(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112639839 A
(43)申请公布日 2021.04.09

(21)申请号 202080004753.3

(22)申请日 2020.05.22

(85)PCT国际申请进入国家阶段日
2021.03.02

(86)PCT国际申请的申请数据
PCT/CN2020/091883 2020.05.22

(71)申请人 深圳市大疆创新科技有限公司
地址 518057 广东省深圳市南山区高新区南区粤兴一道9号香港科大深圳产学研大楼6楼

(72)发明人 韩峰 杨康

(74)专利代理机构 北京龙双利达知识产权代理有限公司 11329
代理人 毋小妮 毛威

(51)Int.Cl.
G06N 3/08(2006.01)
G06N 3/04(2006.01)
G06F 17/16(2006.01)
G06F 9/30(2006.01)

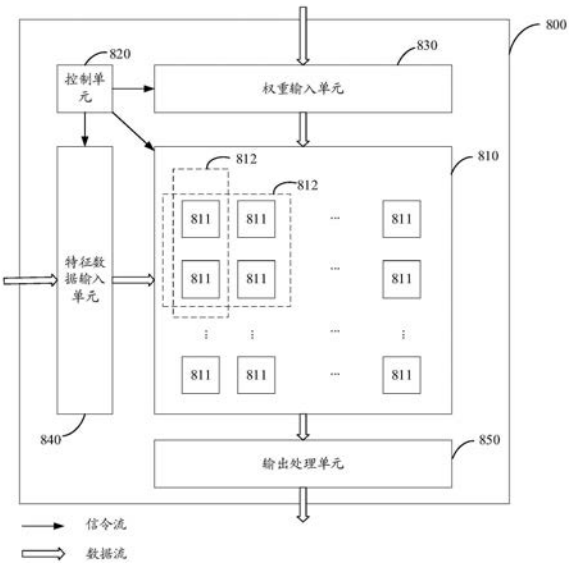
权利要求书3页 说明书19页 附图14页

(54)发明名称

神经网络的运算装置及其控制方法

(57)摘要

提供一种神经网络的运算装置及其控制方法。运算装置包括脉动阵列;脉动阵列的处理单元为第一计算单元,第一计算单元支持乘法操作数的定点数位宽为n比特,n为2的m次方,m为正整数;第一计算单元可进行先移位后累加操作,以使得脉动阵列中2行c列的多个第一计算单元作为一个整体形成支持乘法操作数的定点数位宽为2n比特的第二计算单元,c为1或2。通过设置脉动阵列中的计算单元可以进行先移位后累加操作,可以使得运算装置支持多种定点数位宽,从而可以满足应用中多种定点化精度要求。



CN 112639839 A

1.一种神经网络的运算装置,其特征在于,包括脉动阵列;

所述脉动阵列的处理单元为第一计算单元,所述第一计算单元支持乘法操作数的定点数位宽为n比特,n为2的m次方,m为正整数;

所述第一计算单元可进行先移位后累加操作,以使得所述脉动阵列中2行c列的多个所述第一计算单元作为一个整体形成支持乘法操作数的定点数位宽为2n比特的第二计算单元,c为1或2。

2.如权利要求1所述的运算装置,其特征在于,还包括控制单元,用于:

在所述运算装置需要处理定点数位宽为n比特的输入特征数据的情况下,控制所述第一计算单元不进行先移位后累加操作,以使所述脉动阵列对定点数位宽为n比特的输入特征数据进行处理;

在所述运算装置需要处理定点数位宽为2n比特的输入特征数据的情况下,控制用于形成所述第二计算单元的2行c列的所述第一计算单元中的一个或多个所述第一计算单元进行先移位后累加操作,以使所述脉动阵列对定点数位宽为2n比特的输入特征数据进行处理。

3.如权利要求2所述的运算装置,其特征在于,c为2,所述控制单元用于,在所述运算装置需要对定点数位宽为2n比特的输入特征数据与定点数位宽为2n比特的权重进行运算的情况下,控制所述第二计算单元中所包含的部分所述第一计算单元进行先移位后累加操作,使得所述第二计算单元中所包含的2行所述第一计算单元中后1行的2个所述第一计算单元分别输出所述第二计算单元的4n比特运算结果的低2n比特位与高2n比特位。

4.如权利要求2或3所述的运算装置,其特征在于,所述控制单元还用于,在所述运算装置需要处理定点数位宽为2n比特的输入特征数据的情况下,将定点数位宽为2n比特的输入特征数据的低n比特位与高n比特位分别送入所述第二计算单元中所包含的2行所述第一计算单元中。

5.如权利要求2-4中任一项所述的运算装置,其特征在于,c为2;

所述控制单元还用于,在所述运算装置需要处理定点数位宽为2n比特的权重的情况下,将定点数位宽为2n比特的权重的低n比特位与高n比特位分别送入所述第二计算单元中所包含的2列所述第一计算单元中。

6.如权利要求3所述的运算装置,其特征在于,还包括输出处理单元,用于:

对所述脉动阵列输出的对应于同一个所述第二计算单元的低2n比特位运算结果与高2n比特位运算结果进行拼接,获得所述同一个所述第二计算单元的4n比特运算结果;

对对应于同一个权重矩阵的p个所述第二计算单元的4n比特运算结果进行累加,以获得所述权重矩阵对应的输出特征数据,p等于所述权重矩阵的宽度。

7.如权利要求2-6中任一项所述的运算装置,其特征在于,还包括:

特征数据输入单元,用于缓存待处理的输入特征数据,并根据所述控制单元的控制信令将所述输入特征数据送入所述脉动阵列中;

权重输入单元,用于缓存待处理的权重,并根据所述控制单元的控制信令将所述权重送入所述脉动阵列中。

8.如权利要求3-6中任一项所述的运算装置,其特征在于,所述定点数位宽为2n比特的输入特征数据在外部存储器中的存储格式为:输入特征图中每行输入特征数据的低n比特

位与高n比特位分别集中存储。

9.如权利要求1-8中任一项所述的运算装置,其特征在于,所述运算装置用于执行卷积操作。

10.一种神经网络加速器,其特征在于,包括:

处理模块,所述处理模块为如权利要求1-9中任一项所述神经网络的运算装置;

输入模块,用于将特征数据与权重从外部存储器读出并送入所述处理模块中;

输出模块,用于将所述处理模块输出的输出特征数据存储到所述外部存储器中。

11.一种运算装置的控制方法,其特征在于,所述运算装置包括脉动阵列,所述脉动阵列的处理单元为第一计算单元,所述第一计算单元支持乘法操作数的定点数位宽为n比特,n为2的m次方,m为正整数,所述第一计算单元可进行先移位后累加操作,以使得所述脉动阵列中2行c列的多个所述第一计算单元作为一个整体形成支持乘法操作数的定点数位宽为2n比特的第二计算单元,c为1或2;

所述控制方法包括:

在所述运算装置需要处理定点数位宽为n比特的输入特征数据的情况下,控制所述第一计算单元不进行先移位后累加操作,以使所述脉动阵列对定点数位宽为n比特的输入特征数据进行处理;

在所述运算装置需要处理定点数位宽为2n比特的输入特征数据的情况下,控制用于形成所述第二计算单元的2行c列的所述第一计算单元中的一个或多个所述第一计算单元进行先移位后累加操作,以使所述脉动阵列对定点数位宽为2n比特的输入特征数据进行处理。

12.如权利要求11所述的控制方法,其特征在于,所述在所述运算装置需要处理定点数位宽为2n比特的输入特征数据的情况下,控制用于形成所述第二计算单元的2行c列的所述第一计算单元中的一个或多个所述第一计算单元进行先移位后累加操作,包括:

在所述运算装置需要对定点数位宽为2n比特的输入特征数据与定点数位宽为2n比特的权重进行运算的情况下,控制所述第二计算单元中所包含的部分所述第一计算单元进行先移位后累加操作,使得所述第二计算单元中所包含的2行所述第一计算单元中后1行的2个所述第一计算单元分别输出所述第二计算单元的4n比特运算结果的低2n比特位与高2n比特位。

13.如权利要求11或12所述的控制方法,其特征在于,所述控制方法还包括:

在所述运算装置需要处理定点数位宽为2n比特的输入特征数据的情况下,将定点数位宽为2n比特的输入特征数据的低n比特位与高n比特位分别送入所述第二计算单元中所包含的2行所述第一计算单元中。

14.如权利要求11-13中任一项所述的控制方法,其特征在于,c为2;

所述控制方法还包括:

在所述运算装置需要处理定点数位宽为2n比特的权重的情况下,将定点数位宽为2n比特的权重的低n比特位与高n比特位分别送入所述第二计算单元中所包含的2列所述第一计算单元中。

15.如权利要求12所述的控制方法,其特征在于,所述控制方法还包括:

对所述脉动阵列输出的对应于同一个所述第二计算单元的低2n比特位运算结果与高

2n比特位运算结果进行拼接，获得所述同一个所述第二计算单元的4n比特运算结果；

对对应于同一个权重矩阵的p个所述第二计算单元的4n比特运算结果进行累加，以获得所述权重矩阵对应的输出特征数据，p等于所述权重矩阵的宽度。

16.如权利要求12-15中任一项所述的控制方法，其特征在于，所述定点数位宽为2n比特的输入特征数据在外部存储器中的存储格式为，输入特征图中每行输入特征数据的低n比特位与高n比特位分别集中存储。

17.如权利要求11-16中任一项所述的控制方法，其特征在于，所述运算装置用于执行卷积操作。

18.一种神经网络处理装置，其特征在于，包括：存储器与处理器，所述存储器用于存储指令，所述处理器用于执行所述存储器存储的指令，并且对所述存储器中存储的指令的执行，使得所述处理器用于执行如权利要求11-17中任一项所述的方法。

19.一种计算机存储介质，其特征在于，其上存储有计算机程序，所述计算机程序被计算机执行时，使得所述计算机执行如权利要求11-17中任一项所述的方法。

20.一种包含指令的计算机程序产品，其特征在于，所述指令被计算机执行时使得计算机执行如权利要求11-17中任一项所述的方法。

神经网络的运算装置及其控制方法

[0001] 版权申明

[0002] 本专利文件披露的内容包含受版权保护的材料。该版权为版权所有人所有。版权所有人不反对任何人复制专利与商标局的官方记录和档案中所存在的该专利文件或者该专利披露。

技术领域

[0003] 本申请涉及神经网络领域，并且更为具体地，涉及一种神经网络的运算装置及其控制方法。

背景技术

[0004] 计算机中数值的表示有两种形式，一种是定点数(fixed-point number)，另一种是浮点数(floating-point number)。当前主流的神经网络计算框架中，普遍采用浮点数作为计算单元运算时要求的数据格式，例如，神经网络计算框架训练后得到的权重系数和各层的输出特征数据都是浮点数。由于定点运算装置相比于浮点运算装置占用的面积更小，消耗的功耗更少，所以神经网络加速装置普遍采用定点数作为计算单元运算时要求的数据格式。因此，神经网络计算框架训练得到的权重系数和各层的输出特征数据在神经网络加速装置中部署时，均需要进行定点化。定点化指的是将数据由浮点数转换为定点数的过程。

[0005] 有些深度卷积神经网络为满足运算精度要求需要使用较小位宽的定点数进行定点化，而另一些深度卷积神经网络为满足运算精度要求需要使用较大位宽的定点数进行定点化。

[0006] 但是，当前的神经网络运算装置只支持一种定点数位宽，导致无法满足应用中进行定点化的运算精度要求。

发明内容

[0007] 本申请提供一种神经网络的运算装置及其控制方法，该运算装置可以支持多种定点数位宽，从而可以满足应用中进行定点化的运算精度要求。

[0008] 第一方面，本申请实施例提供一种神经网络的运算装置，所述运算装置包括脉动阵列；所述脉动阵列的处理单元为第一计算单元，所述第一计算单元支持乘法操作数的定点数位宽为n比特，n为2的m次方，m为正整数；所述第一计算单元可进行先移位后累加操作，以使得所述脉动阵列中2行c列的多个所述第一计算单元作为一个整体形成支持乘法操作数的定点数位宽为2n比特的第二计算单元，c为1或2。

[0009] 第二方面，本申请实施例提供一种神经网络加速器，包括：处理模块，所述处理模块为第一方面提供的神经网络的运算装置；输入模块，用于从外部存储器读取输入特征数据与权重到所述处理模块中；输出模块，用于将所述处理模块获得的输出特征数据存储到所述外部存储器中。

[0010] 第三方面，本申请实施例提供一种神经网络的运算装置的控制方法，所述运算装

置包括脉动阵列，所述脉动阵列的处理单元为第一计算单元，所述第一计算单元支持乘法操作数的定点数位宽为n比特，n为2的m次方，m为正整数，所述第一计算单元可进行先移位后累加操作，以使得所述脉动阵列中2行c列的多个所述第一计算单元作为一个整体形成支持乘法操作数的定点数位宽为2n比特的第二计算单元，c为1或2；所述控制方法包括：在所述运算装置需要处理定点数位宽为n比特的输入特征数据的情况下，控制所述第一计算单元不进行先移位后累加操作，以使所述脉动阵列对定点数位宽为n比特的输入特征数据进行处理；在所述运算装置需要处理定点数位宽为2n比特的输入特征数据的情况下，控制用于形成所述第二计算单元的2行c列的所述第一计算单元中的一个或多个所述第一计算单元进行先移位后累加操作，以使所述脉动阵列对定点数位宽为2n比特的输入特征数据进行处理。

[0011] 第四方面，本申请实施例提供一种装置，所述装置用于执行上述第三方面中的方法。

[0012] 第五方面，本申请实施例提供一种装置，所述装置包括存储器和处理器，所述存储器用于存储指令，所述处理器用于执行所述存储器存储的指令，并且对所述存储器中存储的指令的执行使得所述处理器执行第三方面的方法。

[0013] 第六方面，本申请实施例提供一种芯片，所述芯片包括处理模块与通信接口，所述处理模块用于控制所述通信接口与外部进行通信，所述处理模块还用于实现第三方面的方法。

[0014] 第七方面，本申请提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被计算机执行时使得所述计算机实现第一方面的方法。

[0015] 第八方面，本申请提供一种包含指令的计算机程序产品，所述指令被计算机执行时使得所述计算机实现第三方面的方法。具体地，所述计算机可以为上述装置。

[0016] 第九方面，本申请实施例提供一种神经网络处理芯片，其上集成由第一方面提供的神经网络运算装置或由第二方面提供的神经网络加速器。

[0017] 在本申请提供的神经网络的运算装置中，通过设置脉动阵列中的计算单元可以进行先移位后累加操作，可以使得运算装置支持多种定点数位宽，从而可以满足应用中多种定点化精度要求。

附图说明

[0018] 图1为深度卷积神经网络的框架示意图。

[0019] 图2为卷积操作的示意图。

[0020] 图3为神经网络加速装置的架构示意图。

[0021] 图4至图7为采用本申请实施例提供的神经网络处理装置实现卷积操作或平均池化操作的时序示意图。

[0022] 图8为本申请实施例提供的神经网络运算装置的示意性框图。

[0023] 图9为本申请实施例提供的神经网络的运算装置的另一示意性框图。

[0024] 图10为本申请实施例提供的神经网络的运算装置的控制方法的示意图。

[0025] 图11为本申请实施例提供的运算装置中的脉动阵列中的2行2列的第一计算单元等效形成支持2n比特的定点数位宽的第二计算单元的示意图。

[0026] 图12为本申请实施例提供的运算装置中的脉动阵列中的第一计算单元的结构示意图。

[0027] 图13为本申请实施例提供的运算装置中的ACC阵列中的ACC的结构示意图。

[0028] 图14为使用本申请实施例提供的神经网络的运算装置进行定点数位宽为n比特的卷积运算的示意性流程图。

[0029] 图15为使用本申请实施例提供的神经网络的运算装置进行定点数位宽为2n比特的卷积运算的示意性流程图。

[0030] 图16为定点数位宽为2n bits的特征数据在SRAM中的存储格式的示意图。

[0031] 图17为定点数位宽为n bits的特征数据在SRAM中存储的格式的示意图。

[0032] 图18为本申请实施例提供的神经网络加速器的示意性框图。

[0033] 图19为本申请实施例提供的神经网络处理装置的示意性框图。

具体实施方式

[0034] 为了更好地理解本申请实施例,下面先介绍本申请涉及的相关技术与概念。

[0035] 1、深度神经网络(以深度卷积神经网络(Deep Convolutional Neural Network,DCNN)为例)

[0036] 图1是深度卷积神经网络的框架示意图。深度卷积神经网络的输入值(由输入层输入),经隐藏层进行卷积(convolution)、转置卷积(transposed convolution or deconvolution)、归一化(Batch Normalization,BN)、缩放(Scale)、全连接(fully connected)、拼接(Concatenation)、池化(pooling)、元素智能加法(element-wise addition)和激活(activation)等运算后,得到输出值(由输出层输出)。本申请实施例的神经网络的隐藏层可能涉及的运算不仅限于上述运算。

[0037] 深度卷积神经网络的隐藏层可以包括级联的多层。每层的输入为上层的输出,为特征图(feature map),每层对输入的一组或多组特征图进行前述描述的至少一种运算,得到该层的输出。每层的输出也是特征图。一般情况下,各层以实现的功能命名,例如实现卷积运算的层称作卷积层,实现池化运算的层称作池化层。此外,深度卷积神经网络的隐藏层还可以包括转置卷积层、BN层、Scale层、池化层、全连接层、Concatenation层、元素智能加法层和激活层等,此处不进行一一列举。通常情况下,卷积层的后面会紧接着一层激活层。在BN层被提出以后,越来越多的神经网络在卷积层之后先接BN层,再接激活层。

[0038] 作为示例而非限定,卷积层的卷积操作过程如图2所示。卷积层的卷积操作过程为,对一组权重值与一组输入特征数据进行向量内积运算,输出一组输出特征数据。一组权重值可以称为滤波器(filter)或卷积核。一组输入特征数据为输入特征图中的部分特征值。一组输出特征数据为输出特征图中的部分特征值。卷积层的每个输出特征数据由输入特征图中的部分特征值与卷积核中的权重值进行内积运算得到。

[0039] 卷积核、输入特征图与输出特征图均可以被表示为一个多维矩阵。例如,在图2中,卷积核可以被表示为三维矩阵R×R×N,该卷积核的宽度与高度均为R,深度为N;输入特征图可以表示为三维矩阵H×H×M,输入特征图的宽度与高度均为H,深度为M(图中未示出);输出特征图可以表示为三维矩阵E×E×L,输出特征图的宽度与高度均为E,深度为L。

[0040] 深度卷积神经网络中其它各层的运算流程可以参考现有的技术,本文不进行赘

述。

[0041] 深度卷积神经网络的每层(包括输入层和输出层)可以有一个输入和/或一个输出,也可以有多个输入和/或多个输出。例如,在视觉领域的分类和检测任务中,特征图的宽高往往是逐层递减的(如图1所示的输入、特征图#1、特征图#2、特征图#3和输出的宽高是逐层递减的)。又例如,在语义分割任务中,特征图的宽高在递减到一定深度后,有可能会通过转置卷积运算或上采样(upsampling)运算,再逐层递增。当前,需要较多权重参数用于运算的层有:卷积层、全连接层、转置卷积层和BN层。

[0042] 2、神经网络加速装置

[0043] 从图2可知,卷积层的计算耗时较长。深度卷积神经网络中大部分运算都是卷积操作,卷积计算时间占了深度卷积神经网络的大部分计算时间,导致深度卷积神经网络的计算时间很长。为了减少深度卷积神经网络的计算时间,神经网络加速装置被提了出来。神经网络加速装置表示专用于处理神经网络运算的硬件电路。例如,专用于加速卷积层的运算的加速装置可以称为深度卷积神经网络加速装置。

[0044] 作为示例而非限定,图3为神经网络加速装置的架构示意图。神经网络加速装置300包括输入特征数据输入模块(IFM_Loader)310、权重输入模块(或称为滤波器输入模块(Filt_Loader))320、计算模块(或称为乘累加处理模块(MAU))330与输出模块(OFM_Packer)340。

[0045] 输入特征数据输入模块310用于从外部存储器中(图3中以静态随机存取存储器(Static Random-Access Memory,SRAM)为例)读出输入特征数据,并将其送入处理模块330中。

[0046] 权重输入模块320用于从SRAM中读出权重值,并将其送入处理模块330中。

[0047] 计算模块330用于对输入特征数据与权重值进行乘累加操作,获得输出特征数据并输出。

[0048] 输出模块340用于将处理模块330输出的输出特征数据写入SRAM。

[0049] 如图3所示,计算模块330包括脉动阵列331与输出处理单元332。输出处理单元332中包括用来存储神经网络运算的中间结果的存储器。

[0050] 下面以进行卷积运算为例,描述计算模块330采用脉动阵列331进行运算的流程。

[0051] 1)将权重输入模块320送入的权重值装载到脉动阵列331。

[0052] 2)将输入特征数据输入模块310送入的输入特征图数据送入脉动阵列331,将其与在先装载的权重值进行乘累加操作。

[0053] 3)如果存储器缓存了中间结果,输出处理单元332则将脉动阵列331的输出结果与存储器中的中间结果再进行一次累加。若累加的结果仍为中间结果,则输出处理单元332将其继续存储到存储器中,否则将其输出到输出模块340中进行后续处理。

[0054] 3、定点化

[0055] 计算机中数值的表示有两种形式,一种是定点数(fixed-point number),另一种是浮点数(floating-point number)。当前主流的神经网络计算框架中,普遍采用浮点数作为计算单元运算时要求的数据格式,例如,神经网络计算框架训练后得到的权重系数和各层的输出特征数据都是浮点数。由于定点运算装置相比于浮点运算装置占用的面积更小,消耗的功耗更少,所以神经网络加速装置普遍采用定点数作为计算单元运算时要求的数据

格式。因此，神经网络计算框架训练得到的权重系数和各层的输出特征数据在神经网络加速装置中部署时，均需要进行定点化。定点化指的是将数据由浮点数转换为定点数的过程。关于浮点数与定点数以及定点化的概念可以参考现有技术，本文不进行详述。

[0056] 现有部分深度卷积神经网络使用较小位宽的定点数进行定点化后的精度损失较小；但是另一部分深度卷积神经网络采用相同位宽的定点数进行定点化后的精度损失却很大。也就是说，为满足运算精度要求需要，有些深度卷积神经网络使用较小位宽的定点数进行定点化，而另一些深度卷积神经网络需要使用较大位宽的定点数进行定点化。

[0057] 但是，当前神经网络运算装置只支持一种定点数位宽，例如，只支持8比特(bit)的定点数位宽，或者只支持16比特的定点数位宽，导致无法满足应用中进行定点化的运算精度要求。

[0058] 针对上述问题，本申请提出一种可支持多种定点数位宽的神经网络运算装置。

[0059] 为了更好地理解本申请实施例，下面先以图3中所示的脉动阵列331为例，结合图4-图7描述脉动阵列的原理。

[0060] 假设图3中所示的脉动阵列331包括如图4所示的3行3列的计算单元：C00、C01、C02、C10、C11、C12、C20、C21与C22。输出处理单元332与计算单元C20、C21与C22连接，用于根据其输出的运算结果获得输出特征数据。

[0061] 下面以3*3的权重矩阵为 $\begin{bmatrix} W11 & W12 & W13 \\ W21 & W22 & W23 \\ W31 & W32 & W33 \end{bmatrix}$ 与3*3的输入特征矩阵为 $\begin{bmatrix} a11 & a12 & a13 \\ a21 & a22 & a23 \\ a31 & a32 & a33 \end{bmatrix}$ 的卷积操作为例进行描述。即3*3的权重矩阵 $\begin{bmatrix} W11 & W12 & W13 \\ W21 & W22 & W23 \\ W31 & W32 & W33 \end{bmatrix}$ 可以称为卷积核。

[0062] $\begin{bmatrix} a11 & a12 & a13 \\ a21 & a22 & a23 \\ a31 & a32 & a33 \end{bmatrix}$ 与 $\begin{bmatrix} W11 & W12 & W13 \\ W21 & W22 & W23 \\ W31 & W32 & W33 \end{bmatrix}$ 的卷积操作的计算结果应为：

[0063] a11*W11+a12*W12+a13*W13+a21*W21+a22*W22+a23*W23+a31*W31+a32*W32+a33*W33。

[0064] 如图4至图7所示，采用脉动阵列331执行卷积操作的流程如下。

[0065] 参见图4所示，预先将权重 $\begin{bmatrix} W11 & W12 & W13 \\ W21 & W22 & W23 \\ W31 & W32 & W33 \end{bmatrix}$ 加载到计算单元C00、C01、C02、C10、C11、C12、C20、C21与C22中。

[0066] 在第一个周期(T1)，参见图5所示，输入特征数据a11进入计算单元C00，其中，输入特征数据a11从计算单元C00的左边载入并从左向右流动。在第一个周期结束时，计算单元C00的计算结果为a11*W11。计算单元C00的计算结果a11*W11从上向下流动。

[0067] 在第二个周期(T2)，参见图6所示，输入特征数据a11向右流动进入计算单元C01，

计算结果a11*W11向下流动进入计算单元C10;同时输入特征数据a21载入至计算单元C00,输入特征数据a21载入至计算单元C10。在第二个周期结束时,计算单元C00的计算结果为a12*W11,计算单元C01的计算结果为a11*W12,计算单元C10的计算结果为a11*W11+a21*W21。各个计算单元的计算结果从上向下流动。

[0068] 在第三个周期(T3),参见图7所示,输入特征数据a11向右流动进入计算单元C02,a12向右流动进入计算单元C01,a21向右流动进入计算单元C11,计算结果a12*W11向下流动进入计算单元C10,计算结果a12*W12向下流动进入计算单元C11,计算结果a11*W11+a21*W21向下流动进入计算单元C20。同时,a13载入计算单元C00,a22载入计算单元C10,a31载入计算单元C20。在第三个周期结束时,计算单元C00的计算结果为a13*W11,计算单元C01的计算结果为a12*W12,计算单元C02的计算结果为a11*W13,计算单元C10的计算结果为a12*W11+a22*W21,计算单元C11的计算结果为a11*W12+a21*W22,计算单元C20的计算结果为a11*W11+a21*W21+a31*W31。各个计算单元的计算结果从上向下流动。

[0069] 以此类推,在第五个周期结束时计算单元C21会输出计算结果a12*W12+a22*W22+a32*W32,在第七个周期结束时计算单元C22会输出计算结果a13*W13+a23*W23+a33*W33。

[0070] 可知,第三个周期结束时计算单元C20的计算结果a11*W11+a21*W21+a31*W31、第五个周期结束时计算单元C21的计算结果a12*W12+a22*W22+a32*W32、以及第七个周期结束时计算单元C22的计算结果a13*W13+a23*W23+a33*W33的累加值为输入特征数据$\begin{bmatrix} a11 & a12 & a13 \\ a21 & a22 & a23 \\ a31 & a32 & a33 \end{bmatrix}$和权重$\begin{bmatrix} W11 & W12 & W13 \\ W21 & W22 & W23 \\ W31 & W32 & W33 \end{bmatrix}$的卷积操作的计算结果。

[0071] 输出处理单元332用于接收计算单元C20、C21与C22输出的运算结果(应理解,是卷积操作的中间计算结果),并对第三个周期结束时计算单元C20的计算结果、第五个周期结束时计算单元C21的计算结果、以及第七个周期结束时计算单元C22的计算结果作累加,得到输入特征数据$\begin{bmatrix} a11 & a12 & a13 \\ a21 & a22 & a23 \\ a31 & a32 & a33 \end{bmatrix}$和权重$\begin{bmatrix} W11 & W12 & W13 \\ W21 & W22 & W23 \\ W31 & W32 & W33 \end{bmatrix}$的卷积操作的计算结果。

[0072] 图8为本申请实施例提供的神经网络运算装置800的示意性框图。神经网络运算装置800包括脉动阵列810。

[0073] 脉动阵列810的处理单元为第一计算单元811。下文一些实施例中也会将第一计算单元811记为MC。

[0074] 第一计算单元811支持乘法操作数的定点数位宽为n比特,n为2的m次方,m为正整数。

[0075] m例如为1、2、3或其它正整数。即第一计算单元811可以支持定乘法操作数的定点数位宽为2比特、4比特、8比特或16比特或其它2的m次方比特的运算。

[0076] 第一计算单元811支持乘法操作数的定点数位宽为n比特表示,第一计算单元811支持乘法运算x1*y1,其中,x1与y1的定点数位宽均为n比特。

[0077] 第一计算单元811可进行先移位后累加操作,以使得脉动阵列810中2行c列的多个第一计算单元811作为一个整体形成支持乘法操作数的定点数位宽为2n比特的第二计算单

元812，c为1或2。

[0078] 下文一些实施例中也会将第二计算单元811记为MU。

[0079] 第一计算单元811可进行先移位后累加操作，表示，第一计算单元811可对本计算单元的运算结果与其它计算单元的运算结果进行先移位后累加的操作，其中，该其它计算单元可以包括与本计算单元行相邻的计算单元，或与本计算单元列相邻的计算单元，或位于本计算单元的对角线上的计算单元。

[0080] 例如，第一计算单元811可以将本计算单元的运算结果先左移n位，再与与本计算单元行相邻的前一级计算单元的运算结果累加。

[0081] 脉动阵列810中2行c列的多个第一计算单元811作为一个整体形成支持乘法操作数的定点数位宽为2n比特的第二计算单元812，表示，脉动阵列810中2行c列的多个第一计算单元811作为一个整体可以等效为第二计算单元812，且该第二计算单元812支持乘法操作数的定点数位宽为2n比特。

[0082] 第二计算单元812支持乘法操作数的定点数位宽为2n比特，表示，第二计算单元812支持乘法运算x2*y2，x2与y2中定点数位宽最大的一个数的定点数位宽为2n比特，例如，x2与y2的定点数位宽均为2n比特，或者，x2与y2中一个数的定点数位宽为2n比特，另一个数的定点数位宽为n比特。

[0083] 应理解，因为第一计算单元811支持乘法操作数的定点数位宽为n比特，因此，脉动阵列810本身是可以支持定点数位宽为2n比特的运算的。若脉动阵列810中2行c列的多个第一计算单元811作为一个整体形成的第二计算单元812支持乘法操作数的定点数位宽为2n比特，则使得脉动阵列810也可支持定点数位宽为2n比特的运算。

[0084] 因此，脉动阵列810既可以支持定点数位宽为n比特的运算，又可以支持定点数位宽为2n比特的运算，也就是说，本申请实施例提供的神经网络的运算装置既可以支持定点数位宽为n比特的运算，又可以支持定点数位宽为2n比特的运算。

[0085] 上述可知，在本申请提供的神经网络的运算装置中，通过设置脉动阵列中的计算单元可以进行先移位后累加操作，可以使得运算装置支持多种定点数位宽，从而可以满足应用中多种定点化精度要求。

[0086] 需要说明的是，第二计算单元812仅为了便于理解与描述而引入，并非表示脉动阵列810中实际包含第二计算单元812。换句话说，在本文中，将脉动阵列810中2行c列的多个第一计算单元811作为一个整体所形成的计算单元记为第二计算单元812。

[0087] 作为示例而非限定，图9为本申请实施例提供的神经网络的运算装置800的另一示意性框图。

[0088] 在图9中，第一计算单元811表示为MC，第二计算单元812表示为MU。在图9中，c为2，即脉动阵列810中2行2列的第一计算单元(即4个第一计算单元)整体可等效为一个第二计算单元812。

[0089] 例如，在图9中，MC不进行先移位后累加操作，则脉动阵列810可进行乘法操作数的定点数位宽为n比特的运算。

[0090] 又例如，在图9中，同一个MU内的MC通过进行先移位后累加操作，使得MU可以支持乘法操作数的定点数位宽为2n比特，则脉动阵列810可进行乘法操作数的定点数位宽为2n比特的运算。

[0091] 可知，在本申请实施例中，无需额外增加硬件，就可以实现神经网络运算装置支持多种定点数位宽。

[0092] 例如，本申请实施例提供的神经网络的运算装置800可以应用于卷积层，也可应用于池化层。即运算装置800可以用于处理卷积操作，也可以用于处理池化操作。

[0093] 运算装置800可以在控制单元的控制下，在多种不同定点数位宽之间切换。

[0094] 如图8与图9所示，运算装置800还包括控制单元820。控制单元820用于向脉动阵列810发送控制信令，以控制脉动阵列810的运算方式。可以理解到，控制单元820可以向第一计算单元811发送控制信令，以控制第一计算单元811的运算方式。

[0095] 例如，控制单元820用于执行如图10所示的操作S1010与S1020。

[0096] S1010，在运算装置800需要处理定点数位宽为n比特的输入特征数据的情况下，控制第一计算单元811不进行先移位后累加操作，以使脉动阵列810对定点数位宽为n比特的输入特征数据进行处理。

[0097] S1020，在运算装置800需要处理定点数位宽为2n比特的输入特征数据的情况下，控制用于形成第二计算单元812的2行c列的第一计算单元811中的一个或多个第一计算单元811进行先移位后累加操作，使得第二计算单元812支持乘法操作数的定点数位宽为2n比特，从而使得脉动阵列810对定点数位宽为2n比特的输入特征数据进行处理。

[0098] 例如，c为1。控制单元820用于在，在运算装置800需要处理定点数位宽为2n比特的输入特征数据与定点数位宽为n比特的权重的情况下，控制用于形成第二计算单元812的2行1列的第一计算单元811中的第2行的第一计算单元811进行先移位后累加操作，使得第二计算单元812支持定点数位宽为2n比特的输入特征数据与定点数位宽为n比特的权重的运算，从而使得运算装置800可以执行定点数位宽为2n比特的输入特征数据与定点数位宽为n比特的权重的运算。

[0099] 例如，c为2。控制单元820用于在，在运算装置800需要处理定点数位宽为2n比特的输入特征数据与定点数位宽为2n比特的权重的情况下，控制用于形成第二计算单元812的2行2列的第一计算单元811中的部分第一计算单元811进行先移位后累加操作，使得第二计算单元812支持定点数位宽为2n比特的输入特征数据与定点数位宽为2n比特的权重的运算，从而使得运算装置800可以执行定点数位宽为2n比特的输入特征数据与定点数位宽为2n比特的权重的运算。

[0100] 其中，在本例中，控制单元820用于，控制第二计算单元812中所包含的2行2列中的部分第一计算单元811进行先移位后累加操作，使得第二计算单元812中所包含的2行第一计算单元811中后1行的2个第一计算单元811分别输出第二计算单元812的4n比特运算结果的低2n比特位与高2n比特位。参见下文描述的(三)2n bits*2n bits的运算的示例。

[0101] 在运算装置800需要处理定点数位宽为n比特的输入特征数据的情况下，控制单元820还用于，将定点数位宽为n比特的输入特征数据送入第一计算单元811。

[0102] 作为示例，定点数位宽为n比特的输入特征数据被送入脉动阵列810的方式如下文描述的(一)n bits*n bits的运算。

[0103] 在运算装置800需要处理定点数位宽为2n比特的输入特征数据的情况下，控制单元820还用于，将定点数位宽为2n比特的输入特征数据的低n比特位与高n比特位分别送入第二计算单元812中所包含的2行第一计算单元811中。

[0104] 作为示例,定点数位宽为2n比特的输入特征数据被送入脉动阵列810的方式如下文描述的(二)2n bits*n bits的运算与(三)2n bits*2n bits的运算。

[0105] 在一些实施例中,c为2,在运算装置800需要处理定点数位宽为2n比特的权重的情况下,控制单元820还用于,将定点数位宽为2n比特的权重的低n比特位与高n比特位分别送入第二计算单元812中所包含的2列第一计算单元811中。

[0106] 例如,在运算装置800需要进行定点数位宽为2n比特的输入特征数据与权重的运算的情况下,控制单元820用于,将定点数位宽为2n比特的输入特征数据的低n比特位与高n比特位分别送入第二计算单元812中所包含的2行第一计算单元811中;将定点数位宽为2n比特的权重的低n比特位与高n比特位分别送入第二计算单元812中所包含的2列第一计算单元811中。

[0107] 作为示例,定点数位宽为2n比特的输入特征数据与权重被送入脉动阵列810的方式如下文描述的(三)2n bits*2n bits的运算。

[0108] 继续参见图8与图9,运算装置800还包括特征数据输入单元840与权重输入单元830。

[0109] 权重输入单元830,用于缓存待处理的权重,并根据控制单元820的控制信令将权重送入脉动阵列810中。

[0110] 例如,权重输入单元830负责缓存权重值(例如由下文描述的加速器1800中的权重输入模块(Filt_Loader模块)送入),并在控制单元820的控制下为脉动阵列810装载权重。权重输入单元830与脉动阵列810的每一列第一计算单元810有且仅有一个接口,该接口每个时钟周期仅能传输一个权重值。权重装载分为移位和装载两个阶段,在移位阶段,权重输入单元830依次将同一列第一计算单元810需要的权重值通过同一个接口依次送入脉动阵列810。在脉动阵列810中,接收到的权重值从接口处的第一计算单元810依次向下传递。在装载阶段,脉动阵列810中同一列的第一计算单元810同时将缓存的权重值装载到各自的缓存中供后续的乘累加单元使用。权重输入单元830为相邻两列第一计算单元810装载数据时会有一个时钟周期的延迟。

[0111] 特征数据输入单元840,用于缓存待处理的输入特征数据,并根据控制单元820的控制信令将输入特征数据送入脉动阵列810中。

[0112] 例如,特征数据输入单元840负责缓存输入特征数据(例如由下文描述的加速器1800中的特征数据输入模块(IFM_Loader模块)送入),并在控制单元820的控制下为脉动阵列810装载输入特征数据。特征数据输入单元840与脉动阵列810的每一行第一计算单元811仅有一个接口,该接口每个时钟周期仅能传输一个输入特征数据。在脉动阵列810中,接收到的输入特征数据从接口处的第一计算单元811依次向右传递直至最后一个第一计算单元811。特征数据输入单元840为相邻两行第一计算单元811装载数据时会有一个时钟周期的延迟。

[0113] 在本申请实施例中,通过设置脉动阵列中的计算单元可以进行先移位后累加操作,可以使得运算装置支持多种定点数位宽,从而可以根据应用需求控制运算装置在多种定点数位宽之间切换,以满足应用中多种定点化精度要求。

[0114] 以运算装置800执行卷积操作为例,控制单元820负责控制运算装置800中的各个单元以实现卷积运算。首先,控制单元820控制权重输入单元830将权重值装载到脉动阵列

810,然后,控制单元820控制特征数据输入单元840将特征图数据送入脉动阵列810,并控制脉动阵列810进行卷积运算。待所有的特征图数据送入脉动阵列810并完成卷积运算后,再依次重复上述过程,直至完成所有的卷积运算。

[0115] 下面结合图11描述,由脉动阵列810中的2行c列的第一计算单元811等效形成支持2n比特的定点数位宽的第二计算单元的方式。

[0116] 在图11中,将第一计算单元811记为MC,即MC可以完成n bits*n bits的乘累加操作;将第二计算单元812记为MU。

[0117] 在图11中,以c为2为例,即脉动阵列810中的2行2列的第一计算单元811(MC)等效形成一个第二计算单元812(MU)。在图11中,将2行2列的4个第一计算单元811分别标记为MC(U0_0)、MC(U0_1)、MC(U1_0)与MC(U1_1)。即MC(U0_0)、MC(U0_1)、MC(U1_0)与MC(U1_1)可以等效为一个第二计算单元812(MU)。

[0118] 在图11的示例中,MC具有如下几个输入端:bi(n bits)、si、ai(n bits)、ci。MC具有如下几个输出端:bo(n bits)、ar、acr、ao(n bits)、mr。

[0119] MC的各个输入端的含义如下。

[0120] 输入端bi(n bits)被配置为输入n bits的输入权重值。例如,MC(U0_0)的输入端bi(n bits)输入n bits的输入权重值b_lsb。

[0121] 输入端si被配置为输入前级MC输出的累加结果。例如,MC(U0_0)的输入端si输入前级MC输出的累加结果s_lsb。

[0122] 输入端ai(n bits)被配置为输入n bits的输入特征值。例如,MC(U0_0)的输入端ai(n bits)输入n bits的输入特征数据a_lsb。

[0123] 输入端ci被配置为输入相邻的、但非前级的MC的中间结果。如下面描述的(三)2n bits*2n bits的运算,MC(U1_0)的输入端ci输入MC(U0_1)的中间结果“RM(U0_1)[7:0]”,MC(U0_1)的输入端ci输入MC(U1_0)的中间结果“RM(U1_0)[31:8]”,MC(U1_1)的输入端ci输入MC(U1_0)的中间结果“RA(U1_0)[31:16]”。

[0124] MC的各个输出端的含义如下。

[0125] 输出端bo(n bits)被配置为向下级MC输出n bits的输入特征数据。例如,MC(U0_0)的输出端bo(n bits)向MC(U1_0)输出n bits的输入特征数据b_lsb。

[0126] 输出端ar被配置为输出当前MC的计算结果。例如,MC(U0_0)的输出端ar输出MC(U0_0)的计算结果RA(U0_0)。

[0127] 输出端acr被配置为向权重值流动方向上的下一个MC输出当前MU的中间结果。如下面描述的(三)2n bits*2n bits的运算,MC(U1_0)的输出端acr输出MC(U1_0)的中间结果RM(U1_0)[31:8]。

[0128] 输出端ao(n bits)被配置为向权重值流动方向上的下一个MC输出n bits的输入权重值。例如,MC(U0_0)的输出端ao(n bits))向,MC(U0_1)输出n bits的输入权重值。

[0129] 输出端mr被配置为向属于同一个MU中的、位于对角线位置的MC输出当前MU的中间结果。如下面描述的(三)2n bits*2n bits的运算,MC(U1_0)的输出端mr向MC(U1_1)输出MC(U1_0)的中间结果RA(U1_0)[31:16]。

[0130] 图11中所示的MU可同时完成4个n bits*n bits的乘累加操作,或同时完成2个2n bits*n bits的乘累加操作,或1个2n bits*2n bits的乘累加操作。具体描述如下。

[0131] (一)n bits*n bits的运算

[0132] 当MU进行n bits*n bits的运算时,输入端口a_lsb、a_msb、b_lsb和b_msb分别输入四个不同的n bits的操作数。MU中的4个MC单元分别完成四个不同的n bits*n bits的乘累加操作。计算过程(一)如下所示。

[0133] MC(U0_0):RM(U0_0)=a_lsb*b_lsb

[0134] RA(U0_0)=RM(U0_0)+si_lsb

[0135] SO_LSB=RA(U0_0)

[0136] MC(U1_0):RM(U1_0)=a_msb*b_lsb

[0137] RA(U1_0)=RM(U1_0)+RA(U0_0)

[0138] SO_LSB=RA(U1_0)

[0139] MC(U0_1):RM(U0_1)=a_lsb*b_msb

[0140] RA(U0_1)=RM(U0_1)+si_msb

[0141] SO_MSB=RA(U0_1)

[0142] MC(U1_1):RM(U1_1)=a_msb*b_msb

[0143] RA(U1_1)=RM(U1_1)+RA(U0_1)

[0144] SO_MSB=RA(U1_1)

[0145] 其中,a_msb与a_lsb表示两个不同的n bits的输入特征数据;b_msb与b_lsb表示两个不同的n bits的输入权重值。s_msb与s_lsb表示前级MU输出的累加结果。so_lsb与so_msb表示当前MU输出的累加结果。

[0146] (二)2n bits*n bits的运算

[0147] 当MU进行2n bits*n bits的运算时,2n bits的输入特征值a分别通过输入端口a_msb和a_lsb送入MU;其中,输入端口a_lsb为输入特征值a的低n bits,输入端口a_msb为输入特征值a的高n bits。输入端口b_lsb和b_msb分别为两个不同的n bits的输入权重值。四个MC单元分为两组,MC(U0_0)和MC(U1_0)为第一组,MC(U0_1)和MC(U1_1)为第二组。两组MC单元分别完成二个不同的2n bits*n bits的乘累加操作。计算过程(二)如下所示。MC(U0_0):RM(U0_0)=a_lsb*b_lsb

[0148] RA(U0_0)=RM(U0_0)+s_lsb

[0149] MC(U1_0):RM(U1_0)=a_msb*b_lsb

[0150] RA(U1_0)=RM(U1_0)<<8+RA(U0_0)

[0151] SO_LSB=RA(U1_0)

[0152] MC(U0_1):RM(U0_1)=a_lsb*b_msb

[0153] RA(U0_1)=RM(U0_1)+s_msb

[0154] MC(U1_1):RM(U1_1)=a_msb*b_msb

[0155] RA(U1_1)=RM(U1_1)<<8+RA(U0_1)

[0156] SO_MSB=RA(U1_1)

[0157] 其中,{a_msb,a_lsb}为一个2n bits的输入特征值;b_msb与b_lsb为的两个不同的n bits的输入权重值。s_msb与s_lsb为前级MU输出的累加结果。so_lsb和so_msb为当前MU输出的两个累加结果。

[0158] 可选地,在本例中,作为第一组的MC(U0_0)和MC(U1_0)可以整体视为一个第二计

算单元812,作为第二组的MC(U0_1)和MC(U1_1)可以整体视为一个第二计算单元812。

[0159] 可选地,在本例中,MC(U0_0)、MC(U1_0)、MC(U0_1)和MC(U1_1)可以整体视为一个第二计算单元812。

[0160] (三)2n bits*2n bits的运算

[0161] 当MU进行2n bits*2nbits的运算时,2n bits输入特征值a分别通过输入端口a_msb和a_lsb送入MU;其中,输入端口a_lsb为输入特征值a的低n bits,输入端口a_msb为输入特征值a的高n bits。2n bits输入权重值b分别通过输入端口b_msb和b_lsb送入MU;其中,输入端口b_lsb为输入权重值b的低n bits,输入端口b_msb为输入权重值b的高n bits。MC(U0_0)和MC(U1_0)输出2n bits*2n bits乘累加结果的低2n bits,MC(U0_1)和MC(U1_1)输出2n bits*2nbits乘累加结果的高2nbits。计算过程(三)如下所示。

[0162] MC(U0_0):RM(U0_0)=a_lsb*b_lsb

[0163] RA(U0_0)=RM(U0_0)+s_lsb

[0164] MC(U1_0):RM(U1_0)=a_msb*b_lsb

[0165] RA(U1_0)=RA(U0_0)+RM(U1_0)[7:0]<<8+RM(U0_1)[7:0]<<8

[0166] SO_LSB=RA(U1_0)[15:0]

[0167] MC(U0_1):RM(U0_1)=a_lsb*b_msb

[0168] RA(U0_1)=RM(U0_1)[31:8]+RM(U1_0)[31:8]+s_msb

[0169] MC(U1_1):RM(U1_1)=a_msb*b_msb

[0170] RA(U1_1)=RM(U1_1)+RA(U0_1)+RA(U1_0)[31:16]

[0171] SO_MSB=RA(U1_1)

[0172] 其中,{a_msb,a_lsb}表示一个2n bits的输入特征值;{b_msb,b_lsb}表示一个2n bits的输入权重值;{s_msb,s_lsb}表示前级MU输出的累加结果;{so_msb,so_lsb}表示当前MU输出的累加结果。

[0173] 可选地,在图11所示实施例中,控制单元820控制第一计算单元811执行上述计算过程(一),可以使得脉动阵列810执行n bits*n bits的运算,即可以使得运算装置800支持n bits*n bits的运算。

[0174] 可选地,在图11所示实施例中,控制单元820控制第一计算单元811执行上述计算过程(二),可以使得脉动阵列810执行2n bits*n bits的运算,即可以使得运算装置800支持2n bits*n bits的运算。

[0175] 可选地,在图11所示实施例中,控制单元820控制第一计算单元811执行上述计算过程(三),可以使得脉动阵列810执行2n bits*2n bits的运算,即可以使得运算装置800支持2n bits*2n bits的运算。

[0176] 需要说明的是,本申请实施例提供的运算装置800不仅可以支持n比特与2n比特两种定点数位宽,还可支持4n比特、8n比特等其它更多种定点数位宽。说明如下。

[0177] 继续参见图9,MC支持n比特的定点数位宽;由2行2列MC组成的MU支持2n比特的定点数位宽;可以理解到,由2行2列的MU组成的计算单元,即由4行4列的MC组成的计算单元,可以支持4n比特的定点数位宽;由4行4列的MU组成的计算单元,即由8行8列的MC组成的计算单元,可以支持8n比特的定点数位宽,以此类推。

[0178] 实际应用中,可以根据应用需求,通过设置第一计算单元811的运算方式,使得运

算装置800支持能够满足运算精度的定点数位宽。

[0179] 图12为第一计算单元811的结构示意图。

[0180] 第一计算单元811包括权重移位寄存器(Weight Shift Register)、特征图数据移位寄存器(FM Data Shift Register)、权重寄存器(Weight Register)、特征图数据寄存器(FM Data Register)、乘法电路(Mutiplier电路)、乘积寄存器(Product Register)、先移位后累加操作电路(Carry Adder电路)和累加电路(Accumulate Adder电路)。

[0181] 权重移位寄存器负责缓存从权重输入单元830或上一级第一计算单元811(也可以称为前级811)送来的权重值。在权重装载的移位阶段,权重移位寄存器缓存的权重值会向下传递到下一级第一计算单元811(也可以称为后级811)。在权重装载的装载阶段,权重移位寄存器缓存的权重值会被锁存到权重寄存器。

[0182] 特征图数据移位寄存器负责缓存从特征数据输入单元840或左边的第一计算单元811送来的特征图数据。特征图数据移位寄存器存的特征图数据会被锁存到特征图数据寄存器,同时还会被送到右边的第一计算单元811。左边的第一计算单元811表示在输入特征值在脉动阵列中的流动方向上的前一级第一计算单元811。右边的第一计算单元811表示在输入特征值在脉动阵列中的流动方向上的后一级第一计算单元811。

[0183] 乘法电路负责将权重寄存器和特征图数据寄存器缓存的权重值和特征值进行乘法操作,操作结果被送到乘积寄存器中。先移位后累加操作电路负责将乘积寄存器中的数据与当前第一计算单元811所属的第二计算单元812中的其他第一计算单元811的运算结果进行先移位后累加操作。先移位后累加操作电路的累加结果在累加电路中与上一级第一计算单元811(也可称为前一级第一计算单元811)送入的乘累加计算结果再次累加后向下传递到下一级第一计算单元811(也可称为下一级第一计算单元811)。

[0184] 作为一个示例,第一计算单元811充当图11中的MC(U1_0)。当进行前文描述的(二)2n bits*n bits的运算时,第一计算单元811中的先移位后累加操作电路负责如下运算。

[0185] RM(U1_0)<<8+RA(U0_0)。

[0186] 作为另一个示例,第一计算单元811充当图11中的MC(U1_0)。当进行前文描述的(三)2n bits*2n bits的运算时,第一计算单元811中的先移位后累加操作电路负责如下运算。

[0187] RA(U0_0)+RM(U1_0)[7:0]<<8+RM(U0_1)[7:0]<<8。

[0188] 作为又一个示例,第一计算单元811充当图11中的MC(U0_1)。当进行前文描述的(三)2n bits*2n bits的运算时,第一计算单元811中的先移位后累加操作电路负责如下运算。

[0189] RM(U0_1)[31:8]+RM(U1_0)[31:8]。

[0190] 作为再一个示例,第一计算单元811充当图11中的MC(U1_1)。当进行前文描述的(三)2n bits*2n bits的运算时,第一计算单元811中的先移位后累加操作电路负责如下运算。

[0191] RM(U1_1)+RA(U1_0)[31:16]。

[0192] 需要说明的是,在运算装置800需要进行n比特*n比特的运算的情况下,先移位后累加操作电路负责将乘积寄存器中的数据与0进行先移位后累加操作,如图12所示,即这种情形下,先移位后累加操作电路只用于透传数据,而不处理数据。

[0193] 对于不具有后级计算单元的第一计算单元811来说，累加电路将获得的累加结果直接送入输出处理单元850(下文将描述输出处理单元850)。

[0194] 继续参见图8与图9，运算装置800还包括输出处理单元850。

[0195] 输出处理单元850用于处理脉动阵列810输出的运算结果，获得输出特征数据。

[0196] 作为一个示例，假设脉动阵列810包括前文结合图4-图7描述的例子中的计算单元C00、C01、C02、C10、C11、C12、C20、C21与C22，输出处理单元850用于，接收计算单元C20、C21与C22输出的运算结果(应理解，是卷积操作的中间计算结果)，并对第三个周期结束时计算单元C20的计算结果、第五个周期结束时计算单元C21的计算结果、以及第七个周期结束时计算单元C22的计算结果作累加，得到输出特征数据a11*W11+a12*W12+a13*W13+a21*W21+a22*W22+a23*W23+a31*W31+a32*W32+a33*W33。

[0197] 作为示例，输出处理单元850的结构示意图如图9所示。输出处理单元850包括累加(Accumulate，ACC)阵列851、结果处理(Rslt_Proc)单元852与存储(Psum_Mem)单元853。

[0198] ACC阵列851的列大小与脉动阵列810的列大小一致。

[0199] 假设脉动阵列的大小为M*N，即脉动阵列810包括M行N列的第一计算单元811，则ACC阵列851的大小为1*N，即ACC阵列851包括1行N列的ACC。

[0200] 对应于2行c列的第一计算单元811可作为整体形成第二计算单元812，ACC阵列851中每c个ACC作为整体可形成一个ACC组。

[0201] 例如，脉动阵列810中，2行2列的第一计算单元811可作为整体形成第二计算单元812，则ACC阵列851中每2个ACC作为整体可形成一个ACC组(ACC_GRP)单元，即ACC阵列851共有N/2个ACC组(ACC_GRP)单元。

[0202] 需要说明的是，类似于第二计算单元812，ACC组(ACC_GRP)单元仅为了便于理解与描述而引入，并非表示ACC阵列851中实际包含ACC组单元。换句话说，在本文中，将ACC阵列851中每2个ACC作为一个整体所形成的单元记为ACC组单元。

[0203] 结果处理(Rslt_Proc)单元852负责处理ACC阵列851输出的计算结果。

[0204] 以运算装置800用于进行卷积操作为例。如果ACC阵列851输出的计算结果为卷积计算的最终结果，则结果处理单元852会将其输出，例如，将其发送到运算装置800外部的输出模块中进行后续处理。如果ACC阵列851输出的计算结果为卷积计算的中间结果，则结果处理单元852会将输出的计算结果送入存储(Psum_Mem)单元853中。

[0205] 存储(Psum_Mem)单元853负责缓存ACC阵列851输出的中间结果。还以运算装置800用于进行卷积操作为例，存储单元853负责缓存卷积计算的中间结果。

[0206] 作为示例，存储单元853可以包括数量与脉动阵列810的列大小相匹配的FIFO。假设，脉动阵列810的大小为M*N，即列大小为N，则存储单元853可以由N个FIFO组成。

[0207] 存储单元853中的每个FIFO都可以同时进行读写操作。在进行卷积运算时，N个FIFO会根据卷积核的大小分成不同的组。不同的FIFO组缓存不同卷积核的中间计算结果。

[0208] 前文描述的，如果ACC阵列851输出的计算结果为卷积计算的中间结果，则结果处理单元852会将输出的计算结果送入存储(Psum_Mem)单元853中，具体为，结果处理单元852会将输出的计算结果送入存储单元853中对应的FIFO组。

[0209] 继续参见图9，输出处理单元850中的ACC阵列851中还包括拼接单元854，拼接单元854与ACC组单元一一对应。拼接单元854用于对形成ACC组单元的2个ACC的输入数据进行拼

接。

[0210] 在运算装置800需要处理定点数位宽为2n比特的输入特征数据与定点数位宽为2n比特的权重的情况下,输出处理单元850用于执行如下操作。

[0211] 1)对脉动阵列810输出的对应于同一个第二计算单元812的低2n比特位运算结果与高2n比特位运算结果进行拼接,获得同一个第二计算单元812的4n比特运算结果。

[0212] 2)对对应于同一个权重矩阵的p个第二计算单元812的4n比特运算结果进行累加,以获得权重矩阵对应的输出特征数据,p等于权重矩阵的宽度。

[0213] 例如,操作1)由拼接单元854执行;操作2)由ACC组单元中对应输出高2n比特位的第一计算单元811的一个ACC执行。

[0214] 在运算装置800用于执行卷积操作的情形下,权重矩阵为卷积核。

[0215] 在运算装置800用于执行池化操作的情形下,权重矩阵为池化矩阵。

[0216] 输出处理单元850可以根据控制单元820的控制指令,执行包括拼接动作的累加操作,或不包括拼接动作的累加操作。

[0217] 作为一个示例,控制单元820在运算装置800需要进行上文描述的(一)n bits*n bits的运算,或者(二)2n bits*n bits的运算的情况下,向输出处理单元850发送控制指令1,在运算装置800需要进行上文描述的(三)2n bits*2n bits的运算的情况下,向输出处理单元850发送控制指令2。

[0218] 输出处理单元850在接收到控制指令1的情况下,将工作模式切换为模式(MODE)0,如图9中所示的模式(MODE)0;在接收到控制指令2的情况下,将工作模式切换为模式1,如图9中所示的模式1。

[0219] 输出处理单元850在模式0下的操作流程为,ACC阵列851中的ACC从相应的第一计算单元811中获取脉动阵列810的输出结果,然后对这些输出结果进行累加,从而获得输出特征数据。即在模式0下,输出处理单元850不执行拼接动作。

[0220] 输出处理单元850在模式1下的操作流程为,拼接单元854对第二计算单元812中的2个第一计算单元811分别输出的低2n比特位与高2n比特位进行拼接,获得该第二计算单元812的4n比特运算结果,并将该4n比特运算结果送入该拼接单元854所属的ACC组单元中的高位ACC(即对应输出高2n比特位的第一计算单元811的ACC)中;高位ACC对P个第二计算单元812的4n比特运算结果进行累加,获得输出特征数据。

[0221] ACC阵列851中的ACC的结构示意图如图13所示。

[0222] ACC单元包括脉动阵列累加寄存器(mc_psum Register)、ACC累加寄存器(acc_psum Register)、加和寄存器(sum Register)、滤波器电路(Filter Circuitry)、延迟电路(Delay Circuitry)、第一级累加电路(First Stage Adder电路)以及第二级累加电路(Second Stage Adder电路)。

[0223] 滤波器电路(Filter电路)根据卷积计算时输入的参数步长值(Stride值)过滤掉脉动阵列810输出的冗余的累加值(Psum值),同时,会将未经过滤的累加值(Psum值)送入脉动阵列累加寄存器(mc_psum Register)。延迟电路(Delay电路)将左边一级ACC输出的累加值(Psum值)延迟指定时钟周期后送入ACC累加寄存器(acc_psum Register),延迟的时钟周期数由卷积计算时输入的参数膨胀值(Dilation值)计算得到。

[0224] 第一级累加电路(First Stage Adder电路)负责将脉动阵列累加寄存器(mc_psum

Register)和ACC累加寄存器(acc_psum Register)中缓存的数据进行累加后送入加和寄存器(sum Register)。

[0225] 在将卷积运算的卷积核映射到脉动阵列时,连续N个ACC会映射到同一个卷积核,N的大小和卷积核的宽度相同。

[0226] N个ACC中的第一个ACC不需要接收左面一级ACC输出的累加值(Psum值)。例如,第一个ACC可以接收系统预设信号。

[0227] N个ACC中的最后一个ACC也不会将加和寄存器(sum Register)缓存的累加值(Psum值)输出到右面一级的ACC,而是在第二级累加电路(Second Stage Adder电路)中将加和寄存器(sum Register)缓存的累加值(Psum值)与从存储(Psum_Mem)单元853中读回的累加值(Psum值)累加后输出到结果处理(Rslt_Proc)单元852中。

[0228] 为了更好地理解本申请实施例,下面结合图14与图15描述两个例子。在图14与图15,第一计算单元811记为MC,第二计算单元812记为MU,以及以2行2列的MC组成一个MU为例。

[0229] 图14为使用本申请实施例提供的神经网络的运算装置800进行定点数位宽为n比特的卷积运算的示意性流程图。

[0230] 图14中虚线标记的单元(MC、ACC)只负责传递数据,不参与卷积计算。图14中,卷积核的大小为1*3*3。图14中的各个符号的含义如下。

[0231] KhaDb表示输入特征图中对应卷积核第a行的第b个数。Kwc表示卷积核中第c列的权重值向量,它会在卷积运算开始时部署到相应的MC中。KwcDd表示输出特征图对应卷积核第c列的第d个Psum值。Bias表示卷积运算输入的偏置值。SxTy表示第x级ACC在y时刻输出的累加值(Psum值)。

[0232] 卷积运算开始时,卷积核中的权重值向量Kwc会分三个时钟周期送入脉动阵列810(MAC Array),每个MC装载3*3卷积核中对应位置的权重值。权重装载完毕后,输入特征图的特征值按照图14中的顺序依次送入脉动阵列810(MAC Array),这些特征值在脉动阵列810中与权重值进行乘累加。

[0233] 脉动阵列810输出的累加值(Psum值)的顺序如图14所示。从脉动阵列810中输出的累加值(Psum值)送入对应的ACC中继续进行累加。ACC单元每个时刻进行的计算如图14所示,第3级ACC完成累加操作后,即可得到最终的输出特征图的特征值。

[0234] 图15为使用本申请实施例提供的神经网络的运算装置800进行定点数位宽为2n比特的卷积运算的示意性流程图。

[0235] 图15中虚线标记的单元(MC、ACC)只负责传递数据,不参与卷积计算。图15中,卷积核的大小为1*3*3。图15中的各个符号的含义如下。

[0236] KhaDb_LSB表示输入特征图中对应卷积核第a行的第b个数低n bits数;KhaDb_MSB表示输入特征图中对应卷积核第a行的第b个数的高n bits数。Kwc_LSB表示卷积核中第c列权重值向量的低nbits数;Kwc_MSB表示卷积核中第c列权重值向量的高nbits数,它们会在卷积运算开始时部署到相应的MC中。KwcDd_LSB表示输出特征图对应卷积核第c列的第d个Psum值的低位;KwcDd_MSB为输出特征图对应卷积核第c列的第d个Psum值的高位。Bias表示卷积运算输入的偏置值。SxTy表示第x级ACC单元在y时刻输出的Psum值。

[0237] 卷积运算开始时,卷积核中的特征值向量Kwc_LSB和Kwc_MSB会分六个时钟周期送

入脉动阵列810(MAC Array)。每个MC单元装载3*3卷积核中对应位置的权重值的对应n bits数。权重装载完毕后,输入特征图的特征值按照图15中的顺序依次送入脉动阵列810,它们在脉动阵列810中与权重值进行乘累加。

[0238] 脉动阵列810输出的Psum值的顺序如图15所示。从脉动阵列810输出的Psum值在ACC组单元(ACC_GRP)中首先会将Psum值的低位和高位拼装为一个完整的Psum值,该Psum值送入ACC单元继续进行累加。ACC单元每个时刻进行的计算如图15所示。ACC单元每个时钟周期向下一级传递一个Psum值。第3级ACC_GRP单元的第二个ACC模块的累加操作完成后,可得到最终的输出特征图的特征值。

[0239] 可选地,在一些实施例中,定点数位宽为2n比特的输入特征数据在外部存储器中的存储格式为,输入特征图中每行输入特征数据的低n比特位与高n比特位分别集中存储。

[0240] 例如,定点数位宽为2n bits的特征数据在SRAM中存储的格式如图16所示,该特征图中每行特征值的高n bits和低n bits分别集中存放。

[0241] 可选地,在一些实施例中,定点数位宽为2n比特的权重在外部存储器中的存储格式为,权重矩阵中每行权重的低n比特位与高n比特位分别集中存储。

[0242] 定点数位宽为2n bits的权重在SRAM中存储的格式类似于图16所示。

[0243] 应理解,定点数位宽为2n bits的特征数据与权重在SRAM中存储的格式如图16所示,有利于特征数据与权重按照图15所示的顺序送入脉动阵列810,因此有助于提高数据读取与写入的速度。

[0244] 例如,定点数位宽为nbits的特征数据在SRAM中的存储格式如图17所示。定点数位宽为nbits的权重在SRAM中的存储格式类似于图17所示。

[0245] 本申请提供的运算装置800可以应用于深度神经网络加速器。

[0246] 如图18所示,本申请实施例还提供一种神经网络加速器1800。神经网络加速器1800包括处理模块1810、权重输入模块1820、特征数据输入模块1830、输出模块1840。

[0247] 处理模块1810为上文方法实施例提供的神经网络的运算装置800。

[0248] 权重输入模块1820,用于将权重从外部存储器读出并送入处理模块800,例如。参见图8与图9,权重输入模块1820用于将权重从外部存储器读出并送入所述处理模块1810中的权重输入单元830中。

[0249] 特征数据输入模块1830,用于将特征数据从外部存储器读出并送入处理模块800。参见图8与图9,特征数据输入模块1830用于将特征数据从外部存储器读出并送入所述处理模块1810中的特征数据输入单元840中。

[0250] 输出模块1840,用于将处理模块1810输出的输出特征数据存储到外部存储器中。

[0251] 特征数据与权重在外部存储器中的存储格式如图16或图17所示。

[0252] 例如,特征数据为定点数位宽为n比特的数据,则特征数据在外部存储器中的存储格式如图17所示。

[0253] 又例如,特征数据为定点数位宽为2n比特的数据,则特征数据在外部存储器中的存储格式如图16所示。

[0254] 例如,权重为定点数位宽为n比特的数据,则权重在外部存储器中的存储格式如图17所示。

[0255] 又例如,权重为定点数位宽为2n比特的数据,则权重在外部存储器中的存储格式

如图16所示。

[0256] 本申请实施例还提供一种神经网络的运算装置的控制方法。该运算装置包括脉动阵列，脉动阵列的处理单元为第一计算单元，第一计算单元支持乘法操作数的定点数位宽为n比特，n为2的m次方，m为正整数，第一计算单元可进行先移位后累加操作，以使得脉动阵列中2行c列的多个第一计算单元作为一个整体形成支持乘法操作数的定点数位宽为2n比特的第二计算单元。

[0257] 其中，c可以为1或2。即，第一计算单元可进行先移位后累加操作，以使得脉动阵列中2行1列的两个第一计算单元作为一个整体形成支持乘法操作数的定点数位宽为2n比特的第二计算单元；或者，第一计算单元可进行先移位后累加操作，以使得脉动阵列中2行2列的四个第一计算单元作为一个整体形成支持乘法操作数的定点数位宽为2n比特的第二计算单元。

[0258] 所述控制方法包括如图10所示的操作S1010与S1020。详见上文描述，这里不再赘述。

[0259] 可选地，S1020包括：在运算装置需要对定点数位宽为2n比特的输入特征数据与定点数位宽为2n比特的权重进行运算的情况下，控制第二计算单元中所包含的部分第一计算单元进行先移位后累加操作，使得第二计算单元中所包含的2行第一计算单元中后1行的2个第一计算单元分别输出第二计算单元的4n比特运算结果的低2n比特位与高2n比特位。

[0260] 可选地，该控制方法还包括：在运算装置需要处理定点数位宽为2n比特的输入特征数据的情况下，将定点数位宽为2n比特的输入特征数据的低n比特位与高n比特位分别送入第二计算单元中所包含的2行第一计算单元中。

[0261] 可选地，c为2，该控制方法还包括：在运算装置需要处理定点数位宽为2n比特的权重的情况下，将定点数位宽为2n比特的权重的低n比特位与高n比特位分别送入第二计算单元中所包含的2列第一计算单元中。

[0262] 可选地，该控制方法还包括：对脉动阵列输出的对应于同一个第二计算单元的低2n比特位运算结果与高2n比特位运算结果进行拼接，获得同一个第二计算单元的4n比特运算结果；对对应于同一个权重矩阵的p个第二计算单元的4n比特运算结果进行累加，以获得权重矩阵对应的输出特征数据，p等于权重矩阵的宽度。

[0263] 可选地，定点数位宽为2n比特的输入特征数据在外部存储器中的存储格式为，输入特征图中每行输入特征数据的低n比特位与高n比特位分别集中存储。

[0264] 可选地，该运算装置用于执行卷积操作或池化操作。

[0265] 如图19所示，本申请实施例还提供一种神经网络处理装置1900。神经网络处理装置1900包括存储器1910与处理器1920，所述存储器1910用于存储指令，所述处理器1920用于执行所述存储器1910存储的指令，并且对所述存储器1910中存储的指令的执行，使得所述处理器1920用于执行上文方法实施例提供的控制方法。

[0266] 可选地，如图19所示，神经网络处理装置1900还包括数据接口1930，用于与外部设备进行数据传输。

[0267] 本申请实施例还提供一种计算机存储介质，其上存储有计算机程序，所述计算机程序被计算机执行时，使得所述计算机执行上文方法实施例提供的控制方法。

[0268] 本申请实施例还提供一种包含指令的计算机程序产品，其特征在于，所述指令被

计算机执行时使得计算机执行上文方法实施例提供的控制方法。

[0269] 除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同。本文中在本申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本申请。

[0270] 在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其他任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本发明实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线(digital subscriber line,DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质(例如，软盘、硬盘、磁带)、光介质(例如数字视频光盘(digital video disc,DVD))、或者半导体介质(例如固态硬盘(solid state disk,SSD))等。

[0271] 本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

[0272] 在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0273] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0274] 另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

[0275] 以上所述，仅为本申请的具体实施方式，但本申请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申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

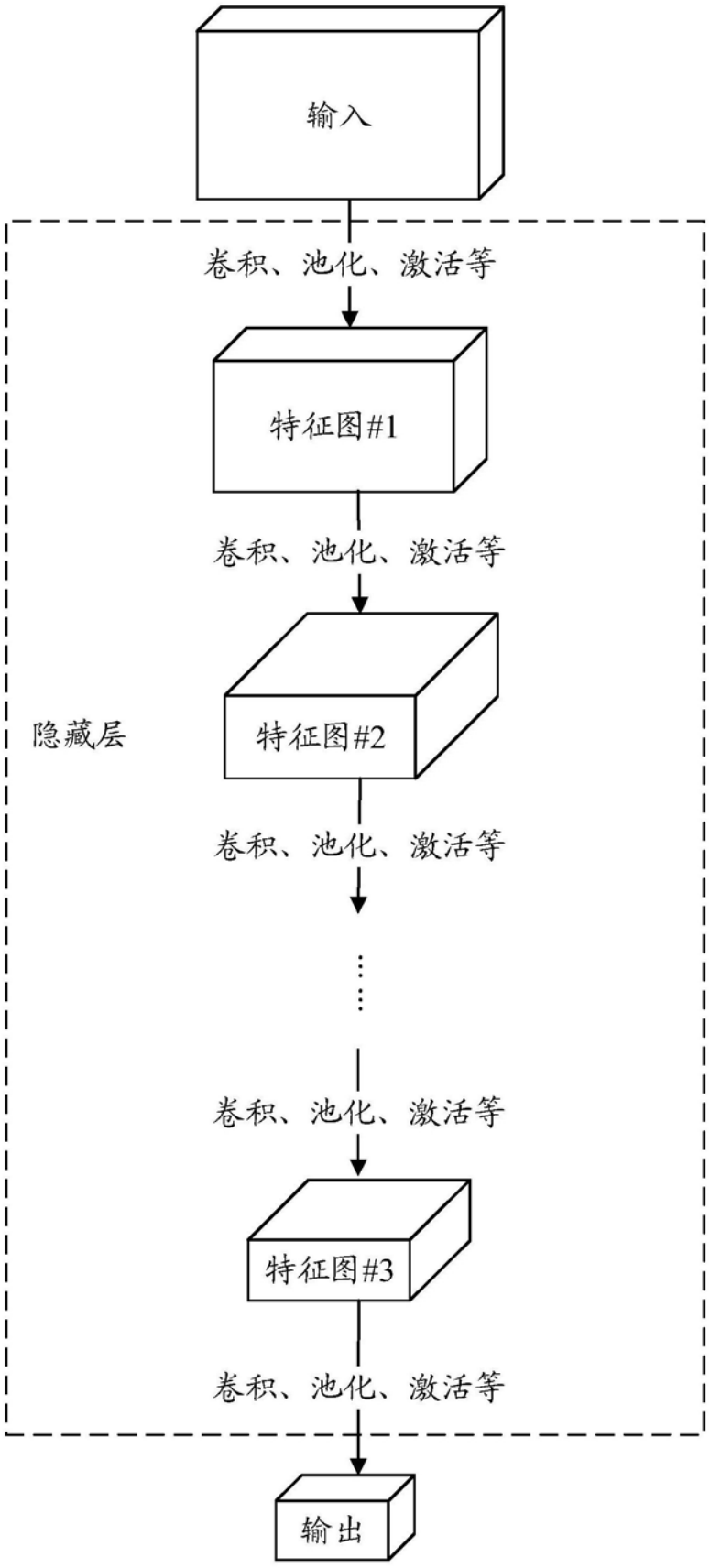

输入
卷积、池化、激活等
特征图#1
卷积、池化、激活等
隐藏层
特征图#2
卷积、池化、激活等
卷积、池化、激活等
特征图#3
卷积、池化、激活等
输出

图1

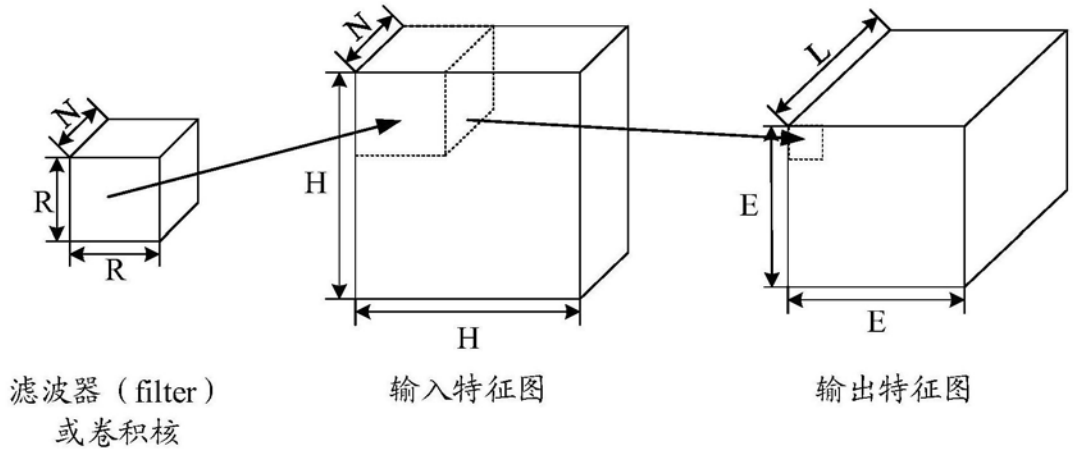

N
R
R
H
N
H
L
E
E
滤波器（filter）
或卷积核
输入特征图
输出特征图

图2

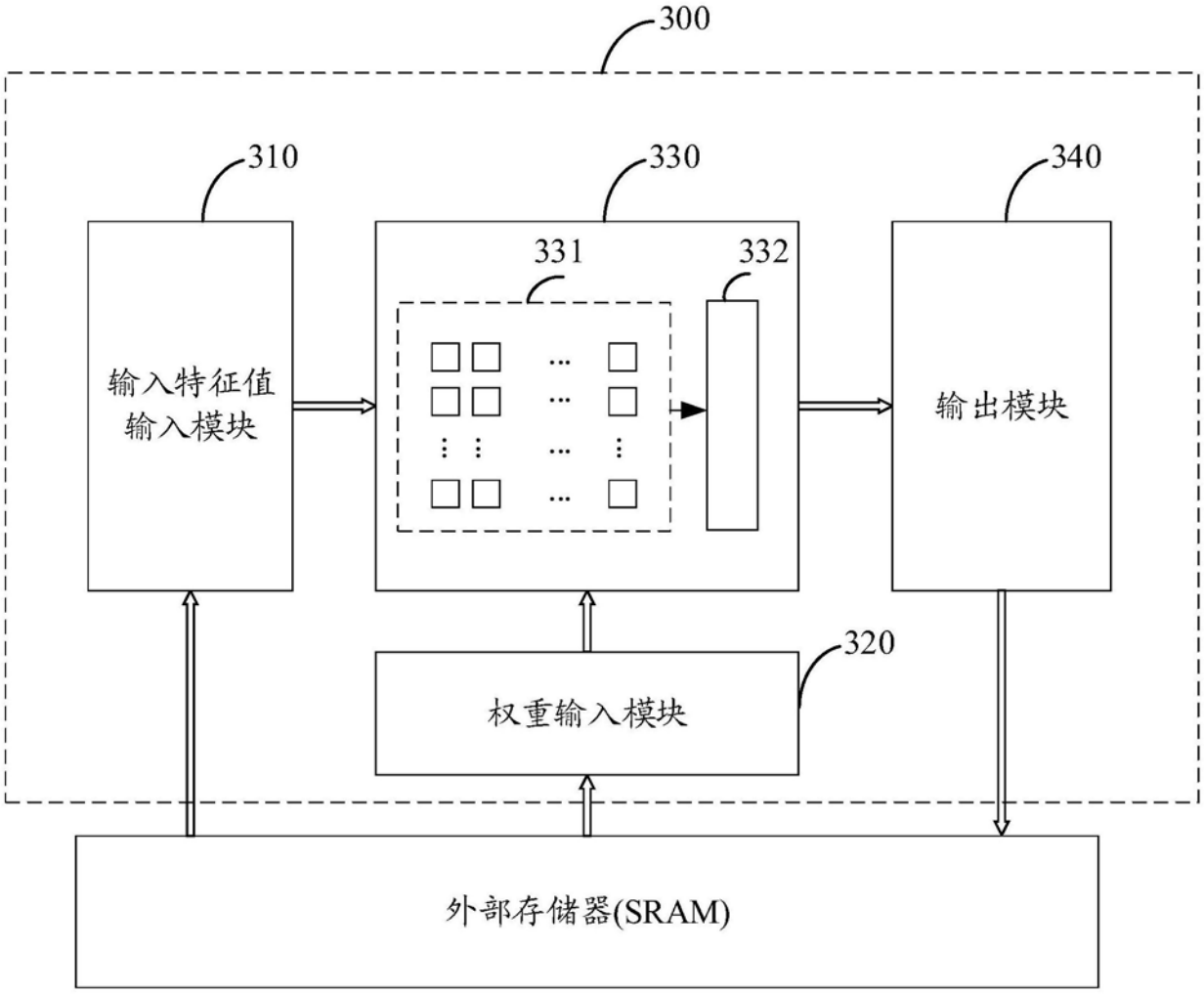

300
310
330
340
331
332
输入特征值
输入模块
输出模块
320
权重输入模块
外部存储器(SRAM)

图3

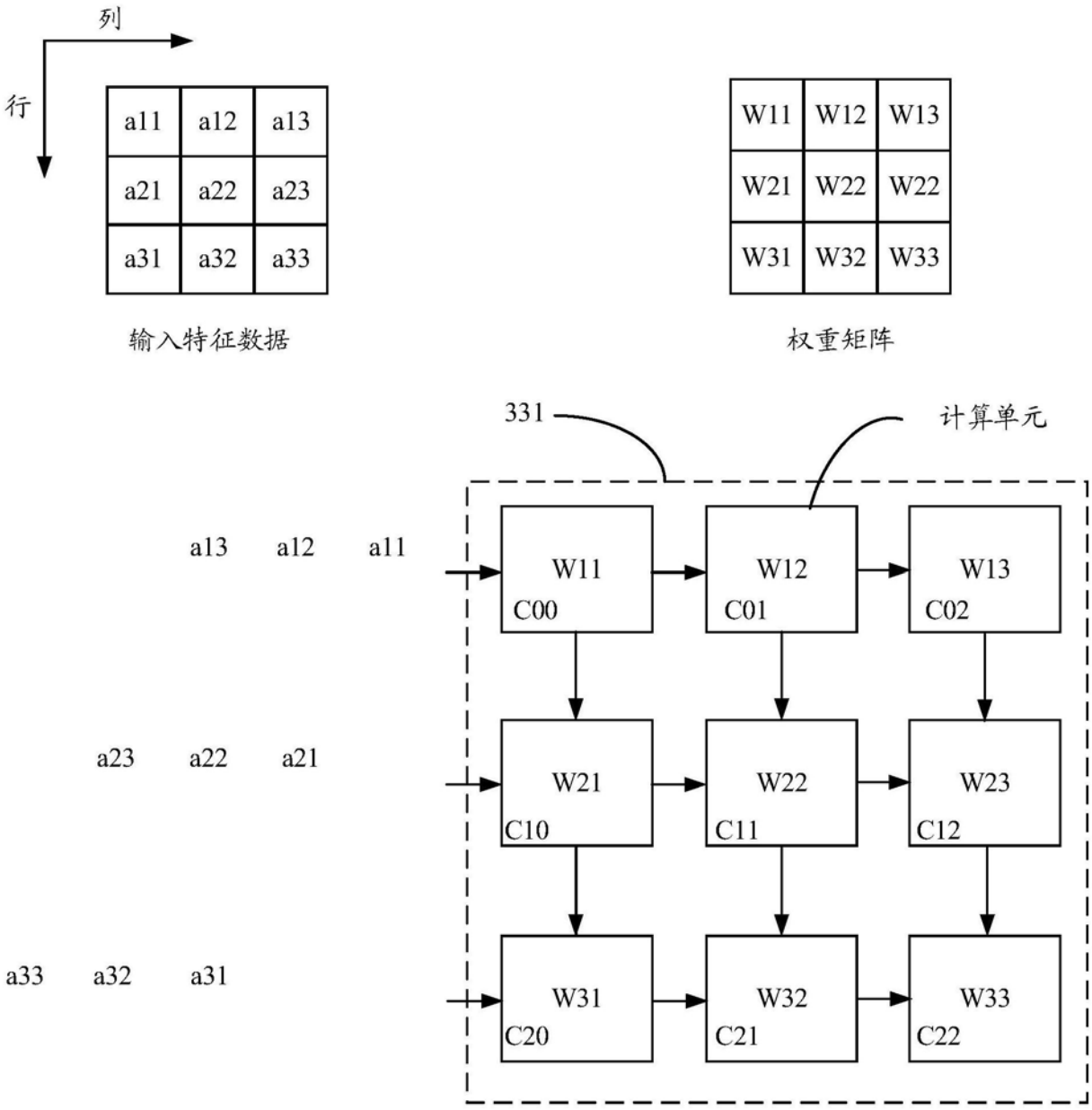

列
行
a11 a12 a13
a21 a22 a23
a31 a32 a33
输入特征数据
W11 W12 W13
W21 W22 W22
W31 W32 W33
权重矩阵
331
计算单元
a13 a12 a11
a23 a22 a21
a33 a32 a31
W11 C00
W12 C01
W13 C02
W21 C10
W22 C11
W23 C12
W31 C20
W32 C21
W33 C22

图4

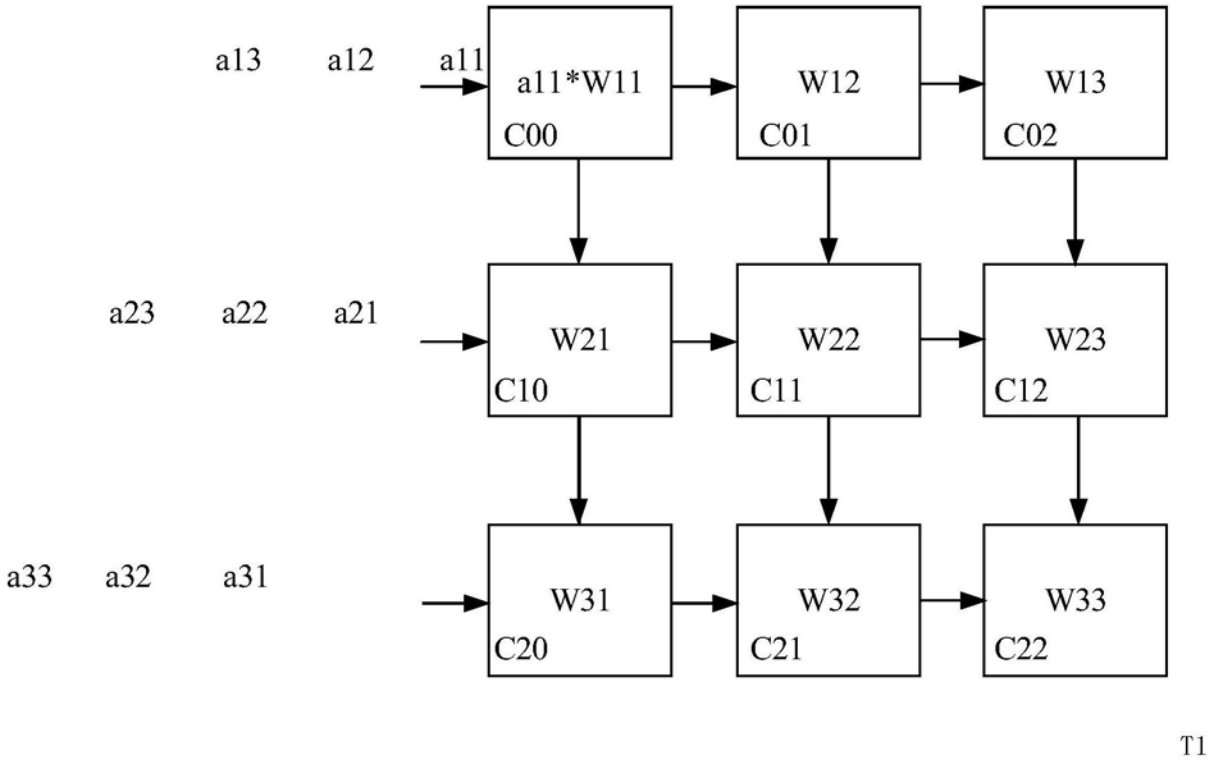

a13
a12
a11
a11*W11
C00
W12
C01
W13
C02
a23
a22
a21
W21
C10
W22
C11
W23
C12
a33
a32
a31
W31
C20
W32
C21
W33
C22
T1

图5

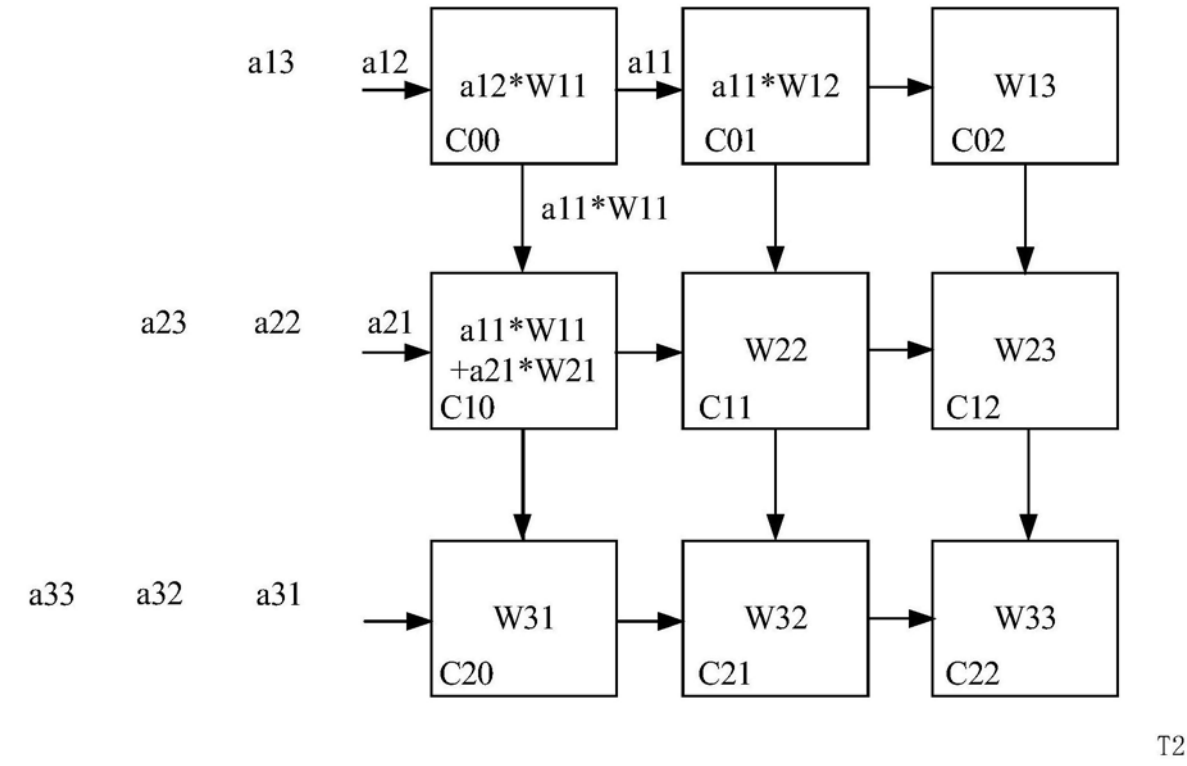

a13
a12
a12*W11
C00
a11
a11*W12
C01
W13
C02
a11*W11
a23
a22
a21
a11*W11
+a21*W21
C10
W22
C11
W23
C12
a33
a32
a31
W31
C20
W32
C21
W33
C22
T2

图6

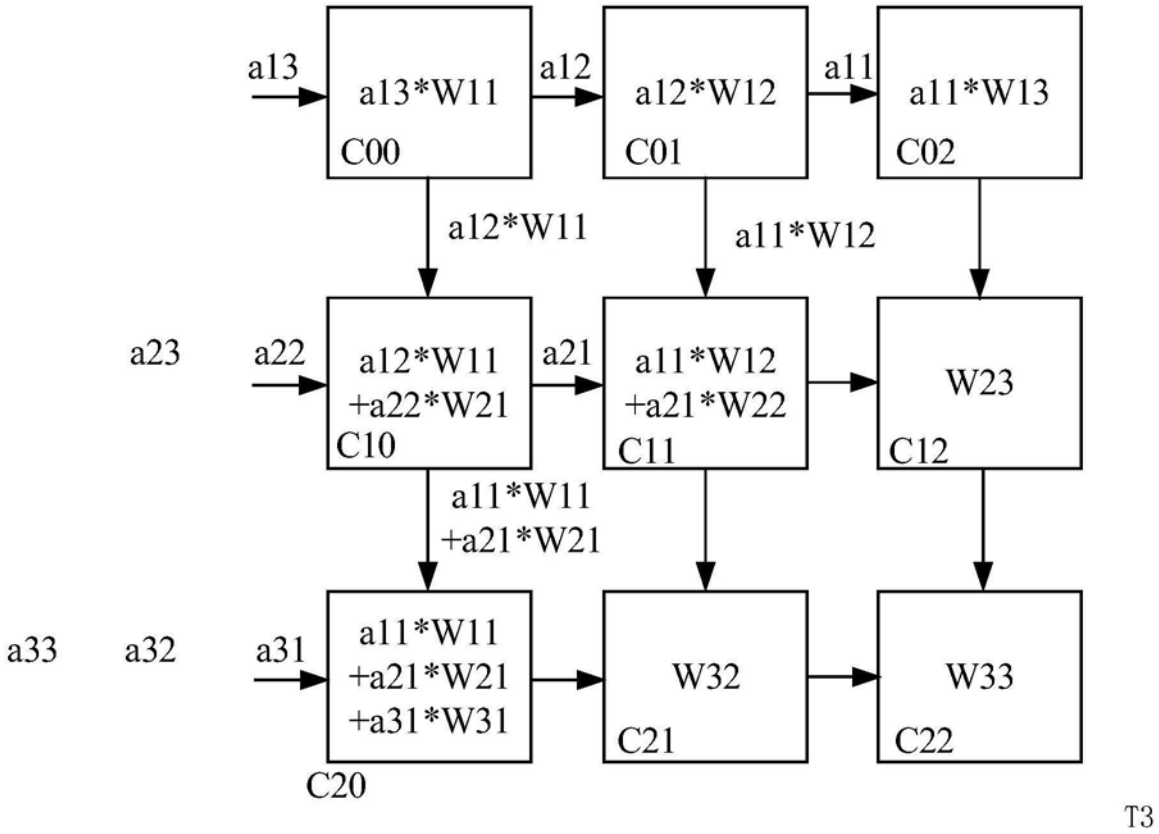

a13
a13*W11
C00
a12
a12*W12
C01
a11
a11*W13
C02
a12*W11
a11*W12
a23
a22
a12*W11
+a22*W21
C10
a21
a11*W12
+a21*W22
C11
W23
C12
a11*W11
+a21*W21
a33
a32
a31
a11*W11
+a21*W21
+a31*W31
C20
W32
C21
W33
C22
T3

图7

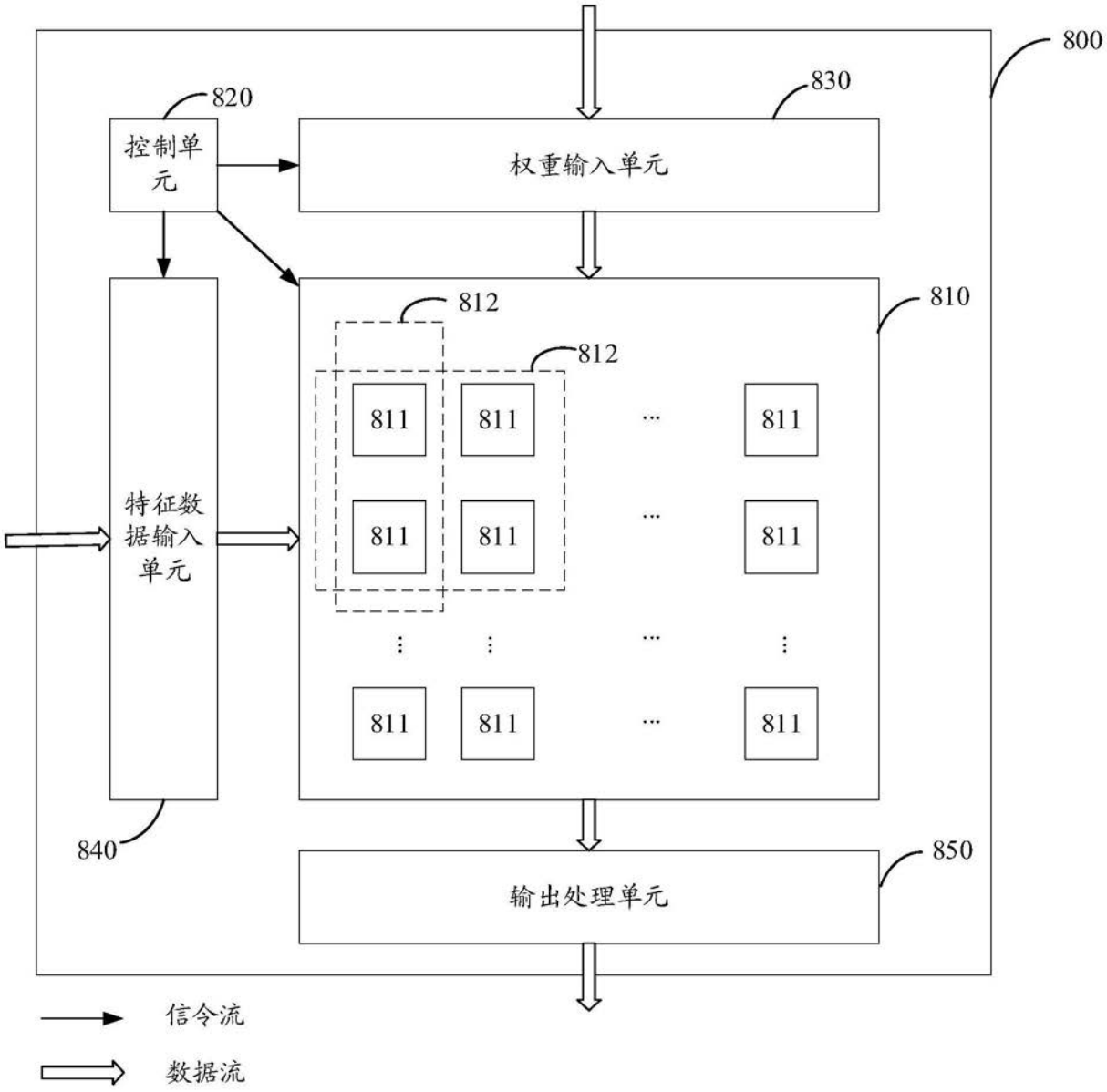

800
820
控制单元
830
权重输入单元
812
812
810
811
特征数据输入单元
840
850
输出处理单元
信令流
数据流

图8

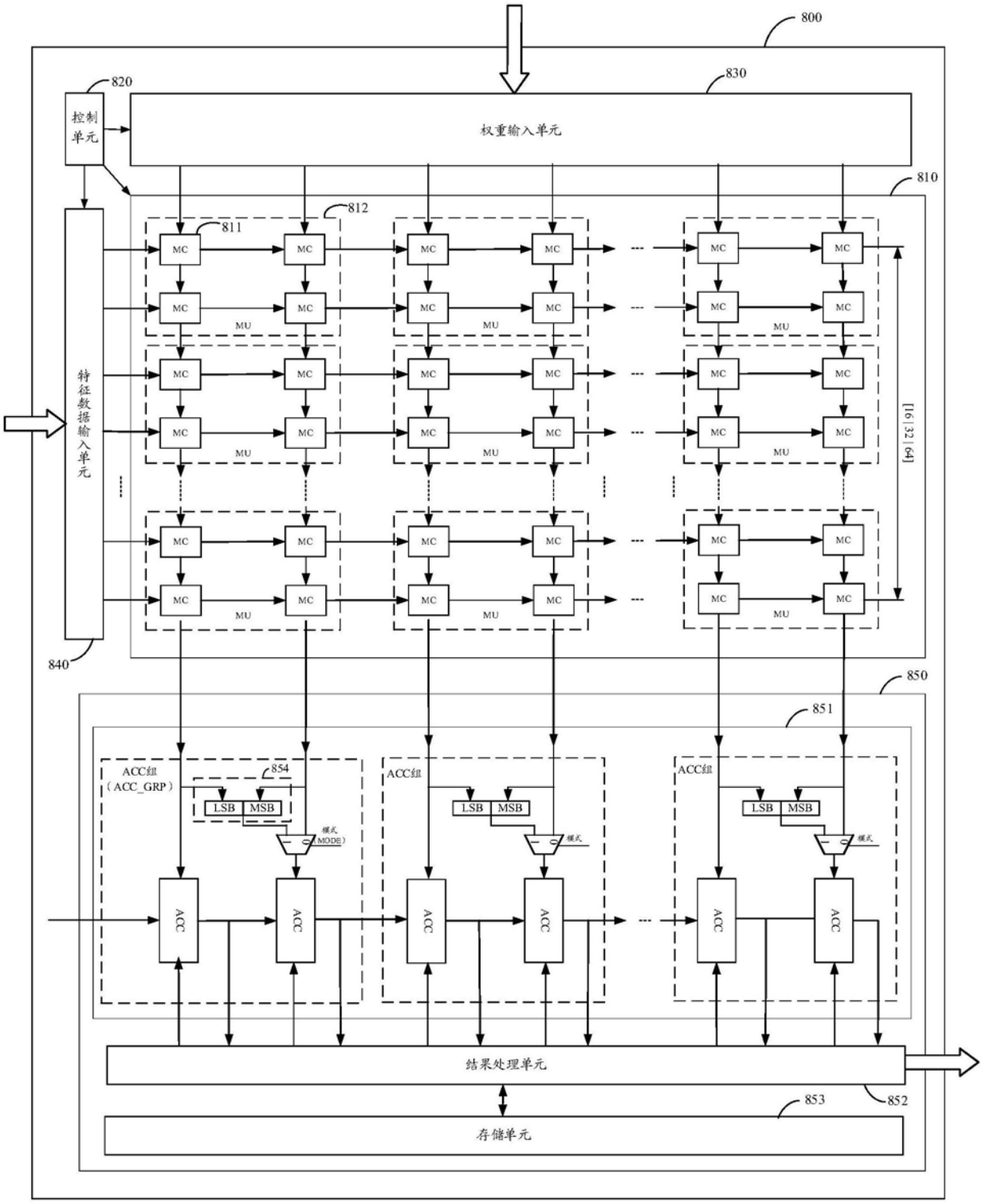

800
820
控制单元
830
权重输入单元
810
812
811
MC
MU
特征数据输入单元
[16|32|64]
840
850
851
ACC组
(ACC_GRP)
854
LSB
MSB
模式
(MODE)
ACC
结果处理单元
852
853
存储单元

图9

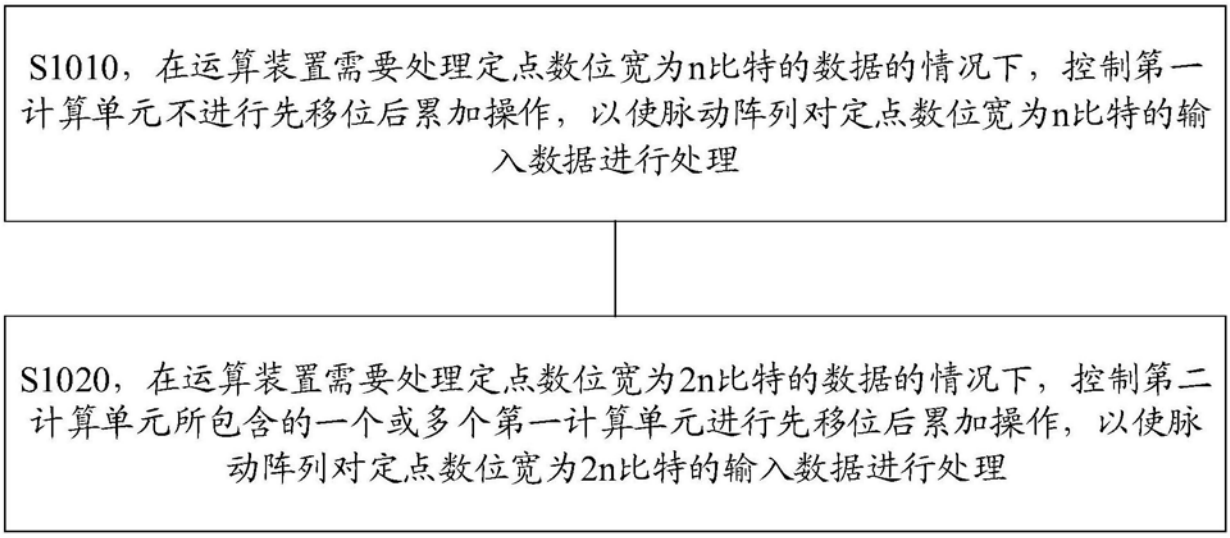

S1010，在运算装置需要处理定点数位宽为n比特的数据的情况下，控制第一计算单元不进行先移位后累加操作，以使脉动阵列对定点数位宽为n比特的输入数据进行处理
S1020，在运算装置需要处理定点数位宽为2n比特的数据的情况下，控制第二计算单元所包含的一个或多个第一计算单元进行先移位后累加操作，以使脉动阵列对定点数位宽为2n比特的输入数据进行处理

图10

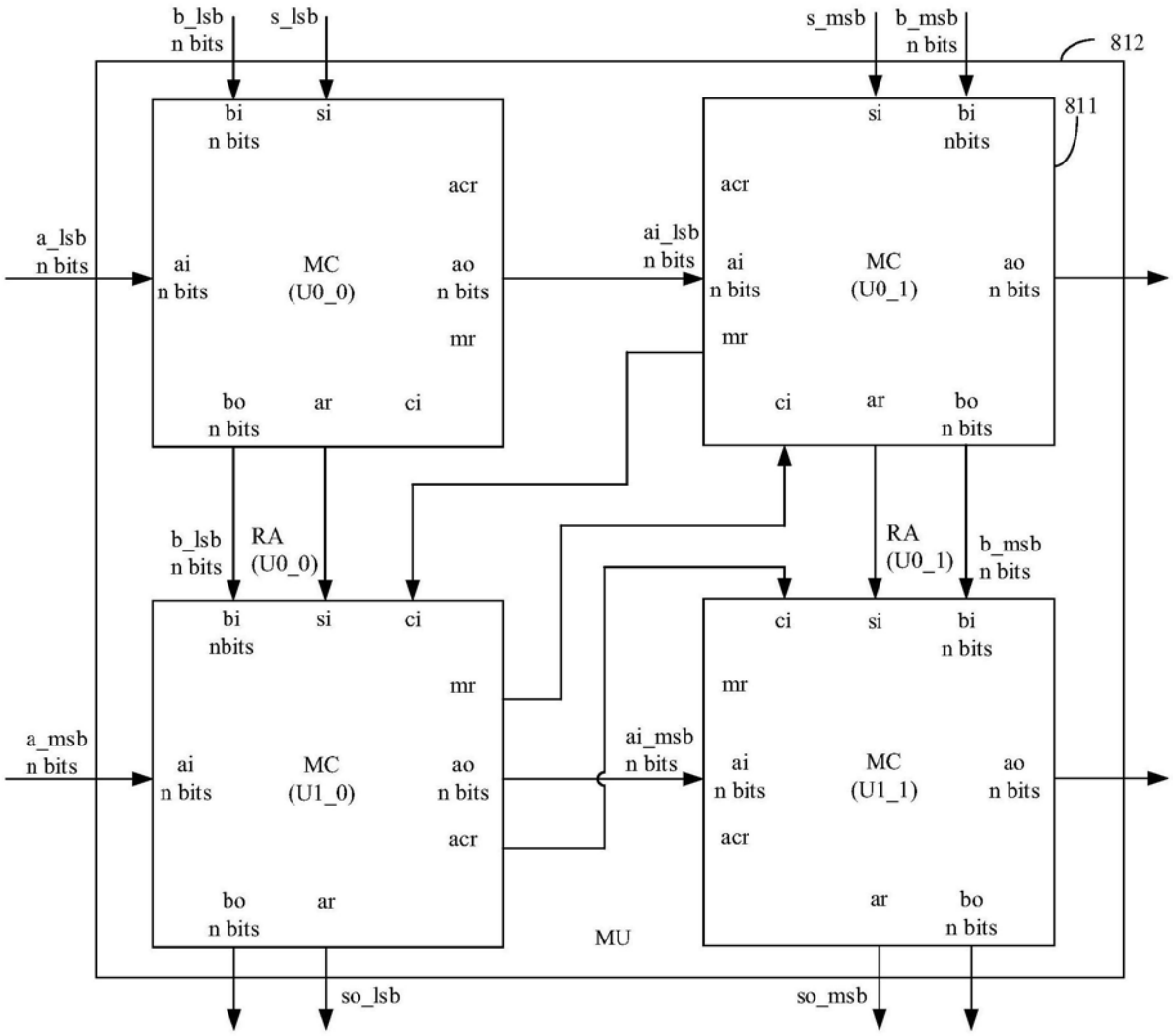

b_lsb
n bits
s_lsb
s_msb
b_msb
n bits
812
811
bi
si
n bits
acr
a_lsb
n bits
ai
n bits
MC
(U0_0)
ao
n bits
mr
bo
n bits
ar
ci
ai_lsb
n bits
MC
(U0_1)
b_lsb
n bits
RA
(U0_0)
RA
(U0_1)
b_msb
n bits
nbits
a_msb
n bits
MC
(U1_0)
ai_msb
n bits
MC
(U1_1)
MU
so_lsb
so_msb

图11

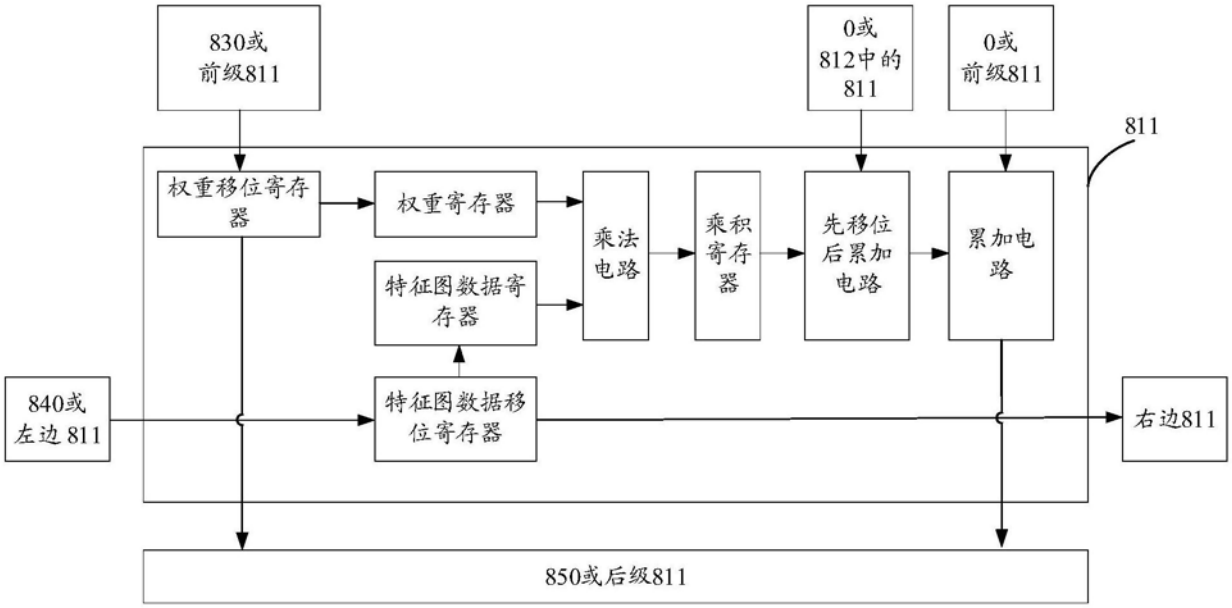

830或
前级811
0或
812中的
811
0或
前级811
811
权重移位寄存
器
权重寄存器
特征图数据寄
存器
乘法
电路
乘积
寄存
器
先移位
后累加
电路
累加电
路
840或
左边 811
特征图数据移
位寄存器
右边811
850或后级811

图12

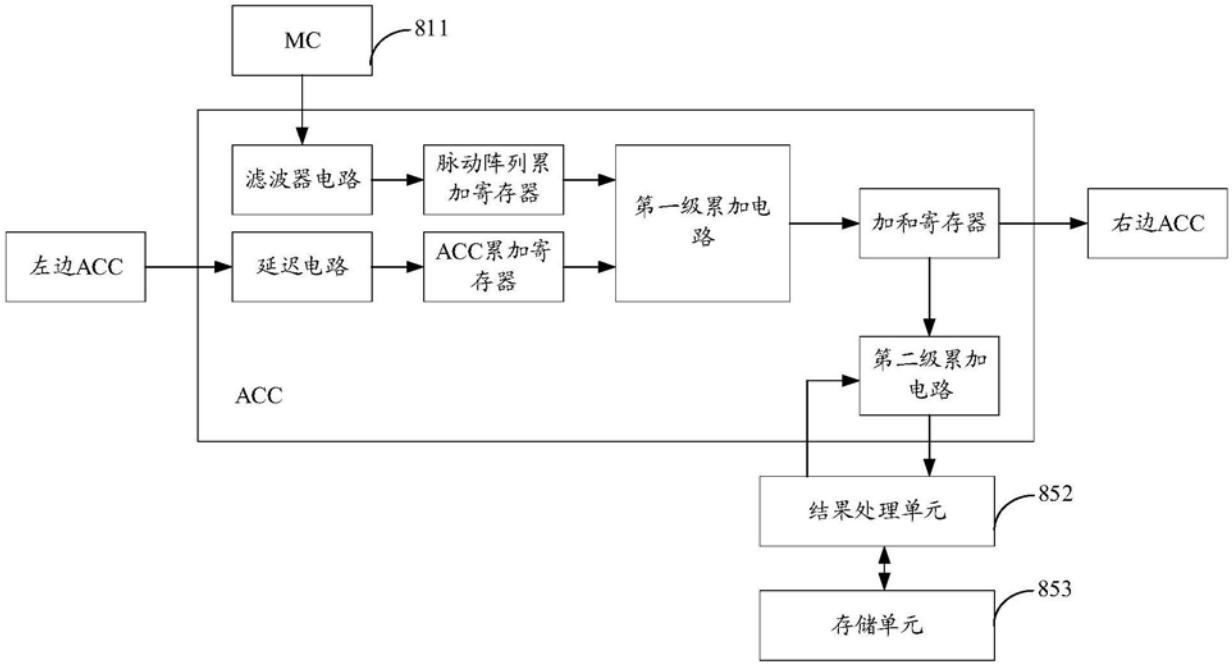

MC
811
滤波器电路
脉动阵列累
加寄存器
第一级累加电
路
加和寄存器
右边ACC
左边ACC
延迟电路
ACC累加寄
存器
第二级累加
电路
ACC
结果处理单元
852
存储单元
853

图13

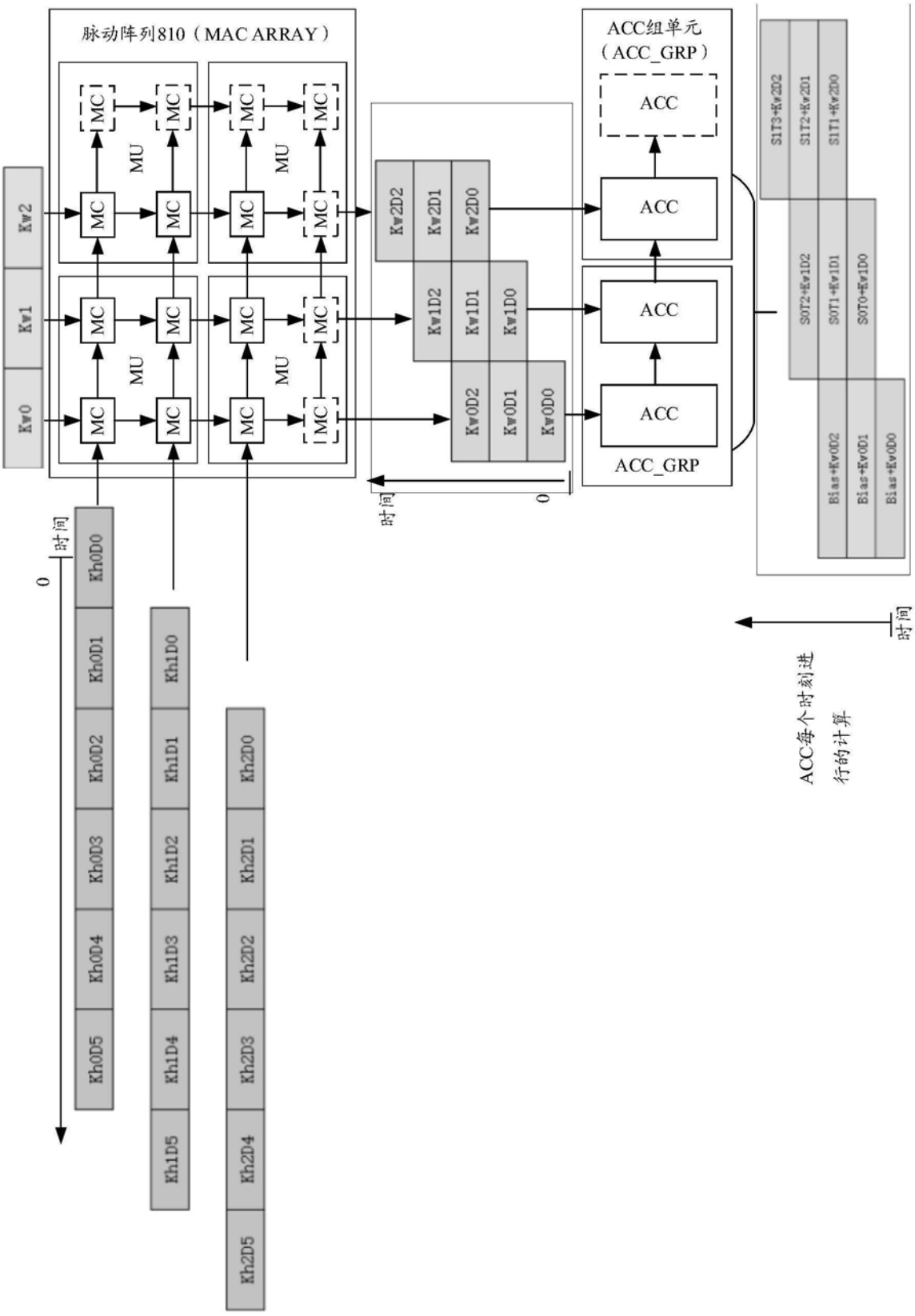

脉动阵列810（MAC ARRAY）
MC
MU
Kw2
Kw1
Kw0
ACC组单元
（ACC_GRP）
ACC
ACC_GRP
Kw2D2
Kw2D1
Kw2D0
Kw1D2
Kw1D1
Kw1D0
Kw0D2
Kw0D1
Kw0D0
S1T3+Kw2D2
S1T2+Kw2D1
S1T1+Kw2D0
S0T2+Kw1D2
S0T1+Kw1D1
S0T0+Kw1D0
Bias+Kw0D2
Bias+Kw0D1
Bias+Kw0D0
时间
0
Kh0D0
Kh0D1
Kh0D2
Kh0D3
Kh0D4
Kh0D5
Kh1D0
Kh1D1
Kh1D2
Kh1D3
Kh1D4
Kh1D5
Kh2D0
Kh2D1
Kh2D2
Kh2D3
Kh2D4
Kh2D5
ACC每个时刻进
行的计算

图14

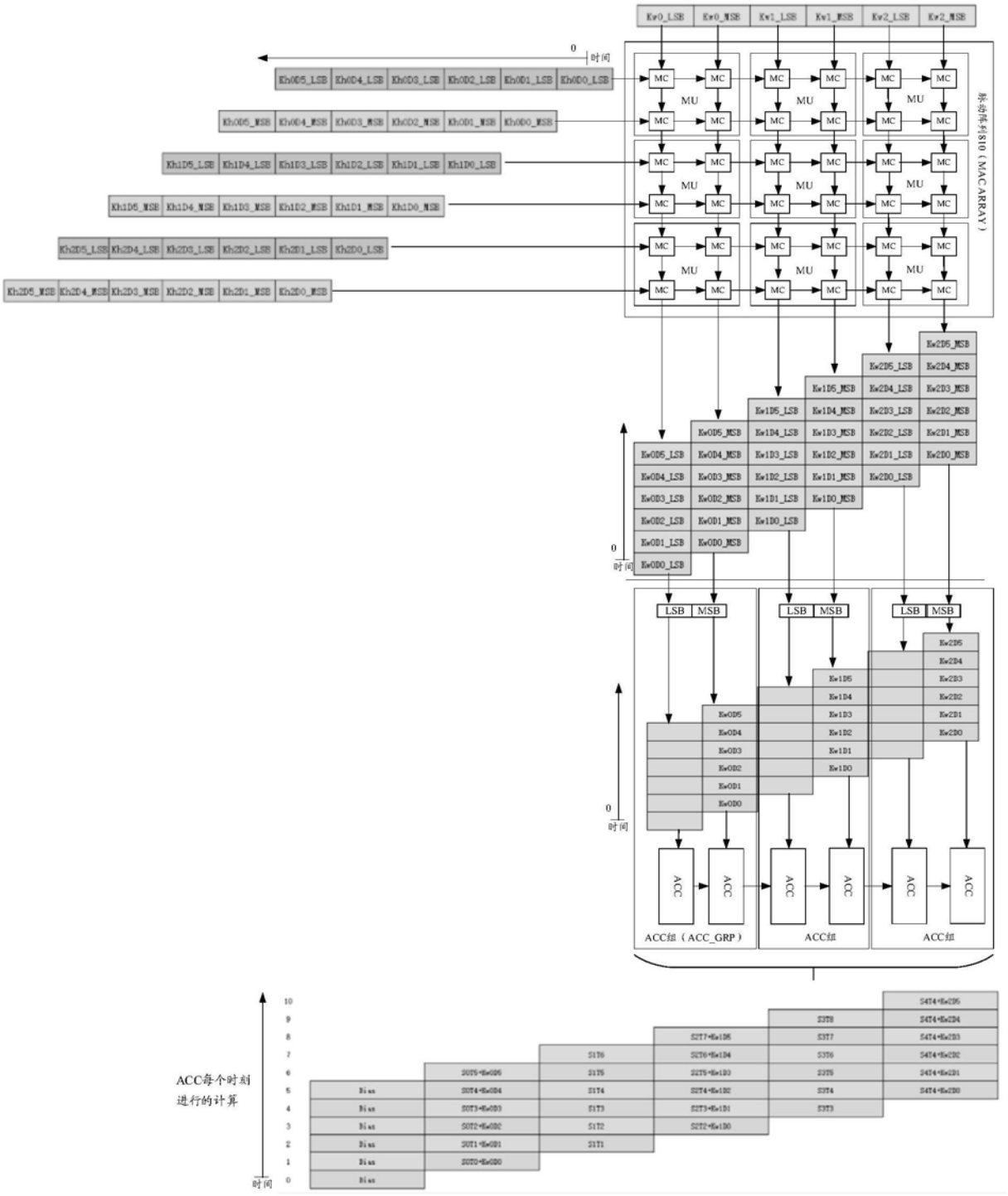


图15

m*n 比特

m	m-1		3	2	1	0
L0Dm_LSB	L0D(m-1)_LSB	——	L0D3_LSB	L0D2_LSB	L0D1_LSB	L0D0_LSB
	L0D(2m-1)_LSB	——	L0D(m+4)_LSB	L0D(m+3)_LSB	L0D(m+2)_LSB	L0D(m+1)_LSB
L0Dm_MSB	L0D(m-1)_MSB	——	L0D3_MSB	L0D2_MSB	L0D1_MSB	L0D0_MSB
	L0D(2m-1)_MSB	——	L0D(m+4)_MSB	L0D(m+3)_MSB	L0D(m+2)_MSB	L0D(m+1)_MSB
L1Dm_LSB	L1D(m-1)_LSB	——	L1D3_LSB	L1D2_LSB	L1D1_LSB	L1D0_LSB
	L1D(2m-1)_LSB	——	L1D(m+4)_LSB	L1D(m+3)_LSB	L1D(m+2)_LSB	L1D(m+1)_LSB
L1Dm_MSB	L1D(m-1)_MSB	——	L1D3_MSB	L1D2_MSB	L1D1_MSB	L1D0_MSB
	L1D(2m-1)_MSB	——	L1D(m+4)_MSB	L1D(m+3)_MSB	L1D(m+2)_MSB	L1D(m+1)_MSB

n 比特

图16

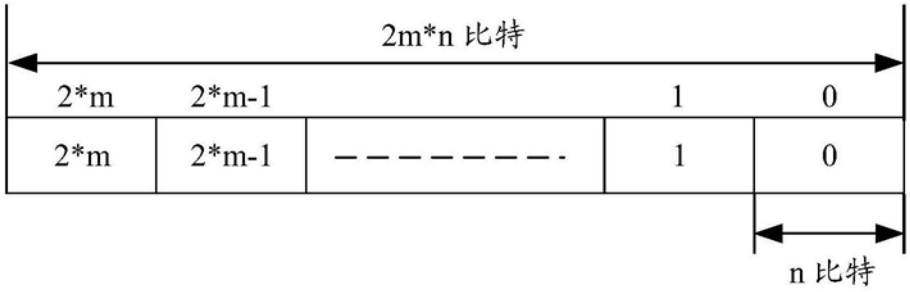


图17

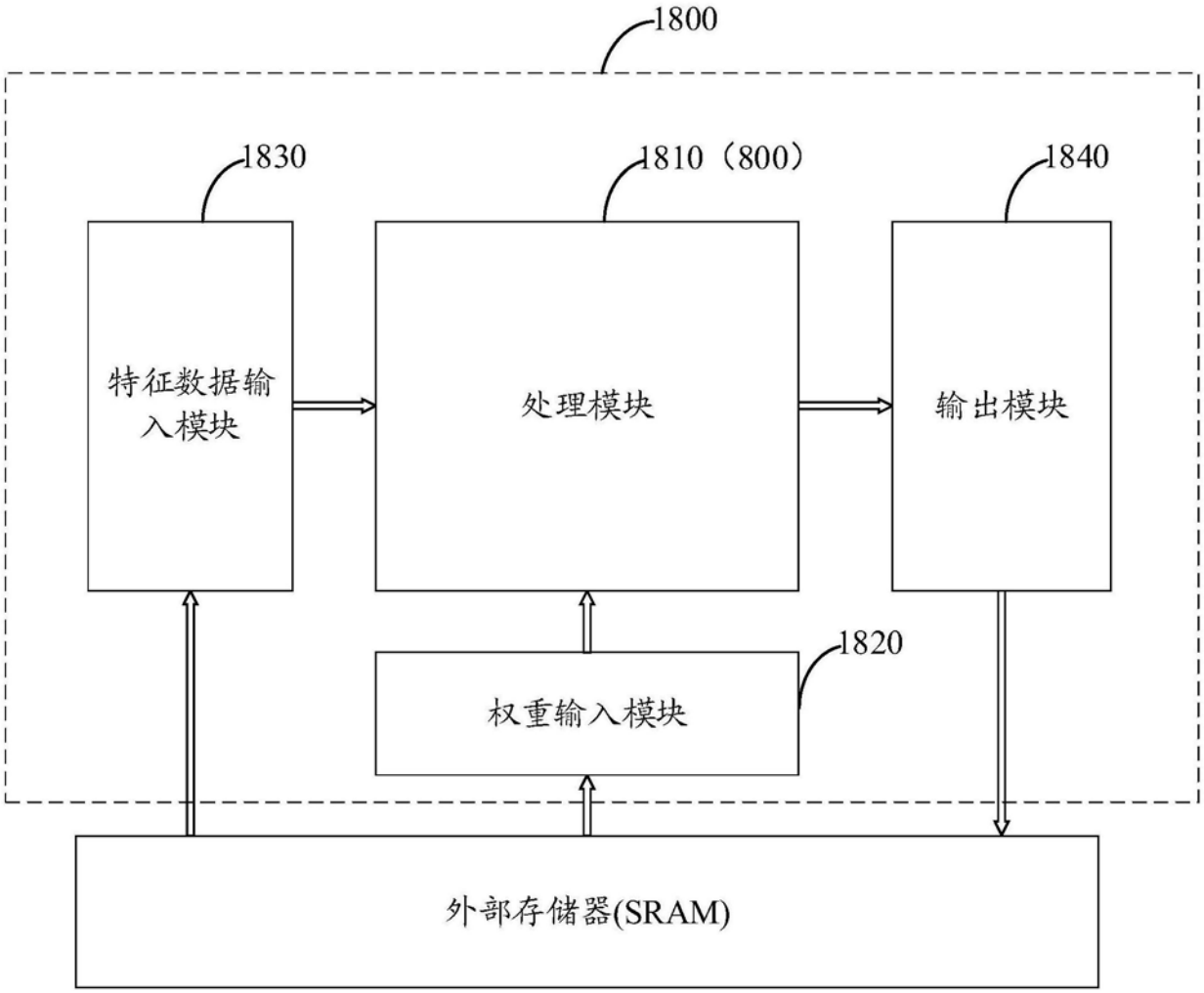

1800
1830
1810（800）
1840
特征数据输入模块
处理模块
输出模块
1820
权重输入模块
外部存储器(SRAM)

图18

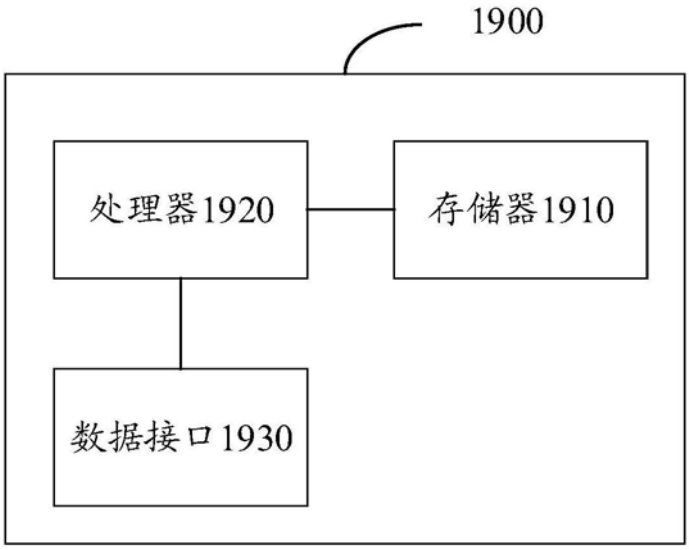

1900
处理器1920
存储器1910
数据接口1930

图19