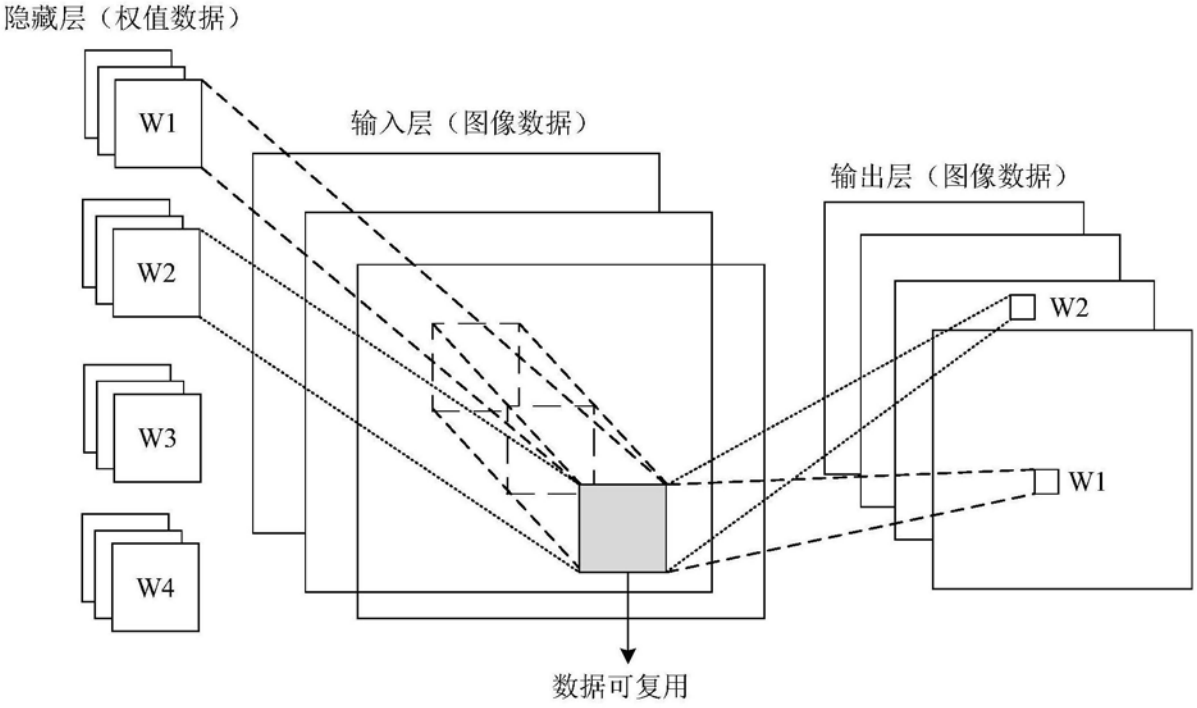

隐藏层（权值数据）
W1
输入层（图像数据）
输出层（图像数据）
W2
W2
W3
W1
W4
数据可复用

图2a

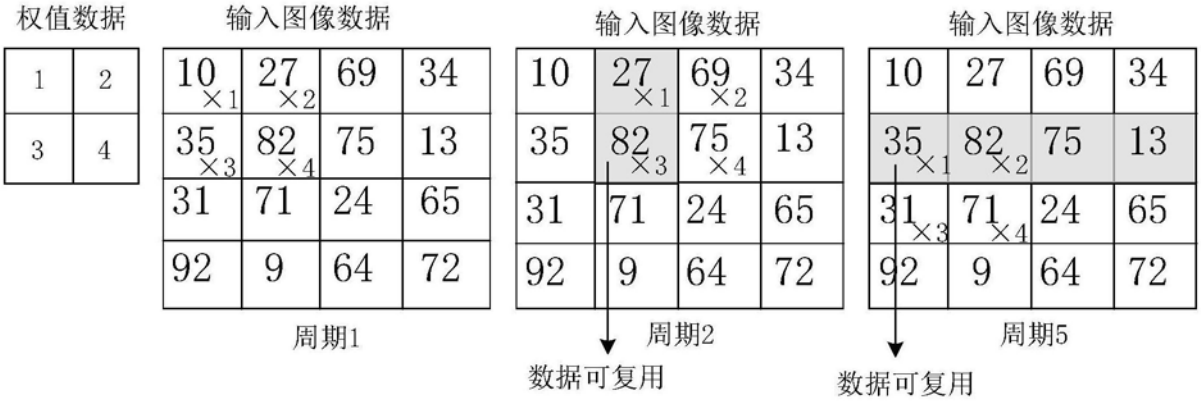

权值数据
输入图像数据
输入图像数据
输入图像数据
1
2
3
4
10 ×1
27 ×2
69
34
35 ×3
82 ×4
75
13
31
71
24
65
92
9
64
72
周期1
周期2
周期5
数据可复用
数据可复用

图2b

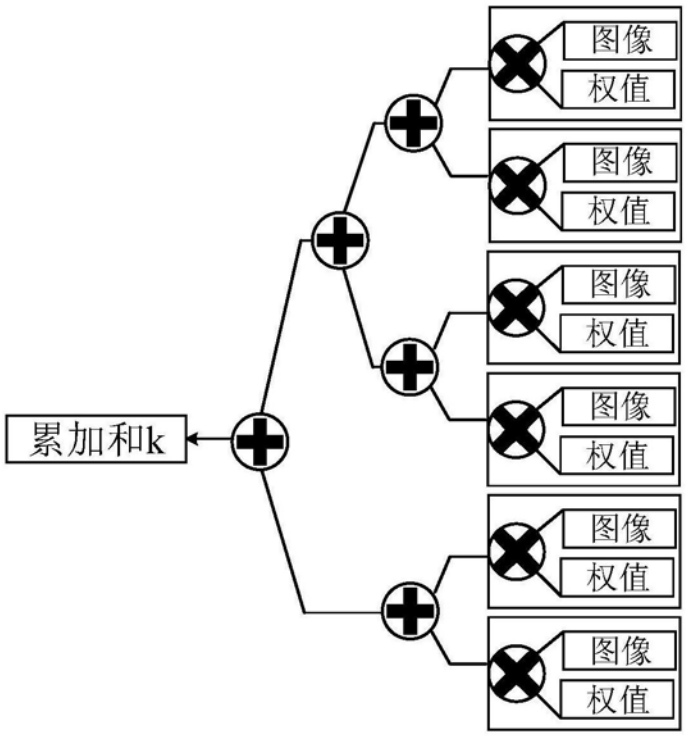

累加和k
图像
权值

图3

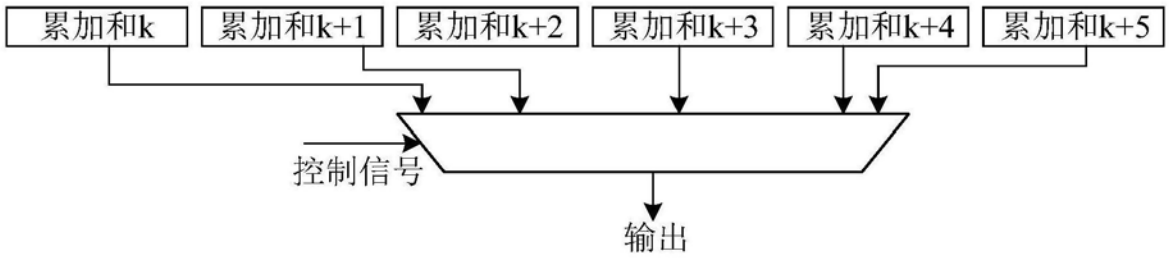

累加和k
累加和k+1
累加和k+2
累加和k+3
累加和k+4
累加和k+5
控制信号
输出

图4

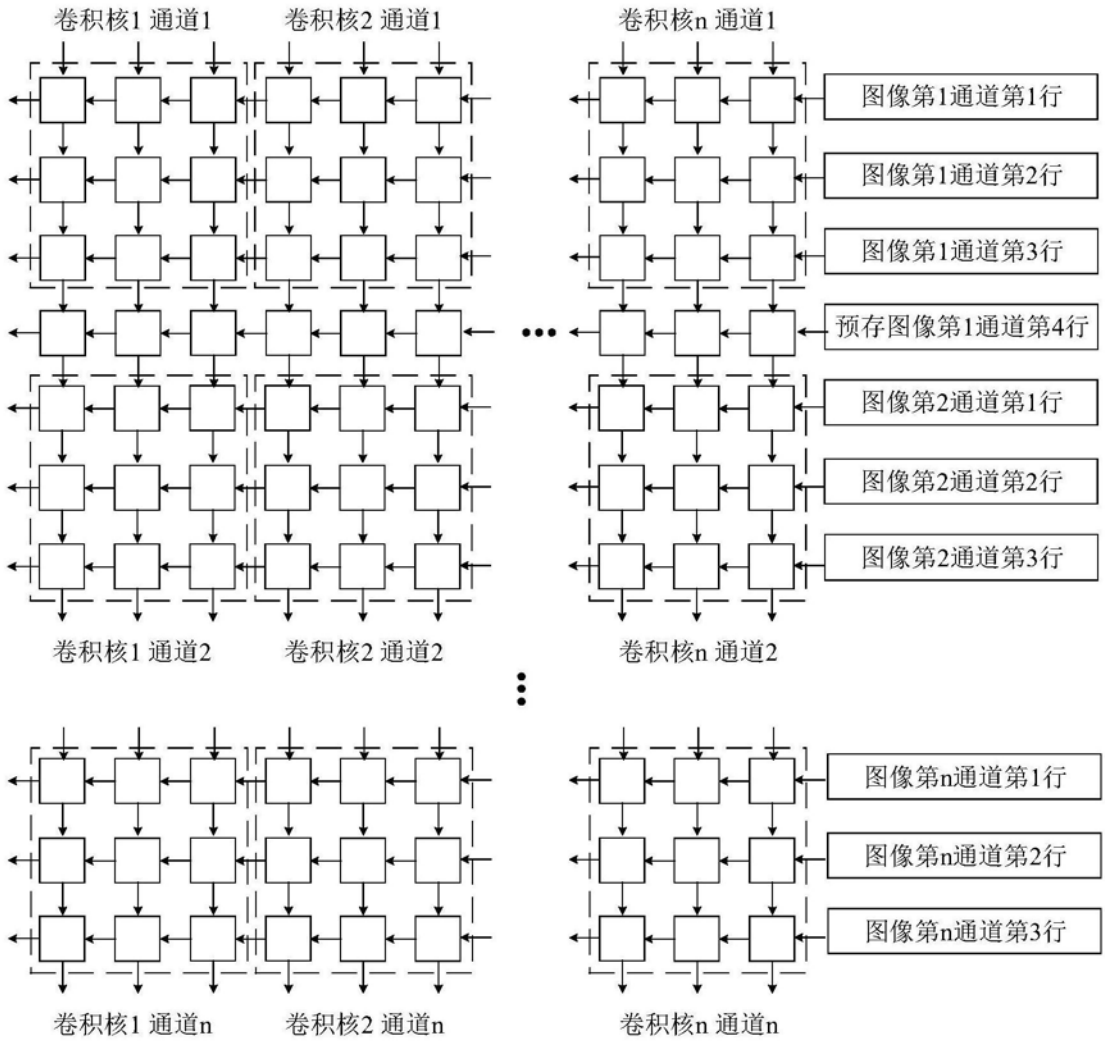

卷积核1 通道1
卷积核2 通道1
卷积核n 通道1
图像第1通道第1行
图像第1通道第2行
图像第1通道第3行
预存图像第1通道第4行
图像第2通道第1行
图像第2通道第2行
图像第2通道第3行
卷积核1 通道2
卷积核2 通道2
卷积核n 通道2
图像第n通道第1行
图像第n通道第2行
图像第n通道第3行
卷积核1 通道n
卷积核2 通道n
卷积核n 通道n

图5a

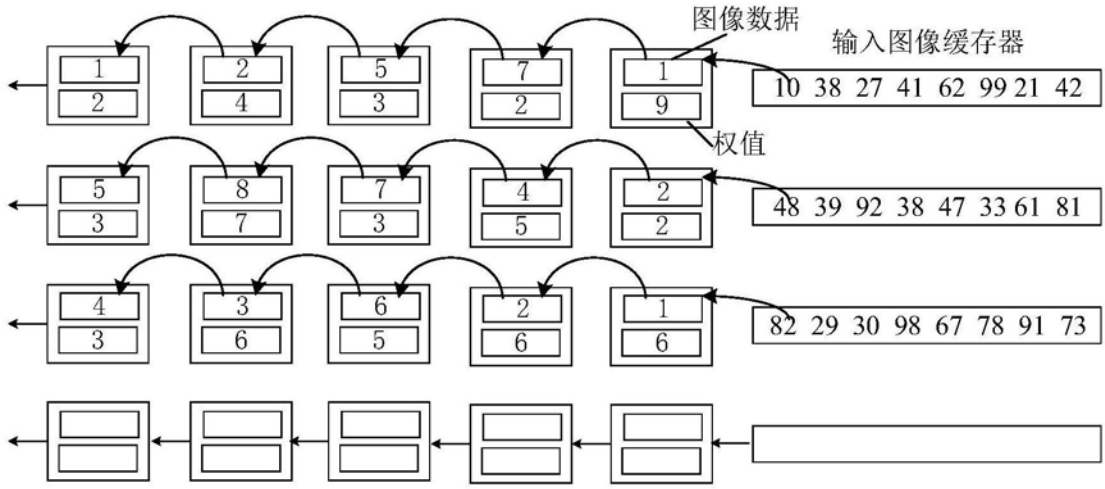

图像数据
输入图像缓存器
10 38 27 41 62 99 21 42
权值
48 39 92 38 47 33 61 81
82 29 30 98 67 78 91 73

图5b

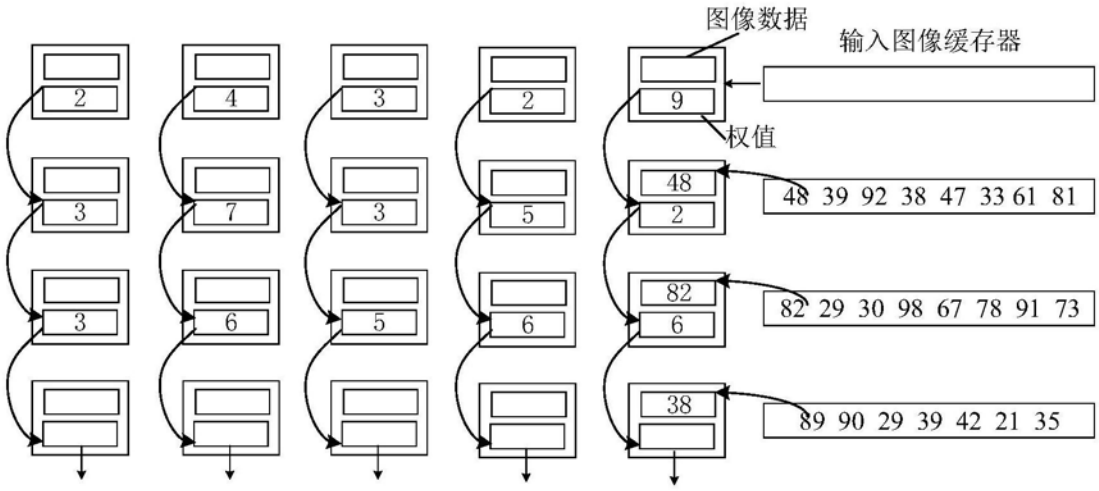

图像数据
输入图像缓存器
权值
48 39 92 38 47 33 61 81
82 29 30 98 67 78 91 73
89 90 29 39 42 21 35

图5c

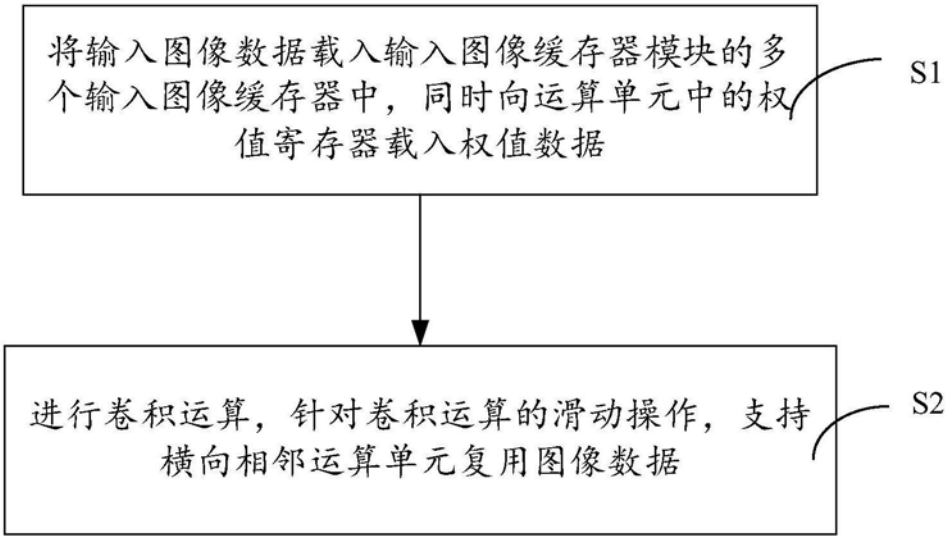

将输入图像数据载入输入图像缓存器模块的多个输入图像缓存器中，同时向运算单元中的权值寄存器载入权值数据
S1
进行卷积运算，针对卷积运算的滑动操作，支持横向相邻运算单元复用图像数据
S2

图6