(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 106844294 B
(45)授权公告日 2019.05.03

(21)申请号 201611243272.X

(22)申请日 2016.12.29

(65)同一申请的已公布的文献号
申请公布号 CN 106844294 A

(43)申请公布日 2017.06.13

(73)专利权人 华为机器有限公司
地址 523808 广东省东莞市松山湖科技产业园区新城大道2号

(72)发明人 徐斌 袁宏辉 何雷骏

(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 魏雪娇 毛威

(51)Int.Cl.
G06F 17/15(2006.01)

(56)对比文件
CN 1682214 A,2005.10.12,
CN 105869016 A,2016.08.17,
CN 106250103 A,2016.12.21,
CN 104915322 A,2015.09.16,
WO 03021423 A2,2003.03.13,
US 2013/0097212 A1,2013.04.18,

审查员 张红云

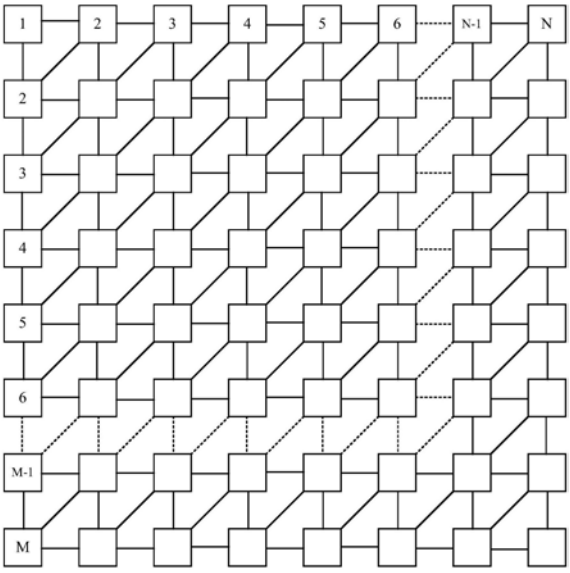
权利要求书3页 说明书19页 附图4页

(54)发明名称

卷积运算芯片和通信设备

(57)摘要

本申请提供了一种卷积运算芯片和通信设备，该卷积运算芯片包括：M×N乘法累加器阵列，包括第一乘法累加器窗口，该第一乘法累加窗口的处理单元$PE_{X,Y}$用于将$PE_{X,Y}$的卷积数据和$PE_{X,Y}$的卷积参数进行乘法运算，并将$PE_{X,Y}$的卷积参数传输至$PE_{X,Y+1}$，将$PE_{X,Y}$的卷积数据传输至$PE_{X-1,Y+1}$，分别作为$PE_{X,Y+1}$和$PE_{X-1,Y+1}$进行乘法运算的乘数；数据缓存模块，用于向第一乘法累加窗口传输卷积数据和卷积参数；输出控制模块，用于输出卷积结果。本申请的卷积运算芯片和通信设备，能够在提高阵列资源利用率的同时降低RAM访问次数，减小RAM访问压力。



CN 106844294 B

1.一种卷积运算芯片,其特征在于,包括数据缓存模块、M×N乘法累加器阵列和输出控制模块,其中,

所述数据缓存模块用于向所述M×N乘法累加器阵列中的第一乘法累加窗口传输用于卷积运算的多个卷积数据和多个卷积参数,其中,所述多个卷积参数由所述数据缓存模块根据第一卷积参数矩阵确定,所述多个卷积数据由所述数据缓存模块根据第一卷积数据矩阵确定,所述第一卷积参数矩阵为A行B列,所述第一卷积数据矩阵为D行E列,所述第一乘法累加窗口为A行C列,A为大于或等于2的整数,B和C均为大于或等于1的整数,D为大于或等于A的正整数,E为大于或等于max(B,C)的整数,M为大于或等于A的正整数,N为大于或等于C的正整数;

所述第一乘法累加窗口包括A×C个处理单元,第i行第j列的处理单元标记为$PE_{i,j}$,i按照从小到大的顺序每次取一个整数,依次从1取值到A,对应于i的每一取值,j按照从小到大的顺序每次取一个整数,依次从1取值到C;所述第一乘法累加窗口的处理单元$PE_{X,Y}$用于将$PE_{X,Y}$的卷积数据和$PE_{X,Y}$的卷积参数进行乘法运算,当C大于或等于2时,所述处理单元$PE_{X,Y}$还用于将所述$PE_{X,Y}$的卷积参数传输至$PE_{X,Y+1}$,将所述$PE_{X,Y}$的卷积数据传输至$PE_{X-1,Y+1}$,分别作为所述$PE_{X,Y+1}$和所述$PE_{X-1,Y+1}$进行乘法运算的乘数,其中,X为大于或等于2且小于或等于A的整数,Y为大于或等于1且小于或等于C-1的整数,所述$PE_{X,Y}$的卷积数据为所述数据缓存模块传输的所述多个卷积数据中的一个卷积数据,所述$PE_{X,Y}$的卷积参数为所述数据缓存模块传输的所述多个卷积参数中的一个卷积参数;

所述第一乘法累加窗口用于将$PE_{i,J}$进行乘法运算得到的乘积进行加法运算以获得卷积结果,其中,J为大于或等于1且小于或等于C的整数;

所述输出控制模块用于输出所述卷积结果。

2.根据权利要求1所述的卷积运算芯片,其特征在于,所述卷积运算芯片还包括:

阵列控制模块,用于从所述M×N乘法累加器阵列中确定用于卷积运算的所述第一乘法累加窗口,其中,根据所述第一卷积参数矩阵的行数确定所述第一乘法累加窗口的行数,根据所述第一卷积参数矩阵的行数和所述第一卷积数据矩阵的行数确定所述第一乘法累加窗口的列数。

3.根据权利要求2所述的卷积运算芯片,其特征在于,所述阵列控制模块具体根据如下公式确定所述第一乘法累加窗口的列数:

C=D-A+1。

4.根据权利要求1至3中任一项所述的卷积运算芯片,其特征在于,所述第一乘法累加窗口具体用于:

第t时钟周期,第1列处理单元$PE_{i,1}$将$PE_{i,1}$的卷积数据和$PE_{i,1}$的卷积参数进行乘法运算获得乘积$X_{i,1}^{t}$,其中,所述$PE_{i,1}$的卷积数据和所述$PE_{i,1}$的卷积参数由所述数据缓存模块传输至所述$PE_{i,1}$而获得;

将$PE_{x,1}$的卷积参数传输至$PE_{x,2}$,将$PE_{x,1}$的卷积数据传输至$PE_{x-1,2}$,分别作为所述$PE_{x,2}$和所述$PE_{x-1,2}$在第t+1时钟周期进行乘法运算的乘数,x按照从小到大的顺序每次取一个整数,依次从2取值到A;

在t分别取[nB+1,nB+B]区间内每一整数的情况下,将对应于t所有取值的所有所述乘

积$X_{i,1}^{t}$利用如下公式进行加法运算获得卷积结果S_1：

$$S_1=\sum_{t=nB+1}^{nB+B}\sum_{i=1}^{A}X_{i,1}^{t},$$

其中，n为大于或等于0并且小于或等于(E-B)的整数。

5.根据权利要求3所述的卷积运算芯片，其特征在于，当C大于或等于2时，所述第一乘法累加窗口具体还用于：

第T时钟周期，第J'列处理单元$PE_{i,J'}$将$PE_{i,J'}$的卷积数据和$PE_{i,J'}$的卷积参数进行乘法运算获得乘积$X_{i,J'}^{T}$，其中，J'为大于或等于2且小于或等于C的整数，所述$PE_{i,J'}$的卷积参数由$PE_{i,J'-1}$的卷积数据传输至所述$PE_{i,J'}$而获得，$PE_{h,J'}$的卷积数据由$PE_{h+1,J'-1}$的卷积数据传输至所述$PE_{h,J'}$而获得，$PE_{A,J'}$的卷积参数和$PE_{A,J'}$的卷积数据由所述数据缓存模块传输至所述$PE_{A,J'}$而获得，h按照从小到大的顺序每次取一个整数，依次从1取值到A-1；

在T分别取[nB+J'，nB+J'+B-1]区间内每一整数的情况下，将对应于T所有取值的所有所述乘积$X_{i,J'}^{T}$通过如下公式进行加法运算获得卷积结果$S_{J'}$：

$$S_{J'}=\sum_{T=nB+J'}^{nB+J'+B-1}\sum_{i=1}^{A}X_{i,J'}^{T},$$

其中，n为大于或等于0并且小于或等于(E-B)的整数。

6.根据权利要求5所述的卷积运算芯片，其特征在于，所述第一卷积数据矩阵为D×E卷积数据矩阵，所述第一卷积参数矩阵为A×B卷积参数矩阵，所述D×E卷积数据矩阵包括D*E个卷积数据$a_{p,q}$，p按照从小到大的顺序每次取一个整数，依次从1取值到D，对应于p的每一取值，q按照从小到大的顺序每次取一个整数，依次从1取值到E，所述A×B卷积参数矩阵包括A*B个卷积参数$b_{p',q'}$，p'按照从小到大的顺序每次取一个整数，依次从1取值到A，对应于p'的每一取值，q'按照从小到大的顺序每次取一个整数，依次从1取值到B，所述数据缓存模块包括：

缓存器，用于缓存所述D*E个卷积数据和所述A*B个卷积参数；

计数器，用于在第nB+P时钟周期，确定所述$PE_{i,1}$的卷积数据为$a_{i,n+P}$，所述$PE_{i,1}$的卷积参数为$b_{i,P}$，其中，P取值为大于或等于1且小于或等于B的整数；

所述计数器还用于在第nB+J'+Z-1时钟周期，确定所述$PE_{A,J'}$的卷积数据为$a_{A+J'-1,n+Z}$，所述$PE_{A,J'}$的卷积参数为$b_{A,Z}$，其中，Z取值为大于或等于1且小于或等于B的整数。

7.根据权利要求1-3、5和6中任一项所述的卷积运算芯片，其特征在于，所述第一乘法累加窗口具体还用于：

第nB+J时钟周期，将乘积$X_{1,J}^{nB+J}$传输至$PE_{2,J}$，与乘积$X_{2,J}^{nB+J}$进行加法运算，获得卷积中间结果Q_1^{nB+J}，其中，所述乘积$X_{1,J}^{nB+J}$为$PE_{1,J}$在第nB+J时钟周期将$PE_{1,J}$的卷积数据和$PE_{1,J}$的卷积参数进行乘法运算获得的乘积，所述乘积$X_{2,J}^{nB+J}$为$PE_{2,J}$在第nB+J时钟周期将$PE_{2,J}$的卷积数据和$PE_{2,J}$的卷积参数进行乘法运算获得的乘积；

将$PE_{f,J}$进行加法运算得到的卷积中间结果Q_{f-1}^{nB+J}传输至$PE_{f+1,J}$，其中，f按照从小到大的顺序每次取一个整数，依次从2取值到A-1；

将所述卷积中间结果Q_{f-1}^{nB+J}与所述$PE_{f+1,J}$进行乘法运算获得的乘积$X_{f+1,J}^{nB+J}$进行加法运算，获得卷积中间结果Q_{f}^{nB+J}；

将在$PE_{A,J}$内获得的卷积中间结果Q_{A-1}^{nB+J}传输给所述输出控制模块用于缓存；

在第nB+J+1时钟周期向所述$PE_{1,J}$传输所述卷积中间结果Q_{A-1}^{nB+J}，作为在第nB+J+1时钟周期进行加法运算的累加初始值；

将第(n+1)B+J-1时钟周期获得的卷积中间结果Q_{A}^{nB+J+1}确定为卷积结果S_J。

8.根据权利要求7所述的卷积运算芯片，其特征在于，当C大于或等于2时，所述第一乘法累加窗口还包括：

第一寄存器，设置于$PE_{X,Y+1}$与所述$PE_{X,Y}$之间，用于所述$PE_{X,Y}$的卷积参数的寄存与传输；

第二寄存器，设置于所述$PE_{X,Y+1}$与$PE_{X+1,Y}$之间，用于$PE_{X+1,Y}$的卷积数据的寄存与传输；

第三寄存器，设置于所述$PE_{X,Y+1}$与$PE_{X+1,Y+1}$之间，用于卷积中间结果的寄存与传输；

其中，所述第一寄存器和所述第二寄存器还用于在所述$PE_{X,Y+1}$在进行乘法运算时使$PE_{X,Y+1}$的卷积数据和$PE_{X,Y+1}$的卷积参数节拍对齐，所述第三寄存器还用于在所述第一乘法累加窗口进行加法运算时使所述$PE_{X,Y+1}$传输的卷积中间结果与所述$PE_{X+1,Y+1}$进行乘法运算获得的乘积节拍对齐。

9.根据权利要求1-3、5和6中任一项所述的卷积运算芯片，其特征在于，所述M×N乘法累加器阵列还包括第二乘法累加窗口，其中，所述第一乘法累加窗口和所述第二乘法累加窗口没有交集。

10.根据权利要求9所述的卷积运算芯片，其特征在于，所述第一卷积数据矩阵与第二卷积数据矩阵相同，所述第二卷积数据矩阵为所述数据缓存模块向所述第二乘法累加窗口传输的卷积数据所属的卷积数据矩阵；

所述第一卷积参数矩阵与第二卷积参数矩阵不同，所述第二卷积参数矩阵为所述数据缓存模块向所述第二乘法累加窗口传输的卷积参数所属的卷积参数矩阵。

11.根据权利要求9所述的卷积运算芯片，其特征在于，所述第一卷积数据矩阵与第二卷积数据矩阵不同，所述第二卷积数据矩阵为所述数据缓存模块向所述第二乘法累加窗口传输的卷积数据所属的卷积数据矩阵；

所述第一卷积参数矩阵与第二卷积参数矩阵相同，所述第二卷积参数矩阵为所述数据缓存模块向所述第二乘法累加窗口传输的卷积参数所属的卷积参数矩阵。

12.一种通信设备，其特征在于，包括通信连接的中央处理器CPU、双倍数据速率同步动态随机存取存储器DDR SDRAM和权利要求1至11中任一项所述的卷积运算芯片，其中，所述CPU用于控制所述卷积运算芯片启动所述卷积运算，所述DDR SDRAM用于向所述卷积运算芯片的所述数据缓存模块输入所述多个卷积数据和所述多个卷积参数。

卷积运算芯片和通信设备

技术领域

[0001] 本申请涉及人工智能领域,更具体地,涉及一种卷积运算芯片和通信设备。

背景技术

[0002] 深度神经网络(Deep Neural Networks,简称“DNN”)技术已经成为人工智能领域的代表性算法,基于深度神经网络技术的字符识别、图像分类或语音识别等关键技术,已经广泛应用于搜索引擎和智能手机等产品中。其中,当前最为有效,且应用最为广泛的神经网络算法是卷积神经网络(Convolutional Neural Network,简称“CNN”)算法,简称“卷积运算”。在现有技术中,CNN算法的核心计算单元是乘加运算,乘法累加器(Multiplication Accumulator,简称“MAC”)阵列常用于矩阵乘法运算,而卷积运算可以转换为矩阵乘法运算,因此业界广泛采用MAC阵列为计算核心的专用加速硬件,例如,现场可编程门阵列(Field－Programmable Gate Array,FPGA)、专用集成电路(Application Specific Integrated Circuits,ASIC)等,以加速卷积运算的运算速度。

[0003] 在现有技术的方案中,一方面,当MAC阵列中存在多个卷积窗口同时进行卷积运算时,这些卷积窗口分布在MAC阵列中的不同位置,使得MAC阵列不是所有处理单元均会利用到。并且,当MAC阵列的尺寸和多个卷积窗口的尺寸不适配时,MAC阵列的利用率会非常低。另一方面,所谓的卷积运算可以转换成矩阵乘法运算,实际上是将有大量交叠的卷积运算平铺成两个大矩阵,该两个大矩阵之间进行乘法运算。由于两个矩阵中存在大量重复数据,而这些数据都需要从随机存取存储器(Random Access Memory,简称“RAM”)中,通过MAC阵列外部的数据通道,被输入到MAC阵列进行计算。因此存在大量重复数据被从RAM输入到MAC阵列,这样会增加RAM的访问次数。

发明内容

[0004] 本申请提供了一种卷积运算芯片和通信设备,能够在提高阵列资源利用率的同时降低RAM访问次数,减小RAM访问压力。

[0005] 第一方面,提供了一种卷积运算芯片,包括数据缓存模块、M×N乘法累加器阵列和输出控制模块,其中,所述数据缓存模块用于向所述M×N乘法累加器阵列中的第一乘法累加窗口传输用于卷积运算的多个卷积数据和多个卷积参数,其中,所述多个卷积参数由所述数据缓存模块根据第一卷积参数矩阵确定,所述多个卷积数据由所述数据缓存模块根据第一卷积数据矩阵确定,所述第一卷积参数矩阵为A行B列,所述第一卷积数据矩阵为D行E列,所述第一乘法累加窗口为A行C列,A为大于或等于2的整数,B和C均为大于或等于1的整数,D为大于或等于A的正整数,E为大于或等于max(B,C)的整数,M为大于或等于A的正整数,N为大于或等于C的正整数;所述第一乘法累加窗口包括A×C个处理单元,第i行第j列的处理单元标记为$PE_{i,j}$,i按照从小到大的顺序每次取一个整数,依次从1取值到A,对应于i的每一取值,j按照从小到大的顺序每次取一个整数,依次从1取值到C;所述第一乘法累加窗口的处理单元$PE_{X,Y}$用于将$PE_{X,Y}$的卷积数据和$PE_{X,Y}$的卷积参数进行乘法运算,当C大于或等于2

时，所述处理单元$PE_{X,Y}$还用于将所述$PE_{X,Y}$的卷积参数传输至$PE_{X,Y+1}$，将所述$PE_{X,Y}$的卷积数据传输至$PE_{X-1,Y+1}$，分别作为所述$PE_{X,Y+1}$和所述$PE_{X-1,Y+1}$进行乘法运算的乘数，其中，X为大于或等于2且小于或等于A的整数，Y为大于或等于1且小于或等于C-1的整数，所述$PE_{X,Y}$的卷积数据为所述数据缓存模块传输的所述多个卷积数据中的一个卷积数据，所述$PE_{X,Y}$的卷积参数为所述数据缓存模块传输的所述多个卷积参数中的一个卷积参数；所述第一乘法累加窗口用于将$PE_{i,J}$进行乘法运算得到的乘积进行加法运算以获得卷积结果，其中，J为大于或等于1且小于或等于C的整数；所述输出控制模块用于输出所述卷积结果。

[0006] 第一方面的卷积运算芯片通过对任意一个处理单元增加一条数据传输通道，使得相邻处理单元之间能够直接传输卷积数据和卷积参数，同时，这些数据在传输过程中都处于第一乘法累加窗口中，不再经过RAM，可以减少RAM的访问次数，降低功耗。

[0007] 在第一方面的一种可能的实现方式中，所述卷积运算芯片还包括：阵列控制模块，用于从所述M×N乘法累加器阵列中确定用于卷积运算的所述第一乘法累加窗口，其中，根据所述第一卷积参数矩阵的行数确定所述第一乘法累加窗口的行数，根据所述第一卷积参数矩阵的行数和所述第一卷积数据矩阵的行数确定所述第一乘法累加窗口的列数。在一种具体的实现方式中，所述阵列控制模块具体根据如下公式确定所述第一乘法累加窗口的列数：C＝D-A+1。在本可能的实现方式中，根据卷积参数矩阵和卷积数据矩阵的尺寸灵活的确定乘法累加窗口的尺寸，可以尽可能地提高MAC阵列的利用率以及卷积运算的效率。

[0008] 在第一方面的一种可能的实现方式中，所述第一乘法累加窗口具体用于：第t时钟周期，第1列处理单元$PE_{i,1}$将$PE_{i,1}$的卷积数据和$PE_{i,1}$的卷积参数进行乘法运算获得乘积$X_{i,1}^{t}$，其中，所述$PE_{i,1}$的卷积数据和所述$PE_{i,1}$的卷积参数由所述数据缓存模块传输至所述$PE_{i,1}$而获得；将$PE_{x,1}$的卷积参数传输至$PE_{x,2}$，将$PE_{x,1}$的卷积数据传输至$PE_{x-1,2}$，分别作为所述$PE_{x,2}$和所述$PE_{x-1,2}$在第t+1时钟周期进行乘法运算的乘数，x按照从小到大的顺序每次取一个整数，依次从2取值到A；在t分别取[nB+1,nB+B]区间内每一整数的情况下，将对应于t所有取值的所有所述乘积$X_{i,1}^{t}$利用如下公式进行加法运算获得卷积结果S_1：

[0009] $$S_1=\sum_{t=nB+1}^{nB+B}\sum_{i=1}^{A}X_{i,1}^{t},$$

[0010] 其中，n为大于或等于0并且小于或等于(E-B)的整数。

[0011] 在第一方面的一种可能的实现方式中，当C大于或等于2时，所述第一乘法累加窗口具体还用于：第T时钟周期，第J'列处理单元$PE_{i,J'}$将$PE_{i,J'}$的卷积数据和$PE_{i,J'}$的卷积参数进行乘法运算获得乘积$X_{i,J'}^{T}$，其中，J'为大于或等于2且小于或等于C的整数，所述$PE_{i,J'}$的卷积参数由$PE_{i,J'-1}$的卷积数据传输至所述$PE_{i,J'}$而获得，$PE_{h,J'}$的卷积数据由$PE_{h+1,J'-1}$的卷积数据传输至所述$PE_{h,J'}$而获得，$PE_{A,J'}$的卷积参数和$PE_{A,J'}$的卷积数据由所述数据缓存模块传输至所述$PE_{A,J'}$而获得，h按照从小到大的顺序每次取一个整数，依次从1取值到A-1；在T分别取[nB+J',nB+J'+B-1]区间内每一整数的情况下，将对应于T所有取值的所有所述乘积$X_{i,J'}^{T}$通过如下公式进行加法运算获得卷积结果$S_{J'}$：

[0012] $$S_{J'}=\sum_{T=nB+J'}^{nB+J'+B-1}\sum_{i=1}^{A}X_{i,J'}^{T},$$

[0013] 其中，n为大于或等于0并且小于或等于(E-B)的整数。

[0014] 在第一方面的一种可能的实现方式中，所述D×E卷积数据矩阵包括D*E个卷积数据$a_{p,q}$，p按照从小到大的顺序每次取一个整数，依次从1取值到D，对应于p的每一取值，q按照从小到大的顺序每次取一个整数，依次从1取值到E，所述A×B卷积参数矩阵包括A*B个卷积参数$b_{p',q'}$，p'按照从小到大的顺序每次取一个整数，依次从1取值到A，对应于p'的每一取值，q'按照从小到大的顺序每次取一个整数，依次从1取值到B，所述数据缓存模块包括：缓存器，用于缓存所述D*E个卷积数据和所述A*B个卷积参数；计数器，用于在第nB+P时钟周期，确定所述$PE_{i,1}$的卷积数据为$a_{i,n+P}$，所述$PE_{i,1}$的卷积参数为$b_{i,P}$，其中，P取值为大于或等于1且小于或等于B的整数；所述计数器还用于在第nB+J'+Z-1时钟周期，确定所述$PE_{A,J'}$的卷积数据为$a_{A+J'-1,n+Z}$，所述$PE_{A,J'}$的卷积参数为$b_{A,Z}$，其中，Z取值为大于或等于1且小于或等于B的整数。

[0015] 在第一方面的一种可能的实现方式中，所述第一乘法累加窗口具体还用于：第nB+J时钟周期，将乘积$X_{1,J}^{nB+J}$传输至$PE_{2,J}$，与乘积$X_{2,J}^{nB+J}$进行加法运算，获得卷积中间结果Q_{1}^{nB+J}，其中，所述乘积$X_{1,J}^{nB+J}$为$PE_{1,J}$在第nB+J时钟周期将$PE_{1,J}$的卷积数据和$PE_{1,J}$的卷积参数进行乘法运算获得的乘积，所述乘积$X_{2,J}^{nB+J}$为$PE_{2,J}$在第nB+J时钟周期将$PE_{2,J}$的卷积数据和$PE_{2,J}$的卷积参数进行乘法运算获得的乘积；将$PE_{f,J}$进行加法运算得到的卷积中间结果Q_{f-1}^{nB+J}传输至$PE_{f+1,J}$，其中，f按照从小到大的顺序每次取一个整数，依次从2取值到A-1；将所述卷积中间结果Q_{f-1}^{nB+J}与所述$PE_{f+1,J}$进行乘法运算获得的乘积$X_{f+1,J}^{nB+J}$进行加法运算，获得卷积中间结果Q_{f}^{nB+J}；将在$PE_{A,J}$内获得的卷积中间结果Q_{A-1}^{nB+J}传输给所述输出控制模块用于缓存；在第nB+J+1时钟周期向所述$PE_{1,J}$传输所述卷积中间结果Q_{A-1}^{nB+J}，作为在第nB+J+1时钟周期进行加法运算的累加初始值；将第(n+1)B+J-1时钟周期获得的卷积中间结果Q_{A}^{nB+J+1}确定为卷积结果S_J。

[0016] 在第一方面的一种可能的实现方式中，当C大于或等于2时，所述第一乘法累加窗口还包括：第一寄存器，设置于$PE_{X,Y+1}$与所述$PE_{X,Y}$之间，用于所述$PE_{X,Y}$的卷积参数的寄存与传输；第二寄存器，设置于所述$PE_{X,Y+1}$与$PE_{X+1,Y}$之间，用于$PE_{X+1,Y}$的卷积数据的寄存与传输；第三寄存器，设置于所述$PE_{X,Y+1}$与$PE_{X+1,Y+1}$之间，用于卷积中间结果的寄存与传输；其中，所述第一寄存器和所述第二寄存器还用于在所述$PE_{X,Y+1}$在进行乘法运算时使$PE_{X,Y+1}$的卷积数据和$PE_{X,Y+1}$的卷积参数节拍对齐，所述第三寄存器还用于在所述第一乘法累加窗口进行加法运算时使所述$PE_{X,Y+1}$传输的卷积中间结果与所述$PE_{X+1,Y+1}$进行乘法运算获得的乘积节拍对齐。

[0017] 在第一方面的一种可能的实现方式中，所述M×N乘法累加器阵列还包括第二乘法累加窗口，其中，所述第一乘法累加窗口和所述第二乘法累加窗口没有交集。

[0018] 在第一方面的一种可能的实现方式中，所述第一卷积数据矩阵与第二卷积数据矩阵相同，所述第二卷积数据矩阵为所述数据缓存模块向所述第二乘法累加窗口传输的卷积数据所属的卷积数据矩阵；所述第一卷积参数矩阵与第二卷积参数矩阵不同，所述第二卷积参数矩阵为所述数据缓存模块向所述第二乘法累加窗口传输的卷积参数所属的卷积参数矩阵。

[0019] 在第一方面的一种可能的实现方式中，所述第一卷积数据矩阵与第二卷积数据矩阵不同，所述第二卷积数据矩阵为所述数据缓存模块向所述第二乘法累加窗口传输的卷积数据所属的卷积数据矩阵；所述第一卷积参数矩阵与第二卷积参数矩阵相同，所述第二卷积参数矩阵为所述数据缓存模块向所述第二乘法累加窗口传输的卷积参数所属的卷积参数矩阵。

[0020] 第二方面，提供了一种通信设备，包括通信连接的中央处理器CPU、双倍数据速率同步动态随机存取存储器DDR SDRAM和第一方面或第一方面的任意一种可能的实现方式所述的卷积运算芯片，其中，所述CPU用于控制所述卷积运算芯片启动所述卷积运算，所述DDR SDRAM用于向所述卷积运算芯片的所述数据缓存模块输入所述多个卷积数据和所述多个卷积参数。

[0021] 第二方面，提供了一种应用于卷积运算芯片中，所述卷积运算芯片包括数据缓存模块、M×N乘法累加器阵列和输出控制模块，所述方法包括：所述数据缓存模块向所述M×N乘法累加器阵列中的第一乘法累加窗口传输用于卷积运算的多个卷积数据和多个卷积参数，其中，所述多个卷积参数由所述数据缓存模块根据第一卷积参数矩阵确定，所述多个卷积数据由所述数据缓存模块根据第一卷积数据矩阵确定，所述第一卷积参数矩阵为A行B列，所述第一卷积数据矩阵为D行E列，所述第一乘法累加窗口为A行C列，A为大于或等于2的整数，B和C均为大于或等于1的整数，D为大于或等于A的正整数，E为大于或等于max(B,C)的整数，M为大于或等于A的正整数，N为大于或等于C的正整数；所述第一乘法累加窗口包括A×C个处理单元，第i行第j列的处理单元标记为$PE_{i,j}$，i按照从小到大的顺序每次取一个整数，依次从1取值到A，对应于i的每一取值，j按照从小到大的顺序每次取一个整数，依次从1取值到C，所述第一乘法累加窗口的处理单元$PE_{X,Y}$将$PE_{X,Y}$的卷积数据和$PE_{X,Y}$的卷积参数进行乘法运算，当C大于或等于2时，所述处理单元$PE_{X,Y}$还将所述$PE_{X,Y}$的卷积参数传输至$PE_{X,Y+1}$，将所述$PE_{X,Y}$的卷积数据传输至$PE_{X-1,Y+1}$，分别作为所述$PE_{X,Y+1}$和所述$PE_{X-1,Y+1}$进行乘法运算的乘数，其中，X为大于或等于2且小于或等于A的整数，Y为大于或等于1且小于或等于C-1的整数，所述$PE_{X,Y}$的卷积数据为所述数据缓存模块传输的所述多个卷积数据中的一个卷积数据，所述$PE_{X,Y}$的卷积参数为所述数据缓存模块传输的所述多个卷积参数中的一个卷积参数；所述第一乘法累加窗口将$PE_{i,J}$进行乘法运算得到的乘积进行加法运算以获得卷积结果，其中，J为大于或等于1且小于或等于C的整数；所述输出控制模块输出所述卷积结果。

[0022] 在第三方面的一种可能的实现方式中，所述卷积运算芯片还包括阵列控制模块，所述方法还包括：所述阵列控制模块从所述M×N乘法累加器阵列中确定用于卷积运算的所述第一乘法累加窗口，其中，所述阵列控制模块根据所述第一卷积参数矩阵的行数确定所

述第一乘法累加窗口的行数，所述阵列控制模块根据所述第一卷积参数矩阵的行数和所述第一卷积数据矩阵的行数确定所述第一乘法累加窗口的列数。

[0023] 在第三方面的一种可能的实现方式中，所述阵列控制模块根据所述第一卷积参数矩阵的行数和所述第一卷积数据矩阵的行数确定所述第一乘法累加窗口的列数，包括：所述阵列控制模块根据如下公式确定所述第一乘法累加窗口的列数：C＝D-A+1。

[0024] 在第三方面的一种可能的实现方式中，所述第一乘法累加窗口将$PE_{i,J}$进行乘法运算得到的乘积进行加法运算以获得卷积结果，包括：第t时钟周期，第1列处理单元$PE_{i,1}$将$PE_{i,1}$的卷积数据和$PE_{i,1}$的卷积参数进行乘法运算获得乘积$X_{i,1}^{t}$，其中，所述$PE_{i,1}$的卷积数据和所述$PE_{i,1}$的卷积参数由所述数据缓存模块传输至所述$PE_{i,1}$而获得；将$PE_{x,1}$的卷积参数传输至$PE_{x,2}$，将$PE_{x,1}$的卷积数据传输至$PE_{x-1,2}$，分别作为所述$PE_{x,2}$和所述$PE_{x-1,2}$在第t+1时钟周期进行乘法运算的乘数，x按照从小到大的顺序每次取一个整数，依次从2取值到A；在t分别取[nB+1,nB+B]区间内每一整数的情况下，将对应于t所有取值的所有所述乘积$X_{i,1}^{t}$利用如下公式进行加法运算获得卷积结果S_1：

[0025] $$S_1=\sum_{t=nB+1}^{nB+B}\sum_{i=1}^{A}X_{i,1}^{t},$$

[0026] 其中，n为大于或等于0并且小于或等于(E-B)的整数。

[0027] 在第三方面的一种可能的实现方式中，当C大于或等于2时，所述第一乘法累加窗口将$PE_{i,J}$进行乘法运算得到的乘积进行加法运算以获得卷积结果，包括：第T时钟周期，第J’列处理单元$PE_{i,J'}$将$PE_{i,J'}$的卷积数据和$PE_{i,J'}$的卷积参数进行乘法运算获得乘积$X_{i,J'}^{T}$，其中，J’为大于或等于2且小于或等于C的整数，所述$PE_{i,J'}$的卷积参数由$PE_{i,J'-1}$的卷积数据传输至所述$PE_{i,J'}$而获得，$PE_{h,J'}$的卷积数据由$PE_{h+1,J'-1}$的卷积数据传输至所述$PE_{h,J'}$而获得，$PE_{A,J'}$的卷积参数和$PE_{A,J'}$的卷积数据由所述数据缓存模块传输至所述$PE_{A,J'}$而获得，h按照从小到大的顺序每次取一个整数，依次从1取值到A-1；在T分别取[nB+J’,nB+J’+B-1]区间内每一整数的情况下，将对应于T所有取值的所有所述乘积$X_{i,J'}^{T}$通过如下公式进行加法运算获得卷积结果$S_{J'}$：

[0028] $$S_{J'}=\sum_{T=nB+J'}^{nB+J'+B-1}\sum_{i=1}^{A}X_{i,J'}^{T},$$

[0029] 其中，n为大于或等于0并且小于或等于(E-B)的整数

[0030] 在第三方面的一种可能的实现方式中，所述D×E卷积数据矩阵包括D*E个卷积数据$a_{p,q}$，p按照从小到大的顺序每次取一个整数，依次从1取值到D，对应于p的每一取值，q按照从小到大的顺序每次取一个整数，依次从1取值到E，所述A×B卷积参数矩阵包括A*B个卷积参数$b_{p',q'}$，p’按照从小到大的顺序每次取一个整数，依次从1取值到A，应于p’的每一取值，q’按照从小到大的顺序每次取一个整数，依次从1取值到B，所述数据缓存模块包括缓存器和计数器，所述多个卷积参数由所述数据缓存模块根据第一卷积参数矩阵确定，所述多个卷积数据由所述数据缓存模块根据第一卷积数据矩阵确定，包括：所述缓存器缓存所述D*E个卷积数据和所述A*B个卷积参数；所述计数器在第nB+P时钟周期，确定所述$PE_{i,1}$的卷

积数据为$a_{i,n+P}$，所述$PE_{i,1}$的卷积参数为$b_{i,P}$，其中，P取值为大于或等于1且小于或等于B的整数；所述计数器在第nB+J'+Z-1时钟周期，确定所述$PE_{A,J'}$的卷积数据为$a_{A+J'-1,n+Z}$，所述$PE_{A,J'}$的卷积参数为$b_{A,Z}$，其中，Z取值为大于或等于1且小于或等于B的整数。

[0031] 在第三方面的一种可能的实现方式中，所述第一乘法累加窗口将$PE_{i,J}$进行乘法运算得到的乘积进行加法运算以获得卷积结果，包括：第nB+J时钟周期，将乘积$X_{1,J}^{nB+J}$传输至$PE_{2,J}$，与乘积$X_{2,J}^{nB+J}$进行加法运算，获得卷积中间结果Q_{1}^{nB+J}，其中，所述乘积$X_{1,J}^{nB+J}$为$PE_{1,J}$在第nB+J时钟周期将$PE_{1,J}$的卷积数据和$PE_{1,J}$的卷积参数进行乘法运算获得的乘积，所述乘积$X_{2,J}^{nB+J}$为$PE_{2,J}$在第nB+J时钟周期将$PE_{2,J}$的卷积数据和$PE_{2,J}$的卷积参数进行乘法运算获得的乘积；将$PE_{f,J}$进行加法运算得到的卷积中间结果Q_{f-1}^{nB+J}传输至$PE_{f+1,J}$，其中，f按照从小到大的顺序每次取一个整数，依次从2取值到A-1；将所述卷积中间结果Q_{f-1}^{nB+J}与所述$PE_{f+1,J}$进行乘法运算获得的乘积$X_{f+1,J}^{nB+J}$进行加法运算，获得卷积中间结果Q_{f}^{nB+J}；将在$PE_{A,J}$内获得的卷积中间结果Q_{A-1}^{nB+J}传输给所述输出控制模块用于缓存；在第nB+J+1时钟周期向所述$PE_{1,J}$传输所述卷积中间结果Q_{A-1}^{nB+J}，作为在第nB+J+1时钟周期进行加法运算的累加初始值；将第(n+1)B+J-1时钟周期获得的卷积中间结果Q_{A}^{nB+J+1}确定为卷积结果S_J。

[0032] 在第三方面的一种可能的实现方式中，当C大于或等于2时，所述第一乘法累加窗口还包括设置于$PE_{X,Y+1}$与$PE_{X,Y}$之间的第一寄存器、设置于$PE_{X,Y+1}$与$PE_{X+1,Y}$之间的第二寄存器和设置于$PE_{X,Y+1}$与$PE_{X+1,Y+1}$之间的第三寄存器，所述方法还包括：所述第一寄存器寄存与传输$PE_{X,Y}$的卷积参数；所述第二寄存器寄存与传输$PE_{X+1,Y}$的卷积数据；所述第三寄存器寄存与传输卷积中间结果；其中，所述第一寄存器和所述第二寄存器在所述$PE_{X,Y+1}$在进行乘法运算时使得$PE_{X,Y+1}$的卷积数据和$PE_{X,Y+1}$的卷积参数节拍对齐，所述第三寄存器在所述第一乘法累加窗口进行加法运算时使得所述$PE_{X,Y+1}$传输的卷积中间结果与所述$PE_{X+1,Y+1}$进行乘法运算获得的乘积节拍对齐。

[0033] 在第三方面的一种可能的实现方式中，所述M×N乘法累加器阵列还包括第二乘法累加窗口，其中，所述第一乘法累加窗口和所述第二乘法累加窗口没有交集。

[0034] 在第三方面的一种可能的实现方式中，所述第一卷积数据矩阵与第二卷积数据矩阵相同，所述第二卷积数据矩阵为所述数据缓存模块向所述第二乘法累加窗口传输的卷积数据所属的卷积数据矩阵；所述第一卷积参数矩阵与第二卷积参数矩阵不同，所述第二卷积参数矩阵为所述数据缓存模块向所述第二乘法累加窗口传输的卷积参数所属的卷积参数矩阵。

[0035] 在第三方面的一种可能的实现方式中，所述第一卷积数据矩阵与第二卷积数据矩阵不同，所述第二卷积数据矩阵为所述数据缓存模块向所述第二乘法累加窗口传输的卷积数据所属的卷积数据矩阵；所述第一卷积参数矩阵与第二卷积参数矩阵相同，所述第二卷

积参数矩阵为所述数据缓存模块向所述第二乘法累加窗口传输的卷积参数所属的卷积参数矩阵。

[0036] 第四方面，提供了一种计算机可读介质，用于存储计算机程序，该计算机程序包括用于执行第三方面或第三方面的任意一种可能的实现方式所述的方法的指令。

附图说明

[0037] 图1为本发明实施例的通信设备的示意性框图。

[0038] 图2为本发明一个实施例的卷积运算芯片的示意性框图。

[0039] 图3为本发明一个实施例的乘法累加器阵列的示意性框图。

[0040] 图4为本发明一个实施例的卷积运算方法的运算过程原理示意图。

[0041] 图5为本发明另一个实施例的卷积运算方法的运算过程原理示意图。

[0042] 图6为本发明又一个实施例的卷积运算方法的运算过程原理示意图。

具体实施方式

[0043] 下面结合附图，对本发明的实施例进行描述。

[0044] 图1为本发明实施例的卷积运算芯片的应用场景图。在一种典型的通信设备中，例如片上系统(System on Chip，简称“SoC”)中，硬件架构包括中央处理器(Central Processing Unit，简称“CPU”)100，双倍速率同步动态随机存储器(Double Date Rate SDRAM，简称“DDR SDRAM”)200以及本发明实施例所述的卷积运算芯片300。CPU 100、DDR SDRAM 200和卷积运算芯片300通信连接。CPU 100控制卷积运算芯片300启动卷积运算，DDR SDRAM 200用于向卷积运算芯片300的数据缓存模块输入多个卷积数据和多个卷积参数，然后卷积运算芯片300根据获取的卷积数据和卷积参数完成卷积运算，得到运算结果，将运算结果写回DDR SDRAM 200约定的内存地址，通知CPU 100卷积运算完成。

[0045] 图2为本发明实施例所述的卷积运算芯片300的示意性框图。如图2所示，卷积运算芯片300包括：数据缓存模块310、M×N乘法累加器阵列320和输出控制模块330。

[0046] 数据缓存模块310用于向M×N乘法累加器阵列320中的第一乘法累加窗口传输用于卷积运算的多个卷积数据和多个卷积参数，其中，多个卷积参数由数据缓存模块310根据第一卷积参数矩阵确定，多个卷积数据由数据缓存模块310根据第一卷积数据矩阵确定，第一卷积参数矩阵为A行B列，第一卷积数据矩阵为D行E列，第一乘法累加窗口为A行C列，A为大于或等于2的整数，B和C均为大于或等于1的整数，D为大于或等于A的正整数，E为大于或等于max(B,C)的整数，M为大于或等于A的正整数，N为大于或等于C的正整数。

[0047] 图3为M×N乘法累加器阵列320的示意性框图。如图3所示，乘法累加器阵列包括M*N个处理单元，第u行第v列的处理单元标记为$PE_{u,v}$，u按照从小到大的顺序每次取一个整数，依次从1取值到M，对应于u的每一取值，v按照从小到大的顺序每次取一个整数，依次从1取值到N。对于某一行而言，例如第U行，U为大于或等于1且小于或等于M的整数，$PE_{U,\alpha}$与$PE_{U,\alpha+1}$之间存在横向的数据传输通道，α为大于或等于1且小于或等于N-1的任意整数。$PE_{\beta,\alpha}$与$PE_{\beta-1,\alpha+1}$之间存在斜向的数据传输通道，α为大于或等于1且小于或等于N-1的任意整数，β为大于或等于2且小于或等于M的任意整数。

[0048] M×N乘法累加器阵列320中的第一乘法累加窗口包括A×C个处理单元，第i行第j

列的处理单元标记为$PE_{i,j}$，i按照从小到大的顺序每次取一个整数，依次从1取值到A，对应于i的每一取值，j按照从小到大的顺序每次取一个整数，依次从1取值到C。第一乘法累加窗口的处理单元$PE_{X,Y}$用于将$PE_{X,Y}$的卷积数据和$PE_{X,Y}$的卷积参数进行乘法运算，当C大于或等于2时，处理单元$PE_{X,Y}$还用于将$PE_{X,Y}$的卷积参数传输至$PE_{X,Y+1}$，将$PE_{X,Y}$的卷积数据传输至$PE_{X-1,Y+1}$，分别作为$PE_{X,Y+1}$和$PE_{X-1,Y+1}$进行乘法运算的乘数，其中，X为大于或等于2且小于或等于A的整数，Y为大于或等于1且小于或等于C-1的整数，$PE_{X,Y}$的卷积数据为数据缓存模块310传输的多个卷积数据中的一个卷积数据，$PE_{X,Y}$的卷积参数为数据缓存模块310传输的多个卷积参数中的一个卷积参数。

[0049] 第一乘法累加窗口用于将$PE_{i,J}$进行乘法运算得到的乘积进行加法运算以获得卷积结果，其中，J为大于或等于1且小于或等于C的整数。

[0050] 输出控制模块330用于输出卷积结果。

[0051] 因此，本发明实施例的卷积运算芯片通过对任意一个处理单元增加一条数据传输通道，使得相邻处理单元之间能够直接传输卷积数据和卷积参数，同时，这些数据在传输过程中都处于第一乘法累加窗口中，不再经过RAM，可以减少RAM的访问次数，降低功耗。

[0052] 应理解，卷积运算芯片300还包括阵列控制模块340，用于从M×N乘法累加器阵列320中根据第一卷积参数矩阵确定用于卷积运算的第一乘法累加窗口，其中，根据第一卷积参数矩阵的行数确定第一乘法累加窗口的行数。具体地，第一卷积参数矩阵为A行B列，则选择用于卷积运算的第一乘法累加窗口的行数也为A，第一乘法累加窗口的列数C可以为大于或等于1且小于或等于N的正整数。

[0053] 在本发明实施例中，为尽可能地提高MAC阵列的利用率以及卷积运算效率，阵列控制模块320可以根据第一卷积参数矩阵的行数和第一卷积数据矩阵的行数确定第一乘法累加窗口的列数。具体而言，阵列控制模块320可以根据如下的公式(1)确定用于卷积运算的第一乘法累加窗口的列数C：

[0054] $C=D-A+1$ (1)。

[0055] C的上述取值是一种可选的方式，当然C也可以选取大于或小于公式(1)所计算得到的值，此时通过调整卷积矩阵和卷积参数的输入，仍然可以实现卷积运算，本发明实施例对此不作限定。

[0056] 因此，使用上述方案确定用于卷积运算的第一乘法累加窗口的行数和列数，使得乘法累加器中处理单元的使用与卷积参数矩阵的大小解耦，可以根据需要灵活调整第一乘法累加窗口的行数和列数，能够提高资源的利用率，从而提高设备的运算性能。

[0057] 例如，假设第一卷积参数矩阵为3行4列，则第一卷积窗口的行数确定为3行，同时，

[0058] 假设第一卷积数据矩阵为3行，则确定第一乘法累加窗口的列数C＝3-3+1＝1列；

[0059] 假设第一卷积数据矩阵为4行，则确定第一乘法累加窗口的列数C＝4-3+1＝2列；

[0060] 假设第一卷积数据矩阵为5行，则确定第一乘法累加窗口的列数C＝5-3+1＝3列；

[0061] 假设第一卷积数据矩阵为6行，则确定第一乘法累加窗口的列数C＝6-3+1＝4列；

[0062] …

[0063] 应理解，如果第一卷积数据矩阵尺寸较大，例如包括256行数据，使得按照以上公式计算出的C超出了乘法累加器的最大列数N，则可以将该256行数据分为多次计算。

[0064] 以乘法累加器阵列为8行8列为例进行说明。3行8列的第一乘法累加窗口可以同时

对10行卷积数据进行卷积，因此对于256行卷积数据可以做26次卷积运算，其中前25次每一次对10行卷积数据进行卷积，第26次对剩余的6行数据进行卷积。

[0065] 为了加快运算效率以及提高阵列使用率，可以同时激活多行处理单元作为乘法累加窗口，例如，可以在M行N列的乘法累加器阵列中选择激活连续的floor((M/A)*A行处理单元。在8行8列的乘法累加器阵列中，可以激活连续的floor((8/3)*3＝6行处理单元，其中3行8列乘法累加窗口可以用于做25次的卷积运算，该25次卷积运算每次针对不同的10行数据进行，剩余的3行4列乘法累加窗口可以对剩余的6行数据进行卷积运算；也可以是相邻的两个3行8列乘法累加窗口分工做该26次卷积运算，每一3行8列乘法累加窗口都可以做针对10行数据的卷积运算，以及都可以做针对剩余6行数据的卷积运算，本发明实施例对此不作限定。

[0066] 假设D×E卷积数据矩阵包括D*E个卷积数据$a_{p,q}$，其中，p按照从小到大的顺序每次取一个整数，依次从1取值到D，对应于p的每一取值，q按照从小到大的顺序每次取一个整数，依次从1取值到E。所述A×B卷积参数矩阵包括A*B个卷积参数$b_{p',q'}$，p'按照从小到大的顺序每次取一个整数，依次从1取值到A，对应于p'的每一取值，q'按照从小到大的顺序每次取一个整数，依次从1取值到B。数据缓存模块310内的缓存器用于将D*E个卷积数据和A*B个卷积参数进行缓存；在第nB+P时钟周期，数据缓存模块310中的计数器确定$PE_{i,1}$的卷积数据为$a_{i,n+P}$，$PE_{i,1}$的卷积参数为$b_{i,P}$，其中，P取值为大于或等于1且小于或等于B的整数，i为变量，i按照从小到大的顺序每次取一个整数，依次从1取值到A。该计数器在第nB+J'+Z-1时钟周期，确定$PE_{A,J'}$的卷积数据为$a_{A+J'-1,n+Z}$，$PE_{A,J'}$的卷积参数为$b_{A,Z}$，其中，Z取值为大于或等于1且小于或等于B的整数。

[0067] 第一乘法累加窗口具体执行以下步骤。第t时钟周期，第1列处理单元$PE_{i,1}$将$PE_{i,1}$的卷积数据和$PE_{i,1}$的卷积参数进行乘法运算获得乘积$X_{i,1}^{t}$，其中，$PE_{i,1}$的卷积数据和$PE_{i,1}$的卷积参数由数据缓存模块310传输至$PE_{i,1}$而获得；将$PE_{x,1}$的卷积参数传输至$PE_{x,2}$，将$PE_{x,1}$的卷积数据传输至$PE_{x-1,2}$，分别作为$PE_{x,2}$和$PE_{x-1,2}$在第t+1时钟周期进行乘法运算的乘数，x按照从小到大的顺序每次取一个整数，依次从2取值到A；在t分别取[nB+1,nB+B]区间内每一整数的情况下，将对应于t所有取值的所有所述乘积$X_{i,1}^{t}$利用如下公式(2)进行加法运算获得卷积结果S_1：

[0068] $$S_1=\sum_{t=nB+1}^{nB+B}\sum_{i=1}^{A}X_{i,1}^{t} \qquad (2)$$

[0069] 其中，i为变量，i按照从小到大的顺序每次取一个整数，依次从1取值到A，n为大于或等于0并且小于或等于(E-B)的整数。

[0070] 由公式(2)可知，加法运算是将t在[nB+1,nB+B]区间的每一整数取值下的A(由变量i变化获得)个$X_{i,1}^{t}$先进行求和，再将t作为变量得到的B个$X_{i,1}^{t}$进行求和。

[0071] 具体地，第一卷积数据矩阵为：

[0072] $$\begin{pmatrix} a_{1,1} & \cdots & a_{1,E} \\ \vdots & \ddots & \vdots \\ a_{D,1} & \cdots & a_{D,E} \end{pmatrix}$$

[0073] 第一卷积参数矩阵为：

[0074] $$\begin{pmatrix} b_{1,1} & \cdots & b_{1,B} \\ \vdots & \ddots & \vdots \\ b_{A,1} & \cdots & b_{A,B} \end{pmatrix}$$

[0075] n＝0时，在第1至B时钟周期，数据缓存模块310向处理单元$PE_{1,1}$传输的卷积数据分别为$a_{1,1}$，$a_{1,2}$，$a_{1,3}$，…，$a_{1,B}$，向处理单元$PE_{1,1}$传输的卷积参数分别为$b_{1,1}$，$b_{1,2}$，$b_{1,3}$，…，$b_{1,B}$；数据缓存模块310向处理单元$PE_{2,1}$传输的卷积数据分别为$a_{2,1}$，$a_{2,2}$，$a_{2,3}$，…，$a_{2,B}$，向处理单元$PE_{2,1}$传输的卷积参数分别为$b_{2,1}$，$b_{2,2}$，$b_{2,3}$，…，$b_{2,B}$；

[0076] …

[0077] 数据缓存模块310向处理单元$PE_{A,1}$传输的卷积数据分别为$a_{A,1}$，$a_{A,2}$，$a_{A,3}$，…，$a_{A,B}$，向处理单元$PE_{A,1}$传输的卷积参数分别为$b_{A,1}$，$b_{A,2}$，$b_{A,3}$，…，$b_{A,B}$。

[0078] n＝1时，在第B+1至2B时钟周期，数据缓存模块310向处理单元$PE_{1,1}$传输的卷积数据分别为$a_{1,2}$，$a_{1,3}$，$a_{1,4}$，…，$a_{1,B+1}$，向处理单元$PE_{1,1}$传输的卷积参数分别为$b_{1,1}$，$b_{1,2}$，$b_{1,3}$，…，$b_{1,B}$；数据缓存模块310向处理单元$PE_{2,1}$传输的卷积数据分别为$a_{2,2}$，$a_{2,3}$，$a_{2,4}$，…，$a_{2,B+1}$，向处理单元$PE_{2,1}$传输的卷积参数分别为$b_{2,1}$，$b_{2,2}$，$b_{2,3}$，…，$b_{2,B}$；

[0079] …

[0080] 数据缓存模块310向处理单元$PE_{A,1}$传输的卷积数据分别为$a_{A,2}$，$a_{A,3}$，$a_{A,4}$，…，$a_{A,B+1}$，向处理单元$PE_{A,1}$传输的卷积参数分别为$b_{A,1}$，$b_{A,2}$，$b_{A,3}$，…，$b_{A,B}$；

[0081] …

[0082] 对于除了第一列以外的其它列的处理单元而言，第T时钟周期，第J’列处理单元$PE_{i,J'}$将$PE_{i,J'}$的卷积数据和$PE_{i,J'}$的卷积参数进行乘法运算获得乘积$X_{i,J'}^{T}$，其中，i按照从小到大的顺序每次取一个整数，依次从1取值到A，J’为大于或等于2且小于或等于C的整数，$PE_{i,J'}$的卷积参数由$PE_{i,J'-1}$的卷积参数传输至$PE_{i,J'}$而获得，$PE_{h,J'}$的卷积数据由$PE_{h+1,J'-1}$的卷积数据传输至$PE_{h,J'}$而获得，$PE_{A,J'}$的卷积参数和$PE_{A,J'}$的卷积数据由数据缓存模块310传输至$PE_{A,J'}$而获得，h按照从小到大的顺序每次取一个整数，依次从1取值到A-1。具体地，计数器在第nB+J’+Z-1时钟周期，确定$PE_{A,J'}$的卷积数据为$a_{A+J'-1,n+Z}$，$PE_{A,J'}$的卷积参数为$b_{A,Z}$，其中，Z取值为大于或等于1且小于或等于B的整数。

[0083] 在T分别取[nB+J’，nB+J’+B-1]区间内每一整数的情况下，将对应于T所有取值的所有所述乘积$X_{i,J'}^{T}$，通过如下公式(3)进行加法运算获得卷积结果$S_{J'}$：

[0084] $$S_{J'}=\sum_{T=nB+J'}^{nB+J'+B-1}\sum_{i=1}^{A}X_{i,J'}^{T} \qquad (3)$$

[0085] 由公式(3)可知，加法运算是将T在[nB+J’，nB+J’+B-1]区间的每一整数取值下的A(由变量i变化获得)个$X_{i,J'}^{T}$先进行求和，再将T作为变量得到的B个$X_{i,1}^{t}$进行求和。

[0086] 具体地，第T时钟周期，处理单元$PE_{1,2}$中的卷积参数由上一时钟周期内$PE_{1,1}$的卷积参数传输至$PE_{1,2}$而获得，处理单元$PE_{1,2}$中的卷积数据由上一时钟周期内处理单元$PE_{2,1}$的卷积数据传输而来；处理单元$PE_{2,2}$中的卷积参数由上一时钟周期内$PE_{2,1}$的卷积参数传输至$PE_{2,2}$而获得，处理单元$PE_{2,2}$中的卷积数据由上一时钟周期内处理单元$PE_{3,1}$的卷积数据传输而来；

[0087] …

[0088] 处理单元$PE_{A-1,2}$中的卷积参数由上一时钟周期内$PE_{A-1,1}$的卷积参数传输至$PE_{A-1,2}$而获得，处理单元$PE_{A-1,2}$中的卷积数据由上一时钟周期内处理单元$PE_{A,1}$的卷积数据传输而获得；处理单元$PE_{A,2}$的卷积参数和卷积数据由数据缓存模块310传输而获得。

[0089] 处理单元$PE_{1,J'}$中的卷积参数由上一时钟周期内$PE_{1,J'-1}$的卷积参数传输至$PE_{1,J'}$而获得，处理单元$PE_{1,J'}$中的卷积数据由上一时钟周期内处理单元$PE_{2,J'-1}$的卷积数据传输而获得；处理单元$PE_{2,J'}$中的卷积参数由上一时钟周期内$PE_{2,J'-1}$的卷积参数传输至$PE_{2,J'}$而获得，处理单元$PE_{2,J'}$中的卷积数据由上一时钟周期内处理单元$PE_{3,J'-1}$的卷积数据传输而来；

[0090] …

[0091] 处理单元$PE_{A-1,J'}$中的卷积参数由上一时钟周期内$PE_{A-1,J'-1}$的卷积参数传输至$PE_{A-1,J'}$而获得，处理单元$PE_{A-1,J'}$中的卷积数据由上一时钟周期内处理单元$PE_{A,J'-1}$的卷积数据传输而获得；处理单元$PE_{A,J'}$的卷积参数和卷积数据由数据缓存模块310传输而获得。

[0092] 对于第一乘法累加窗口中的任意一列处理单元，例如，第J列处理单元，在第nB+J时钟周期，将乘积$X_{1,J}^{nB+J}$传输至$PE_{2,J}$，与乘积$X_{2,J}^{nB+J}$进行加法运算，获得卷积中间结果Q_1^{nB+J}，其中，乘积$X_{1,J}^{nB+J}$为$PE_{1,J}$在第nB+J时钟周期将$PE_{1,J}$的卷积数据和$PE_{1,J}$的卷积参数进行乘法运算获得的乘积，乘积$X_{2,J}^{nB+J}$为$PE_{2,J}$在第nB+J时钟周期将$PE_{2,J}$的卷积数据和$PE_{2,J}$的卷积参数进行乘法运算获得的乘积；将$PE_{f,J}$进行加法运算得到的卷积中间结果Q_{f-1}^{nB+J}传输至$PE_{f+1,J}$，其中，f按照从小到大的顺序每次取一个整数，依次从2取值到A-1；将卷积中间结果Q_{f-1}^{nB+J}与$PE_{f+1,J}$进行乘法运算获得的乘积$X_{f+1,J}^{nB+J}$进行加法运算，获得卷积中间结果Q_f^{nB+J}；将在$PE_{A,J}$内获得的卷积中间结果Q_{A-1}^{nB+J}传输给输出控制模块330用于缓存；在第nB+J+1时钟周期向$PE_{1,J}$传输卷积中间结果Q_{A-1}^{nB+J}，作为在第nB+J+1时钟周期进行加法运算的累加初始值。

[0093] 将第(n+1)B+J-1时钟周期获得的卷积中间结果Q_{A-1}^{nB+J+1}确定为第J卷积结果S_J。

[0094] 图4至图6为本发明实施例的的卷积运算芯片进行卷积运算的过程原理图。假设第一卷积数据矩阵为如下矩阵：

[0095] $$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 & 25 \\ 26 & 27 & 28 & 29 & 30 & 31 & 32 & 33 & 34 & 35 & 36 & 37 & 38 \\ 39 & 40 & 41 & 42 & 43 & 44 & 45 & 46 & 47 & 48 & 49 & 50 & 51 \\ 52 & 53 & 54 & 55 & 56 & 57 & 58 & 59 & 60 & 61 & 62 & 63 & 64 \end{bmatrix},$$

[0096] 第一卷积参数矩阵为如下矩阵：

[0097] $$\begin{bmatrix} k11 & k12 & k13 \\ k21 & k22 & k23 \\ k31 & k32 & k33 \end{bmatrix},$$

[0098] M×N乘法累加器阵列320为8行8列的平铺阵列，即M等于8，N等于8。可以看出，第一卷积数据矩阵为一个5行13列的矩阵，第一卷积参数矩阵为一个3行3列的矩阵，在此例中，A＝3，B＝3，D＝5，E＝13，M＝8，N＝8。

[0099] 可以知道，3行3列的第一卷积参数矩阵与5行13列的第一卷积数据矩阵经过卷积运算后，得到的卷积结果矩阵的尺寸为3行11列。

[0100] 阵列控制模块340根据第一卷积数据矩阵和第一参数矩阵的尺寸确定第一乘法累加窗口的尺寸，具体如下：

[0101] 如果第一卷积参数矩阵的尺寸为3行3列，则将会采用连续的3行处理单元作为卷积运算的处理单元，即确定第一乘法累加窗口的行数为A＝3行；同时，第一卷积数据矩阵为5行13列，则根据公式C＝D-A+1，确定第一乘法累加窗口的列数为C＝3列。因此，在8行8列的乘法累加器阵列中，第一乘法累加窗口为3行3列。

[0102] 应理解，3×3第一乘法累加窗口在8×8乘法累加器320中的位置不受限制，可以位于阵列的边缘，也可以位于阵列的中间，本发明实施例对此不作限制。

[0103] 第一卷积参数矩阵和第一卷积数据矩阵分别输入数据缓存模块310后，数据缓存模块310即对上述矩阵进行缓存和并输入第一乘法累加窗口。

[0104] 数据缓存模块310将第一卷积参数矩阵中的卷积参数按照时钟周期的顺序进行排列，即：

[0105] 数据缓存模块310将第一卷积参数矩阵中的第一行卷积参数按照时钟周期的顺序排列为k11,k12,k13,k11,k12,k13,k11…

[0106] 数据缓存模块310将第一卷积参数矩阵中的第一行卷积参数按照时钟周期的顺序排列为k21,k22,k23,k21,k22,k23,k21…

[0107] 数据缓存模块310将第一卷积参数矩阵中的第一行卷积参数按照时钟周期的顺序排列为k31,k32,k33,k31,k32,k33,k31…

[0108] 也就是说，在第一时钟周期，第一乘法累加窗口的第一列处理单元中的卷积参数从第一行到第三行分别为k11,k21和k31；在第二时钟周期，第一乘法累加窗口的第一列处理单元中的卷积参数从第一行到第三行分别为k12,k22和k32；在第三时钟周期，第一乘法累加窗口的第一列处理单元中的卷积参数从第一行到第三行分别为k13,k23和k33；

[0109] …

[0110] 数据缓存模块310将第一卷积数据矩阵中的卷积数据按照时钟周期的顺序进行排

列,即:

[0111] 数据缓存模块310将第一卷积数据矩阵中的第一行卷积数据按照时钟周期的顺序排列为0,1,2,1,2,3,2,3,4,3,4…

[0112] 数据缓存模块310将第一卷积数据矩阵中的第二行卷积数据按照时钟周期的顺序排列为13,14,15,14,15,16,15,16,17,16,17…

[0113] 数据缓存模块310将第一卷积数据矩阵中的第三行卷积数据按照时钟周期的顺序排列为26,27,28,27,28,29,28,29,30,29,30…

[0114] 数据缓存模块310将第一卷积数据矩阵中的第四行卷积数据按照时钟周期的顺序排列为39,40,41,40,41,42,41,42,43,42,43…

[0115] 数据缓存模块310将第一卷积数据矩阵中的第五行卷积数据按照时钟周期的顺序排列为52,53,54,53,54,55,54,55,56,56,57…

[0116] 也就是说,在第一时钟周期,第一乘法累加窗口的第一列处理单元中的卷积数据从第一行到第三行分别为0,13和26;在第二时钟周期,第一乘法累加窗口的第一列处理单元中的卷积数据从第一行到第三行分别为1,14和27;在第三时钟周期,第一乘法累加窗口的第一列处理单元中的卷积参数从第一行到第三行分别为2,15和28。

[0117] 图4为发明实施例的卷积运算芯片在第一时钟周期的运算过程原理图。如图4所示,在第一时钟周期,将第一列处理单元$PE_{i,1}$(其中,i的分别取值1,2,3)中的卷积数据和卷积参数分别进行乘法运算,得到的乘法结果分别为

$X^1_{1,1}$=0*k11, $X^1_{2,1}$=13*k21, $X^1_{3,1}$=26*k31;其中,$X^1_{1,1}$、$X^1_{2,1}$和$X^1_{3,1}$分别为处理单元$PE_{1,1}$、$PE_{2,1}$和$PE_{3,1}$在第一时钟周期内的乘法结果;

[0118] 将$X^1_{1,1}$通过数据通道传输至处理单元$PE_{2,1}$,与$X^1_{2,1}$相加得到卷积中间结果

$Q^1_{1(1)}$=0*k11+13*k21;

[0119] 将得到的卷积中间结果Q^1_1通过数据通道传输到处理单元$PE_{3,1}$,与$X^1_{3,1}$相加得到卷积中间结果$Q^1_{2(1)}$=0*k11+13*k21+26*k31;并将卷积中间结果$Q^1_{2(1)}$传输给输出控制模块330用于缓存;

[0120] 第一列处理单元$PE_{i,1}$(其中,i的分别取值1,2,3)中的卷积数据和卷积参数在第一时钟周期参与卷积运算后,分别沿着不同的数据通道传输到其它的处理单元。第一列处理单元$PE_{i,1}$(其中,i的分别取值1,2,3)中的卷积参数分别通过数据通道传输到第二列处理单元$PE_{i,2}$(其中,i的分别取值1,2,3)的对应位置,即$PE_{1,1}$中的卷积参数k11传输至$PE_{1,2}$,$PE_{2,1}$中的卷积参数k21传输至$PE_{2,2}$,$PE_{3,1}$中的卷积参数k31传输至$PE_{3,2}$分别作为$PE_{1,2}$、$PE_{2,2}$和$PE_{3,2}$在下一时钟周期内进行卷积运算的乘数。同时将第一列处理单元$PE_{i,1}$(其中,i的分别取值1,2,3)中的卷积数据分别通过不同的数据通道传输到第二列处理单元的对应位置,即$PE_{2,1}$中的卷积数据13传输至$PE_{1,2}$,$PE_{3,1}$中的卷积数据26传输至$PE_{2,2}$,数据缓存模块通过数据通道向$PE_{3,2}$传输卷积数据39,分别作为$PE_{1,2}$、$PE_{2,2}$和$PE_{3,2}$在下一时钟周期进行卷积运算的另一个乘数;

[0121] 应理解,处理单元$PE_{1,1}$的卷积数据0被传输至第一乘法累加窗口以外,不再参与后

续卷积运算。

[0122] 图5为本发明实施例的的卷积运算芯片在第二时钟周期的运算过程原理图。如图5所示，在第二时钟周期，在第一列处理单元$PE_{i,1}$(其中，i的分别取值1,2,3)中将数据缓存模块310输入的卷积数据和卷积参数按照上述方法进行乘法运算，得到乘法结果$X_{1,1}^{2}$、$X_{2,1}^{2}$和$X_{3,1}^{2}$分别为1*k12、14*k22和27*k32；

[0123] 将输出控制模块330中缓存的卷积中间结果Q_{2}^{1}在此时钟周期内被传输到处理单元$PE_{1,1}$中，与$X_{1,1}^{2}$相加得到卷积中间结果$Q_{1(1)}^{2}$=0*k11+13*k21+26*k31+1*k12;

[0124] 同理，将第一列卷积中间结果Q_{1}^{2}传输至$PE_{2,1}$，与$X_{2,1}^{2}$相加，可以得到卷积中间结果$Q_{2(1)}^{2}$；将卷积中间结果$Q_{2(1)}^{2}$传输至$PE_{3,1}$，与$X_{3,1}^{2}$相加，可以得到卷积中间结果$Q_{3(1)}^{2}$=0*k11+13*k21+26*k31+1*k12+14*k22+27*k32，将卷积中间结果$Q_{3(1)}^{2}$传输至输出控制模块用于缓存；

[0125] 同时，第二列处理单元$PE_{i,2}$(其中，i的分别取值1,2,3)在第二时钟周期开始进行卷积运算，过程与第一列处理单元的运算过程类似，在此不作详述，在此时钟周期内第二列处理单元$PE_{i,2}$(其中，i的分别取值1，2，3)输出卷积中间结果$Q_{2(2)}^{2}$=13*k11+26*k21+39*k31，传输至输出控制模块330用于缓存；

[0126] 第二列处理单元$PE_{i,2}$(其中，i的分别取值1,2,3)中的卷积数据和卷积参数在参与卷积运算后，分别沿着不同的数据通道传输到其它的处理单元。第二列处理单元$PE_{i,2}$(其中，i的分别取值1,2,3)中的卷积参数分别通过数据通道传输到第三列处理单元$PE_{i,3}$(其中，i的分别取值1,2,3)的对应位置，即$PE_{1,2}$中的卷积参数k11传输至$PE_{1,3}$，$PE_{2,2}$中的卷积参数k21传输至$PE_{2,3}$，$PE_{3,2}$中的卷积参数k31传输至$PE_{3,3}$分别作为$PE_{1,3}$、$PE_{2,3}$和$PE_{3,3}$在下一时钟周期内进行卷积运算的乘数。同时将第二列处理单元$PE_{i,2}$(其中，i的分别取值1,2,3)中的卷积数据分别通过不同的数据通道传输到第三列处理单元的对应位置，即$PE_{2,2}$中的卷积数据26传输至$PE_{1,3}$，$PE_{3,2}$中的卷积数据39传输至$PE_{2,3}$，数据缓存模块310通过数据通道向$PE_{3,3}$传输卷积数据52，分别作为$PE_{1,3}$、$PE_{2,3}$和$PE_{3,3}$在下一时钟周期进行卷积运算的另一个乘数。

[0127] 应理解，处理单元$PE_{1,2}$的卷积数据13被传输至第一乘法累加窗口以外，不再参与后续卷积运算。

[0128] 同时，第一列处理单元$PE_{i,1}$(其中，i的分别取值1,2,3)中的卷积数据和卷积参数在第二时钟周期参与卷积运算后，也分别沿着不同的数据通道按照前述的类似方式传输到第二列的处理单元，在此不作详述。

[0129] 图6为本发明实施例的卷积运算芯片在第三时钟周期的运算过程原理图。如图6所示，在第三时钟周期，在第一列处理单元$PE_{i,1}$(其中，i的分别取值1,2,3)中将数据缓存模块310输入的卷积数据和卷积参数按照上述方法进行乘法运算，得到乘法结果$X_{1,1}^{3}$、$X_{2,1}^{3}$和

$X_{3,1}^{3}$分别为2*k13、15*k23和28*k33；将输出控制模块330中缓存的卷积中间结果$Q_{3(1)}^{2}$传输至处理单元$PE_{1,1}$中，与$X_{1,1}^{3}$相加得到卷积中间结果

$Q_{1(1)}^{3}$=0*k11+13*k21+26*k31+1*k12+14*k22+27*k32+2*k13；将第一列卷积中间结果$Q_{1(1)}^{3}$传输至$PE_{2,1}$，与$X_{2,1}^{3}$相加，可以得到卷积中间结果$Q_{2(1)}^{3}$；将卷积中间结果$Q_{2(1)}^{3}$传输至$PE_{3,1}$，与$X_{3,1}^{3}$相加，可以得到卷积中间结果

$Q_{3(1)}^{3}$=0*k11+13*k21+26*k31+1*k12+14*k22+27*k32+2*k13+15*k23+28*k33。

[0130] 在第三时钟周期第一列处理单元经过卷积运算输出的卷积中间结果$Q_{3(1)}^{3}$即为第一卷积结果，该第一卷积结果输出后作为卷积结果矩阵的第一行第一列的元素。

[0131] 同理，第二列处理单元$PE_{i,2}$(其中，i的分别取值1,2,3)在第三时钟周期接收从第一列处理单元$PE_{i,1}$(其中，i的分别取值1,2,3)传输的卷积参数为k12,k22和k32，第二列处理单元$PE_{i,2}$(i＝1,2)在第三时钟周期接收从第一列处理单元$PE_{i,1}$(i＝2,3)传输的卷积数据14和27，第二列处理单元$PE_{3,2}$从数据缓存模块接收卷积数据40，经过类似运算过程得到卷积中间结果$Q_{3(2)}^{3}$=13*k11+26*k21+39*k31+14*k12+27*k22+40*k32；第三列处理单元$PE_{i,3}$(i＝1,2,3)在第三时钟周期开始进行卷积运算，第三列处理单元$PE_{i,3}$(其中，i的分别取值1,2,3)在第三时钟周期接收从第二列处理单元$PE_{i,2}$(其中，i的分别取值1,2,3)传输的卷积数据k11,k21和k31，第三列处理单元$PE_{i,3}$(i＝1,2)在第三时钟周期接收从第二列处理单元$PE_{i,2}$(i＝2,3)传输的卷积数据26和39，第三列处理单元$PE_{3,3}$从数据缓存模块310接收卷积数据52，经过类似运算过程输出的卷积中间结果为$Q_{2(3)}^{3}$=26*k11+39*k21+52*k31。

[0132] 类似地，在第四时钟周期，第一列处理单元从数据缓存模块310接收的卷积参数从$PE_{1,1}$到$PE_{3,1}$分别为k11,k21和k31，接收的卷积数据从$PE_{1,1}$到$PE_{3,1}$分别为1,14和27，进行卷积运算后输出卷积中间结果$Q_{2(1)}^{4}$；

[0133] 第二列处理单元$PE_{i,2}$(i＝1,2)接收从第一列处理单元$PE_{i,1}$(i＝2,3)传输的卷积数据15和28，第二列处理单元$PE_{3,2}$从数据缓存模块310传输的卷积数据41，第二列处理单元$PE_{i,2}$(其中，i的分别取值1,2,3)接收从第一列处理单元$PE_{i,1}$(i＝1,2,3)传输的卷积参数为k13,k23和k33，经过类似的卷积运算过程后输出卷积中间结果

[0134]

$Q_{3(2)}^{4}$=13*k11+26*k21+39*k31+14*k12+27*k22+40*k32+15*k13+28*k23+41*k33。

[0135] 在第四时钟周期第二列处理单元经过卷积运算输出的卷积中间结果$Q_{3(2)}^{4}$即为第二卷积结果，该第二卷积结果输出后作为卷积结果矩阵的第二行第一列的元素。

[0136] 第三列处理单元$PE_{i,3}$(其中,i的分别取值1,2,3)在第三时钟周期接收从第二列处理单元$PE_{i,2}$(其中,i的分别取值1,2,3)传输的卷积参数为k12,k22和k32,第三列处理单元$PE_{i,2}$(i=1,2)在第四时钟周期接收从第二列处理单元$PE_{i,1}$(i=2,3)传输的卷积数据27和40,第三列处理单元$PE_{3,3}$从数据缓存模块310接收卷积数据53,经过类似运算过程得到卷积中间结果$Q^4_{3(3)}$=26*k11+39*k21+52*k31+27*k12+40*k22+53*k32。

[0137] 类似地,在第五时钟周期,第一列处理单元从数据缓存模块310接收的卷积参数从$PE_{1,1}$到$PE_{3,1}$分别为k12,k22和k32,接收的卷积数据从$PE_{1,1}$到$PE_{3,1}$分别为2,15和28,进行卷积运算后输出卷积中间结果$Q^5_{2(1)}$;

[0138] 第二列处理单元$PE_{i,2}$(i=1,2)接收从第一列处理单元$PE_{i,1}$(i=2,3)传输的卷积数据14和27,第二列处理单元$PE_{3,2}$从数据缓存模块310传输的卷积数据40,第二列处理单元$PE_{i,2}$(其中,i的分别取值1,2,3)接收从第一列处理单元$PE_{i,1}$(其中,i的分别取值1,2,3)传输的卷积参数为k11,k21和k31,经过类似的卷积运算过程后输出卷积中间结果$Q^4_{3(2)}$=14*k11+27*k21+40*k31。

[0139] 第三列处理单元$PE_{i,3}$(其中,i的分别取值1,2,3)在第五时钟周期接收从第二列处理单元$PE_{i,2}$(i=1,2,3)传输的卷积参数为k13,k23和k33,第三列处理单元$PE_{i,2}$(i=1,2)在第五时钟周期接收从第二列处理单元$PE_{i,1}$(i=2,3)传输的卷积数据28和41,第三列处理单元$PE_{3,3}$从数据缓存模块310接收卷积数据54,经过类似运算过程得到卷积中间结果

[0140]

$$Q^5_{3(3)}=26*k11+39*k21+52*k31+27*k12+40*k22+53*k32+28*k13+41*k23+54*k33。$$

[0141] 在第五时钟周期第三列处理单元经过卷积运算输出的卷积中间结果$Q^5_{3(3)}$即为第三卷积结果,该第三卷积结果输出后作为卷积结果矩阵的第三行第一列的元素。

[0142] 因此,第一列处理单元在第一至第三时钟周期内经过卷积运算输出第一卷积结果,作为卷积结果矩阵第一行第一列的元素;第二列处理单元在第二至第四时钟周期内经过卷积运算输出第二卷积结果,作为卷积结果矩阵第二行第一列的元素;第三列处理单元在第三至第五时钟周期内经过卷积运算输出第三卷积结果,作为卷积结果矩阵第三行第一列的元素。

[0143] 本发明实施例的卷积运算芯片中通过对任意一个处理单元增加一条数据传输通道,使得相邻处理单元之间能够直接传输卷积数据和卷积参数,同时,这些数据在传输过程中都处于第一乘法累加窗口中,没有经过RAM,可以减少RAM的访问次数,降低功耗。

[0144] 应理解,卷积结果还可以通过其它的方式得到。例如,在第一列处理单元中,在第一时钟周期分别通过乘法运算得到$X^1_{1,1}$、$X^1_{2,1}$和$X^1_{3,1}$,将上述乘法结果传输到输出控制模块当中用于缓存;在第二时钟周期分别通过乘法运算得到$X^2_{1,1}$、$X^2_{2,1}$和$X^2_{3,1}$,将上述乘法结果传输到输出控制模块当中用于缓存;在第三时钟周期分别通过乘法运算得到

$X_{1,1}^{3}$、$X_{2,1}^{3}$和$X_{3,1}^{3}$，设置在第三时钟周期在处理单元$PE_{3,1}$内进行加法运算，将$\sum_{T=1}^{3}\sum_{i=1}^{3}X_{i,1}^{T}$作为卷积结果输出；再例如，在第一列处理单元中，在第一时钟周期分别通过乘法运算得到$X_{1,1}^{1}$、$X_{2,1}^{1}$和$X_{3,1}^{1}$，将上述乘法结果传输到输出控制模块当中用于缓存；在第二时钟周期分别通过乘法运算得到$X_{1,1}^{2}$、$X_{2,1}^{2}$和$X_{3,1}^{2}$，将输出控制模块中缓存的$X_{1,1}^{1}$、$X_{2,1}^{1}$和$X_{3,1}^{1}$分别传输到$PE_{1,1}$、$PE_{2,1}$和$PE_{3,1}$中，与$X_{1,1}^{2}$、$X_{2,1}^{2}$和$X_{3,1}^{2}$分别进行加法运算，在$PE_{1,1}$内得到卷积中间结果$X_{1,1}^{1}+X_{1,1}^{2}$，在$PE_{2,1}$得到卷积中间结果$X_{2,1}^{1}+X_{2,1}^{2}$，在$PE_{3,1}$内得到卷积中间结果$X_{3,1}^{1}+X_{3,1}^{2}$，分别将上述卷积中间结果传输到输出控制模块当中用于缓存；在第三时钟周期分别通过乘法运算得到$X_{1,1}^{3}$、$X_{2,1}^{3}$和$X_{3,1}^{3}$，将输出控制模块中缓存的卷积中间结果分别传输到$PE_{1,1}$、$PE_{2,1}$和$PE_{3,1}$中，与$X_{1,1}^{3}$、$X_{2,1}^{3}$和$X_{3,1}^{3}$分别相加，在$PE_{1,1}$内得到卷积中间结果$X_{1,1}^{1}+X_{1,1}^{2}+X_{1,1}^{3}$，在$PE_{2,1}$得到卷积中间结果$X_{2,1}^{1}+X_{2,1}^{2}+X_{2,1}^{3}$，在$PE_{3,1}$内得到卷积中间结果$X_{3,1}^{1}+X_{3,1}^{2}+X_{3,1}^{3}$，将第三时钟周期得到的卷积中间结果设置在$PE_{3,1}$内进行加法运算，将最终加法结果$\sum_{T=1}^{3}\sum_{i=1}^{3}X_{i,1}^{T}$作为卷积结果输出。

[0145] 应理解，在第一乘法累加窗口中的任意两个处理单元之间的数据通道上可以设置寄存器。寄存器可以分为第一寄存器，第二寄存器和第三寄存器。第一寄存器设置于处理单元$PE_{X,Y+1}$与处理单元$PE_{X,Y}$之间，用于处理单元$PE_{X,Y}$的卷积参数的寄存与传输；第二寄存器设置于处理单元$PE_{X,Y+1}$与处理单元$PE_{X+1,Y}$之间，用于处理单元$PE_{X+1,Y}$的卷积数据的寄存与传输；第三寄存器设置于处理单元$PE_{X,Y+1}$与处理单元$PE_{X+1,Y+1}$之间，用于卷积中间结果的寄存与传输，其中，X为大于或等于2且小于或等于A的整数，Y为大于或等于1且小于或等于C-1的整数。

[0146] 具体地，处理单元$PE_{X+1,Y}$中的卷积参数和卷积数据在一个时钟周期内进行卷积运算后，该处理单元$PE_{X+1,Y}$(最后一列处理单元除外)的卷积参数通过数据通道传输到第一寄存器中寄存，用于参与下一个时钟周期处理单元$PE_{X+1,Y+1}$的卷积运算；处理单元$PE_{X+1,Y}$(第一行或最后一列处理单元除外)的卷积数据通过数据通道传输到第二寄存器中寄存，用于参与下一个时钟周期处理单元$PE_{X,Y+1}$的卷积运算。同理，处理单元$PE_{X,Y}$中的卷积参数和卷积数据在一个时钟周期内进行卷积运算后，该处理单元$PE_{X,Y}$(最后一列处理单元除外)的卷积参数通过数据通道传输到第一寄存器中寄存，用于参与下一个时钟周期处理单元$PE_{X,Y+1}$的卷积运算；处理单元$PE_{X,Y}$(第一行或最后一列处理单元除外)的卷积数据通过数据通道传输到第二寄存器中寄存，用于参与下一个时钟周期处理单元$PE_{X-1,Y+1}$的卷积运算。在处理单元$PE_{X,Y}$进行乘法运算完成后，将乘法结果或卷积中间结果寄存在处理单元$PE_{X,Y}$与处理单元$PE_{X+1,Y}$之间的第三寄存器中，使得在处理单元$PE_{X,Y}$中得到的乘法结果或卷积中间结果与处理单元$PE_{X+1,Y}$的乘法结果能够节拍对齐。

[0147] 也就是说，处理单元$PE_{X,Y+1}$与处理单元$PE_{X,Y}$之间的第一寄存器用于寄存处理单元

$PE_{X,Y}$传输过来的卷积参数，处理单元$PE_{X,Y+1}$与处理单元$PE_{X+1,Y}$之间的第二寄存器用于寄存处理单元$PE_{X+1,Y}$传输过来的卷积数据，第一寄存器和第二寄存器还用于在处理单元$PE_{X,Y+1}$进行乘法运算时使处理单元$PE_{X,Y}$传输过来的卷积参数和处理单元$PE_{X+1,Y}$传输过来的卷积数据节拍对齐，确保在下一个时钟周期内二者能够进行卷积运算。

[0148] 因此，通过在处理单元之间的数据通道上设置寄存器，使得对于任意一个处理单元，在进行卷积运算时卷积数据和卷积参数能够节拍对齐，确保了卷积运算的顺利进行；同时，由于除第一列处理单元外的处理单元的卷积数据和卷积参数都是通过相邻处理单元之间的数据通道上传输，不需要占用外部总线来传输，所以能够降低外部传输的带宽。

[0149] 应理解，在M行N列的乘法累加器阵列中，除了包括第一乘法累加窗口外，还可以同时包括第二乘法累加窗口，且第一乘法累加窗口和第二乘法累加窗口之间没有共同的处理单元，也就是说在乘法累加器阵列中可以同时进行多组不同的卷积运算。

[0150] 当第一卷积数据矩阵与第二卷积数据矩阵相同，第一卷积参数矩阵与第二卷积参数矩阵不同，其中，第二卷积数据矩阵为数据缓存模块310向所述第二乘法累加窗口传输的卷积数据所属的卷积数据矩阵，第二卷积参数矩阵为数据缓存模块310向第二乘法累加窗口传输的卷积参数所属的卷积参数矩阵。例如，第二卷积参数矩阵为：

[0151] $$\begin{bmatrix} r11 & r12 & r13 & r14 \\ r21 & r22 & r23 & r24 \end{bmatrix}$$

[0152] 第一卷积数据矩阵为5行13列，第二卷积参数矩阵的元素个数、元素值与第一卷积参数矩阵的元素个数、元素值均不相同。第二卷积参数矩阵与第二卷积数据矩阵在乘法累加器中进行卷积运算时，确定的第二乘法累加窗口为2行4列，得到的卷积结果矩阵为4行9列。

[0153] 当第一卷积数据矩阵与第二卷积数据矩阵不同，第一卷积参数矩阵与第二卷积参数矩阵相同时，其中，第二卷积数据矩阵为数据缓存模块310向第二乘法累加窗口传输的卷积数据所属的卷积数据矩阵，第二卷积参数矩阵为数据缓存模块310向第二乘法累加窗口传输的卷积参数所属的卷积参数矩阵。例如，第二卷积数据矩阵为：

[0154] $$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 & 25 \\ 26 & 27 & 28 & 29 & 30 & 31 & 32 & 33 & 34 & 35 & 36 & 37 & 38 \\ 39 & 40 & 41 & 42 & 43 & 44 & 45 & 46 & 47 & 48 & 49 & 50 & 51 \\ 52 & 53 & 54 & 55 & 56 & 57 & 58 & 59 & 60 & 61 & 62 & 63 & 64 \\ 65 & 66 & 67 & 68 & 69 & 70 & 71 & 72 & 73 & 74 & 75 & 76 & 77 \end{bmatrix}$$

[0155] 第一卷积参数矩阵为3行3列，第二卷积数据矩阵中的元素个数、元素值与第一卷积数据矩阵中的元素个数、元素值均不相同。第二卷积参数矩阵与第二卷积数据矩阵在乘法累加器中进行卷积运算时，确定的第二乘法累加窗口为3行4列，得到的卷积结果矩阵为4行11列。

[0156] 应理解，在M行N列的乘法累加器阵列中可以同时包括多个第一乘法累加窗口和多个第二乘法累加窗口，本发明在此不作限定。

[0157] 当然，在M行N列的乘法累加器阵列中还可以同时进行卷积数据矩阵和卷积参数矩

阵均不相同的卷积运算，用于卷积运算的乘法累加窗口均独立于第一乘法累加窗口和第二乘法累加窗口。

[0158] 因此，通过灵活设置乘法累加窗口的尺寸，使得在同一乘法累加器阵列中可以同时进行多种不同的卷积运算，提升了阵列的利用率。

[0159] 应理解，在实际产品型SoC中，通常包括4片本发明实施例的乘法累加器阵列，每个乘法累加器阵列为15行14列，其中，4片乘法累加器阵列既可以相互关联也可以相互独立。

[0160] 应理解，在本发明实施例中，“与A相应的B”表示B与A相关联，根据A可以确定B。但还应理解，根据A确定B并不意味着仅仅根据A确定B，还可以根据A和/或其它信息确定B。

[0161] 另外，本文中术语“系统”和“网络”在本文中常被可互换使用。本文中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。

[0162] 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中描述的各方法步骤和单元，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各实施例的步骤及组成。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。本领域普通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发明的范围。

[0163] 结合本文中所公开的实施例描述的方法或步骤可以用硬件、处理器执行的软件程序，或者二者的结合来实施。软件程序可以置于随机存取存储器(RAM)、内存、只读存储器(ROM)、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

[0164] 尽管通过参考附图并结合优选实施例的方式对本发明进行了详细描述，但本发明并不限于此。在不脱离本发明的精神和实质的前提下，本领域普通技术人员可以对本发明的实施例进行各种等效的修改或替换，而这些修改或替换都应在本发明的涵盖范围内。

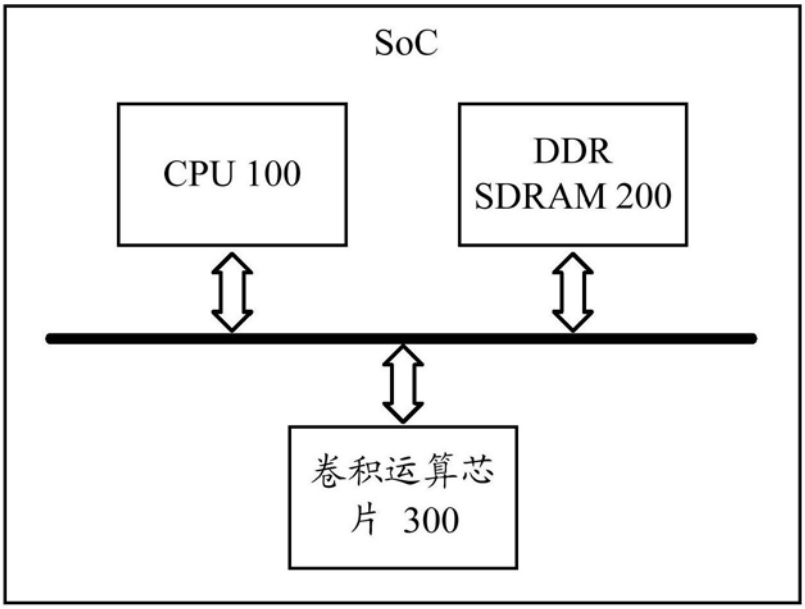

SoC
CPU 100
DDR SDRAM 200
卷积运算芯片 300

图1

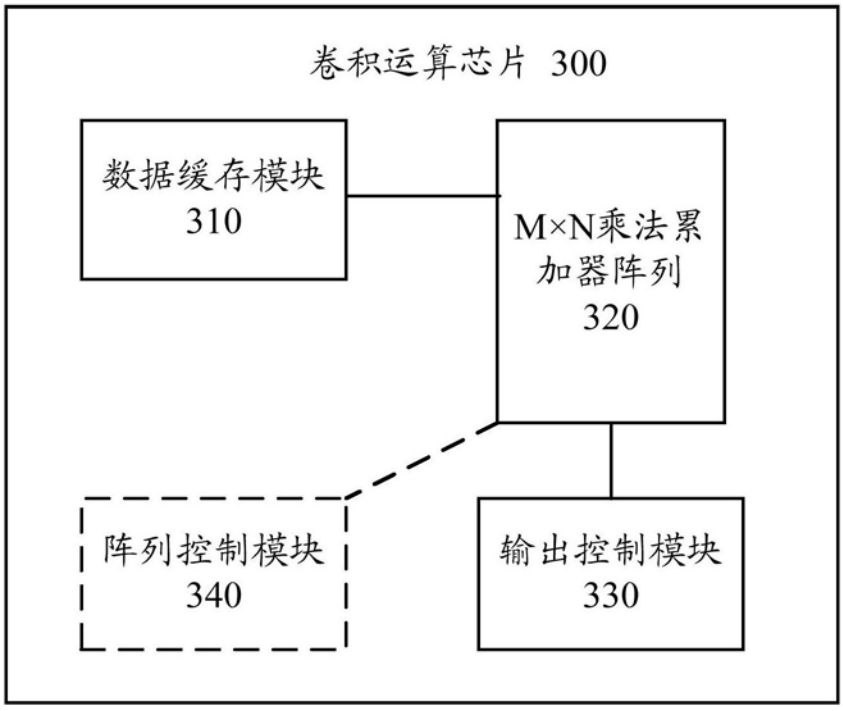

卷积运算芯片 300
数据缓存模块 310
M×N乘法累加器阵列 320
阵列控制模块 340
输出控制模块 330

图2

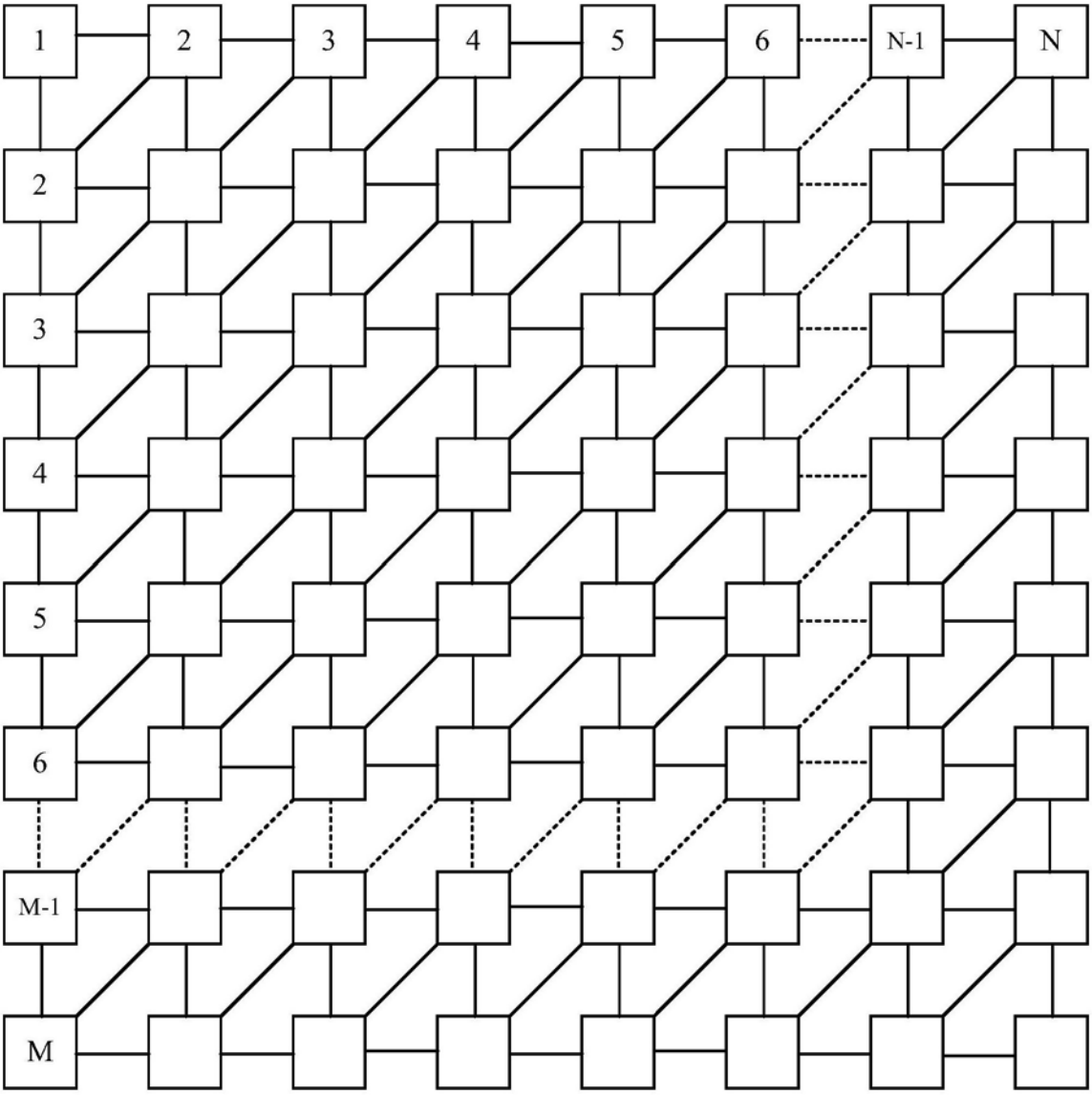


图3

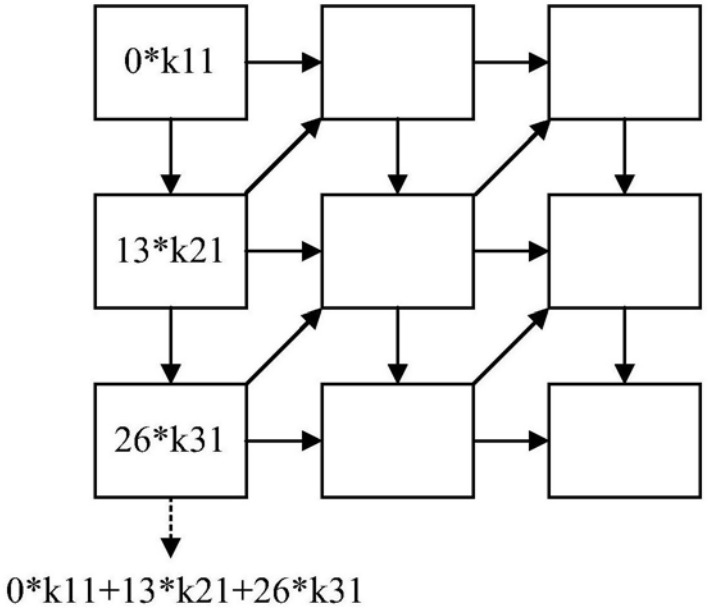

0*k11
13*k21
26*k31
0*k11+13*k21+26*k31

图4

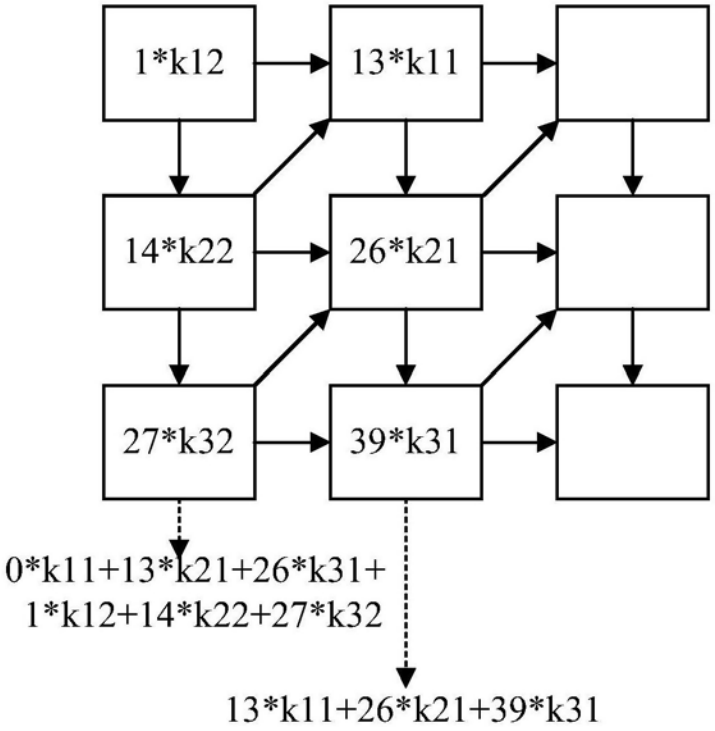

1*k12
13*k11
14*k22
26*k21
27*k32
39*k31
0*k11+13*k21+26*k31+
1*k12+14*k22+27*k32
13*k11+26*k21+39*k31

图5

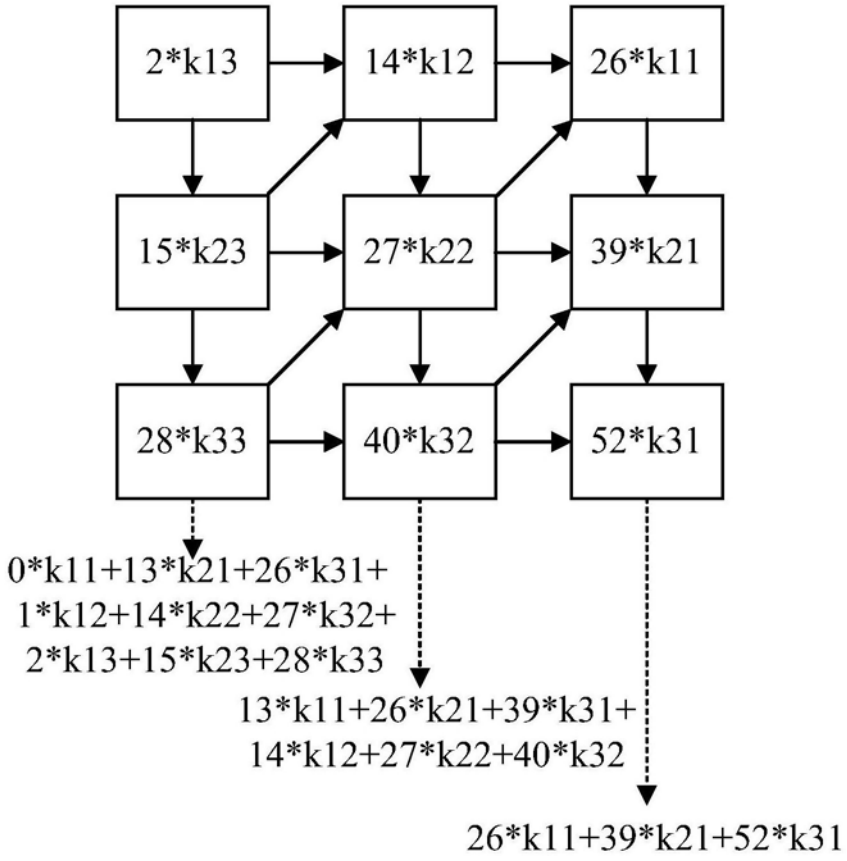

2*k13
14*k12
26*k11
15*k23
27*k22
39*k21
28*k33
40*k32
52*k31
0*k11+13*k21+26*k31+
1*k12+14*k22+27*k32+
2*k13+15*k23+28*k33
13*k11+26*k21+39*k31+
14*k12+27*k22+40*k32
26*k11+39*k21+52*k31

图6