(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110991627 B
(45)授权公告日 2025.01.07

(21)申请号 201910933454.7

(22)申请日 2019.09.29

(65)同一申请的已公布的文献号
申请公布号 CN 110991627 A

(43)申请公布日 2020.04.10

(30)优先权数据
2018-188612 2018.10.03 JP

(73)专利权人 佳能株式会社
地址 日本东京都大田区下丸子3-30-2

(72)发明人 陈则玮

(74)专利代理机构 北京怡丰知识产权代理有限公司 11293
专利代理师 迟军

(51)Int.Cl.
G06N 3/063(2023.01)

(56)对比文件
US 2017116495 A1,2017.04.27

审查员 郭彬瑜

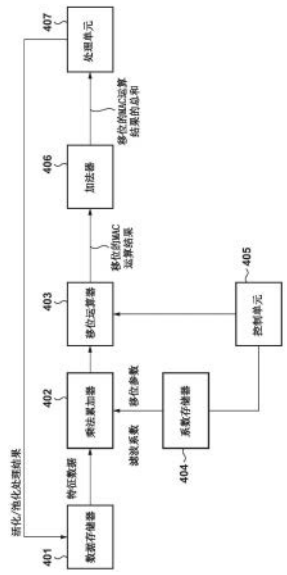
权利要求书2页 说明书12页 附图14页

(54)发明名称

信息处理装置、信息处理方法

(57)摘要

本发明提供信息处理装置、信息处理方法。该信息处理装置包括:控制单元,其被构造为,针对包括多个层的网络的各层,基于数据的位宽设置移位量;多个MAC(乘法累加)单元,其被构造为,对层的多个数据和多个滤波系数执行乘法累加运算;多个移位运算单元,其被构造为,基于移位量对通过所述多个MAC单元获得的多个MAC运算结果进行移位;以及加法单元,其被构造为,计算通过所述多个移位运算单元移位的所述多个MAC运算结果的总和。



CN 110991627 B

1.一种信息处理装置,其包括:

多个乘法累加单元,其被构造为,对包括多个层的网络的层的多个数据和多个滤波系数执行乘法累加运算;

多个移位运算单元,其被构造为,基于移位量通过对各个乘法累加运算结果求2的幂来对通过所述多个乘法累加单元获得的多个乘法累加运算结果进行移位;

控制单元,其被构造为,针对网络的各个层,基于输入数据中的数字和输入数据的位宽来设置与多个移位运算单元相对应的移位量;以及

加法单元,其被构造为,计算通过所述多个移位运算单元移位的所述多个乘法累加运算结果的总和,

其中,所述多个乘法累加单元的数量与所述多个移位运算单元的数量相对应。

2.根据权利要求1所述的信息处理装置,其中,所述多个乘法累加单元在同一电路上执行所述多个层中的各个的乘法累加运算。

3.根据权利要求2所述的信息处理装置,其中,所述多个乘法累加单元在同一电路上依次执行所述多个层中的各个的乘法累加运算。

4.根据权利要求3所述的信息处理装置,其中,所述控制单元针对要处理的各层设置移位量。

5.根据权利要求1所述的信息处理装置,其中,所述多个层包括处理具有彼此不同位宽的数据集的两个或更多个层。

6.根据权利要求1所述的信息处理装置,其中,所述控制单元根据层的数据的位宽,来切换滤波系数的传输计数。

7.根据权利要求1所述的信息处理装置,其中,所述移位运算单元基于多个移位量对所述多个乘法累加运算结果进行移位。

8.根据权利要求1所述的信息处理装置,其中,所述多个乘法累加单元和所述多个移位运算单元并行操作。

9.根据权利要求1所述的信息处理装置,其中,所述加法单元将预定层的总和作为所述预定层之后的层的数据,存储在存储器中。

10.根据权利要求1所述的信息处理装置,所述信息处理装置还包括:

活化单元,其被构造为对总和进行活化处理。

11.根据权利要求10所述的信息处理装置,所述信息处理装置还包括:

池化单元,其被构造为对活化处理的结果进行池化处理。

12.根据权利要求11所述的信息处理装置,其中,池化单元将通过对预定层进行池化处理而获得的结果作为所述预定层之后的层的数据,存储在存储器中。

13.根据权利要求9所述的信息处理装置,所述信息处理装置还包括如下单元,该单元被构造为,基于存储在存储器中的数据,对运动图像的各帧执行图像处理和图像识别中的至少一个。

14.根据权利要求1所述的信息处理装置,其中,所述网络针对各层具有不同的数据位宽。

15.根据权利要求1所述的信息处理装置,其中,数据的位宽是2位、4位或8位。

16.根据权利要求1所述的信息处理装置,其中,在滤波器大小等于1×1的情况下,乘法

累加单元计算数据与滤波系数的乘积。

17.一种信息处理装置,其包括:

多个移位运算单元,其被构造为,基于移位量通过对包括多个层的网络的层的多个数据的各个数据求2的幂来对所述多个数据进行移位;

控制单元,其被构造为,针对所述网络的各个层,基于输入数据中的数字和输入数据的位宽来设置与多个移位运算单元相对应的移位量;

多个乘法累加单元,其被构造为,对多个滤波系数和通过所述多个移位运算单元移位的所述多个数据执行乘法累加运算;以及

加法单元,其被构造为,计算通过所述多个乘法累加单元计算的多个乘法累加运算结果的总和,

其中,所述多个乘法累加单元的数量与所述多个移位运算单元的数量相对应。

18.一种信息处理方法,其包括:

对包括多个层的网络的层的多个数据和多个滤波系数执行乘法累加运算;

基于移位量通过对各个乘法累加运算结果求2的幂来对所获得的多个乘法累加运算结果进行移位;

针对网络的各个层,基于输入数据中的数字和输入数据的位宽来设置与移位相对应的移位量;以及

计算移位的所述多个乘法累加运算结果的总和,

其中,执行乘法累加运算的单元的数量与对多个乘法累加运算结果进行移位的单元的数量相对应。

19.一种信息处理方法,其包括:

基于移位量通过对包括多个层的网络的层的多个数据的各个数据求2的幂来对所述多个数据进行移位;

针对所述网络的各个层,基于输入数据中的数字和输入数据的位宽来设置与移位相对应的移位量;

对多个滤波系数和移位的所述多个数据执行乘法累加运算;以及

对计算出的多个乘法累加运算结果的总和进行计算,

其中,执行乘法累加运算的单元的数量与对多个乘法累加运算结果进行移位的单元的数量相对应。

20.一种非暂时性计算机可读存储介质,其存储使计算机用作权利要求1至17中的任一项中定义的信息处理装置的各单元的程序。

信息处理装置、信息处理方法

技术领域

[0001] 本发明涉及信息处理装置和信息处理方法,尤其涉及具有多个层的网络中的算术技术。

背景技术

[0002] 近年来,由于深度学习的发展,图像识别的精度正在提高。卷积神经网络(CNN)作为深度学习中使用的一种方法为人所知。在CNN中,多个层被分层连接,并且在各个层中包括多个特征图像。图2示出了如下网络的示例,其中存在四个层(层1-4),并且各个层中存在四个特征图像。在图2中,特征图像(i,j)表示层i中的第j个特征图像。通过使用学习的滤波系数和特征图像像素(特征数据)来计算滤波器处理结果。滤波处理是乘法累加(MAC)运算,并且包括多个乘法和累加和。图2所示的各个箭头表示MAC运算。

[0003] 通过使用前层的特征图像和与前层相对应的滤波系数来计算当前层的特征图像。为了计算当前层的一个特征图像,需要先前层的多个特征图像的信息。用于计算当前层的各个特征图像的卷积运算的等式如下。

[0004] $$O_{i,j}(n)=\sum_{m=1}^{M}\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}\left(I_{i+x,j+y}(m)\times C_{x,y}(m,n)\right) \quad (1)$$

[0005] 其中$O_{i,j}(n)$是表示与当前层中的第n个特征图像中的位置(i,j)相对应的MAC运算结果的变量。在等式(1)中,在先前层中有M个特征图像,并且$I_{i,j}(m)$表示在第m个特征图像中的位置(i,j)处的特征数据。存在X×Y个滤波系数$C_{1,1}(m,n)$至$C_{X,Y}(m,n)$,并且各个特征图像的滤波系数不同。用于计算当前层中第n个特征图像的MAC运算被进行M×X×Y次。在执行卷积运算之后,通过使用MAC运算结果$O_{i,j}(n)$执行诸如活化和池化的处理,来计算当前层的特征图像。

[0006] 由于CNN需要大量的MAC运算,因此当CNN应用于诸如移动终端、车载设备等嵌入式系统时,需要高效率的数据并行处理装置。由于减小处理数据的位宽将降低计算卷积运算结果的算术运算单元的成本,因此可以增大算术运算单元的并行度(DOP)。在Y.Li,et al.,A 7.663-TOPS 8.2-W Energy-efficient FPGA Accelerator for Binary Convolutional Neural Networks,Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays,Pages 290-291,Feb.2017中提出了一种用于处理各层具有不同数据位宽的网络的硬件布置。

[0007] 在Y.Li,et al.,A 7.663-TOPS 8.2-W Energy-efficient FPGA Accelerator for Binary Convolutional Neural Networks,Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays,Pages 290-291,Feb.2017中描述的方法中,使用不同种类的算术运算符来处理各层具有不同位宽的CNN。在输入层的特征数据的位宽为8位并且中间层的特征数据的位宽为2位的情况下,需要专用于8位数据的卷积运算单元和专用于2位数据的卷积运算单元。

[0008] 尽管可以通过流水线处理8位数据层和2位数据层以并行处理数据,但是当卷积处理的计算量对于各个层不同时,硬件使用效率会降低。另外,当要处理具有在2位与8位之间的位宽(例如4位)的特征数据时,由于将不得不使用专用于8位数据的卷积运算单元(这是因为不存在针对该位宽的卷积运算单元),因此效率将降低。

[0009] 在K.Lee,et al.,A 502-GOPS and 0.984-mW Dual-Mode Intelligent ADAS SoC With Real-Time Semiglobal Matching and Intention Prediction for Smart Automotive Black Box System,IEEE Journal of Solid-State Circuits,Vol.52,No.1,Pages 139-150,Jan.2017中描述的方法中,提出了一种RNN(递归神经网络)专用硬件,该硬件具有SIMD(单指令多数据)构造,该构造能够处理具有多个位宽的特征数据集。尽管可以通过使用相同的硬件来处理8位数据、16位数据和32位数据,但这将在计算并行输出的数据的总和时增加处理时间,这是因为在数据临时保持在存储器中后需要再次执行SIMD命令。

发明内容

[0010] 本发明提供即使在多层网络中存在具有多个位宽的数据集的情况下也实现高效率的处理的技术。

[0011] 根据本发明的第一方面,提供了一种信息处理装置,其包括:控制单元,其被构造为,针对包括多个层的网络的各层,基于数据的位宽设置移位量;多个MAC(乘法累加)单元,其被构造为,对层的多个数据和多个滤波系数执行乘法累加运算;多个移位运算单元,其被构造为,基于移位量对通过所述多个乘法累加单元获得的多个乘法累加运算结果进行移位;以及加法单元,其被构造为,计算通过所述多个移位运算单元移位的所述多个乘法累加运算结果的总和。

[0012] 根据本发明的第二方面,提供了一种信息处理装置,其包括:控制单元,其被构造为,针对包括多个层的网络的各层,基于数据的位宽设置移位量;多个移位运算单元,其被构造为,基于移位量对层的多个数据进行移位;多个乘法累加单元,其被构造为,对多个滤波系数和通过所述多个移位运算单元移位的所述多个数据执行乘法累加运算;以及加法单元,其被构造为,计算通过所述多个乘法累加单元计算的多个乘法累加运算结果的总和。

[0013] 根据本发明的第三方面,提供了一种信息处理方法,其包括:针对包括多个层的网络的各层,基于数据的位宽设置移位量;对层的多个数据和多个滤波系数执行乘法累加运算;基于移位量对多个乘法累加运算结果进行移位;以及计算移位的所述多个乘法累加运算结果的总和。

[0014] 根据本发明的第四方面,提供了一种信息处理方法,其包括:针对包括多个层的网络的各层,基于数据的位宽设置移位量;基于移位量对层的多个数据进行移位;对多个滤波系数和移位的所述多个数据执行乘法累加运算;以及对计算出的多个乘法累加运算结果的总和进行计算。

[0015] 通过以下参照附图对示例性实施例的描述,本发明的其他特征将变得清楚。

附图说明

[0016] 图1是数据处理的流程图;

[0017] 图2是示出处理对象网络的布置的示例的图;

[0018] 图3是示出信息处理装置的硬件布置的示例的框图;

[0019] 图4是示出数据处理单元305的布置的示例的框图;

[0020] 图5是示出乘法累加器402和移位运算器403的布置的示例的框图;

[0021] 图6A至图6C是示出数据与处理时间之间的关系的表;

[0022] 图7是示出数据处理单元305的布置的示例的框图;

[0023] 图8是示出乘法累加器702和移位运算器701的布置的示例的框图;

[0024] 图9是数据处理的流程图;并且

[0025] 图10A至图10C是分别示出8位特征数据、2位特征数据和4位特征数据的示例的图。

具体实施方式

[0026] 现在将参照附图描述本发明的实施例。请注意,下面要描述的各实施例是本发明的详细实施的示例,并且是所附权利要求中描述的布置的详细实施例。

[0027] [第一实施例]

[0028] 首先将参照图3的框图描述根据实施例的信息处理装置的硬件构造的示例。诸如PC(个人计算机)、平板终端设备、智能电话等的计算机装置可应用为信息处理装置。另外,该信息处理装置可以是要被嵌入在这样的设备中的嵌入式设备。

[0029] 输入单元301由诸如键盘、鼠标、触摸面板等的用户接口形成,并且当由用户操作时可以将各种指令输入至CPU 306。

[0030] 数据存储单元302是诸如硬盘驱动设备等的大容量信息存储设备。数据存储单元302存储要在信息处理装置中使用的各种信息(例如OS(操作系统))、由CPU 306执行的各种计算机程序、在CPU 306执行各种处理时要使用的数据等。数据存储单元302中存储的数据包括将由图像处理单元309处理的图像。注意,以下将被描述为“已知信息”的信息也存储在数据存储单元302中。通过CPU 306、数据处理单元305和图像处理单元309将存储在数据存储单元302中的计算机程序和数据加载到RAM 308等中,并且这些计算机程序和数据成为CPU 306、数据处理单元305和图像处理单元309的处理对象。

[0031] 注意,数据存储单元302可以是存储介质(例如,软盘、CD-ROM、CD-R、DVD、存储卡、CF卡、智能介质、SD卡、记忆棒、xD图片卡、USB存储器等)。在这种情况下,信息处理装置需要包括从这样的存储介质读出信息并将信息写入该存储介质的设备。

[0032] 通信单元303用作用于与外部设备进行数据通信的通信接口。可以进行设置,使得通信单元303将从外部设备获得执行信息处理装置中的处理所需的信息。通信单元303可以将由信息处理装置进行的处理的结果发送到外部设备。

[0033] 显示单元304由液晶屏或触摸面板屏形成,并且可以显示图像、字符等,以显示由CPU 306、数据处理单元305或图像处理单元309获得的处理结果。注意,显示单元304可以是诸如投影仪的投影设备。输入单元301和显示单元304可以被集成并形成既具有指令输入接受功能又具有显示功能的设备,例如触摸屏设备。

[0034] 数据处理单元305通过使用由图像处理单元309写入RAM 308中的图像,通过执行根据图1的流程图的处理来执行CNN计算,并且将获得的计算结果输出到数据存储单元302、RAM 308等。注意,将成为数据处理单元305的处理对象的图像不限于由图像处理单元309写入RAM308的图像,并且例如可以是由其他装置输入的图像。稍后将参照图4描述数据处理单

元305。

[0035] CPU 306通过使用存储在ROM 307或RAM 308中的计算机程序和数据来执行各种处理。这允许CPU 306控制信息处理装置的整体操作。

[0036] ROM 307存储不需要重写的信息,例如信息处理装置的设置数据、激活程序等。RAM 308包括用于存储从数据存储单元302和ROM 307加载的计算机程序和数据、由通信单元303从外部设备接收的信息等的区域。RAM 308包括由CPU 306、数据处理单元305和图像处理单元309使用以执行各种处理的工作区域。RAM 308可以以此方式适当地提供各种区域。

[0037] 图像处理单元309在根据来自CPU 306的指令对图像的各个像素执行像素值范围调整之后,读出存储在数据存储单元302中的图像,并将该图像写入RAM 308。

[0038] 上述输入单元301、数据存储单元302、通信单元303、显示单元304、数据处理单元305、CPU 306、ROM 307、RAM 308和图像处理单元309都连接到总线310。

[0039] 注意,信息处理装置的硬件布置不限于图3所示的布置。例如,图3的布置可以由多个装置实现。另外,信息处理装置无需总是包括诸如输入单元301、数据存储单元302和显示单元304的设备,并且可以进行设置,使得这些设备将经由通信路径连接到信息处理装置。

[0040] 另外,被描述为存储在RAM 308中的一些或所有信息可以被存储在数据存储单元302中,并且被描述为存储在数据存储单元302中的一些或所有信息可以被存储在RAM 308中。另选地,可以进行设置,使得将RAM 308的一部分用作数据存储单元302,或者可以以虚拟方式布置,使得通信单元303的通信对方(partner)设备的存储设备经由通信单元303用作数据存储单元。

[0041] 此外,尽管在图3中仅示出了一个CPU 306,但是信息处理装置中包括的CPU 306的数量不限于一个,并且信息处理装置中可以包括多个CPU。另外,数据处理单元305和图像处理单元309可以被实现为硬件或可以被实现为计算机程序。在后者的情况下,这些计算机程序将被存储在数据存储单元302中,并且数据处理单元305和图像处理单元309的功能将通过CPU 306执行相应的计算机程序来执行。

[0042] 注意,基于数据处理单元305的处理结果,CPU 306将对从通信单元303或数据存储单元302获得的运动图像的各帧进行图像处理和/或图像识别。通过CPU 306进行的图像处理或图像识别的结果被存储在RAM308或数据存储单元302中,或者经由通信单元303输出到外部设备。此外,通过CPU 306进行的图像处理或图像识别的结果可以作为图像或字符而被显示在显示单元304上,或者在信息处理装置具有音频输出功能的情况下作为音频被输出。

[0043] <处理对象网络>

[0044] 本实施例使用CNN作为处理对象网络。图2示出了处理对象网络的布置的示例。图2的处理对象网络的细节如上所述。注意,处理对象网络的信息(诸如各MAC运算的计算量、各个特征图像的大小、特征图像的数量、各个特征图像的位数等)被存储在数据存储单元302等中。

[0045] 图2所示的处理对象网络中的层数为四(层1-4),并且各层中有四个特征图像。如上所述,特征图像(i,j)表示层i的第j个特征图像。而且,层的各个特征图像的位宽根据层而变化。层1的各个特征图像的位宽是8位,层2的各个特征图像的位宽是2位,层3的各个特征图像的位宽是4位,并且层4的各个特征图像的位宽是8位。由于第一层(层1)和最后层(层4)保持输入/输出图像信息,因此第一层和最后层的位宽(8位)倾向于大于中间层(层2和层

3)的位宽(2位和4位)。各个特征图像由多个像素(特征数据)形成。

[0046] 以下将描述由数据处理单元305执行的层1至层4中的各个层的特征图像的计算(生成)。通过使用滤波系数和层1的8位特征图像(1,1)、(1,2)、(1,3)和(1,4)执行根据上述等式(1)的MAC运算。随后,作为MAC运算的结果,生成层2的2位特征图像(2,1)、(2,2)、(2,3)和(2,4)。

[0047] 接下来,使用层2的2位特征图像(2,1)、(2,2)、(2,3)和(2,4)以及滤波系数来进行根据上述等式(1)的MAC运算。随后,作为MAC运算的结果,生成层3的4位特征图像(3,1)、(3,2)、(3,3)和(3,4)。

[0048] 接下来,使用层3的4位特征图像(3,1)、(3,2)、(3,3)和(3,4)和滤波系数来进行根据上述等式(1)的MAC运算。随后,作为MAC运算的结果,生成层4的8位特征图像(4,1)、(4,2)、(4,3)和(4,4)。

[0049] <数据处理单元305的布置示例>

[0050] 数据处理单元305的布置的示例在图4中示出。数据存储器401保持各层的各个特征图像的特征数据,并且系数存储器404保持滤波系数。乘法累加器402通过使用系数存储器404中保持的滤波系数和数据存储器401中保持的特征数据执行MAC运算,来计算各个MAC运算结果。移位运算器403对通过乘法累加器402获得的各个MAC运算结果进行移位,加法器406通过将多个移位的MAC运算结果相加来获得“移位的MAC运算结果的总和”。基于加法器406获得的“移位的MAC运算结果的总和”,处理单元407计算活化/池化处理结果,并将计算出的活化/池化处理结果存储在数据存储器401中。控制单元405控制数据处理单元305的整体操作。

[0051] 将根据图1的流程图描述数据处理单元305进行的数据处理。在步骤S101中,控制单元405从RAM 308中读出多个输入特征图像的特征数据和滤波系数,将读出的特征数据存储在数据存储器401中,并且将读出的特征系数存储在系数存储器404中。

[0052] 在步骤S102中,控制单元405开始各层的循环,并将未处理的层中的一个层设置为处理对象层。由于在该示例中将层1至4依次设置为处理对象,因此,层1将首先作为处理对象层。

[0053] 在步骤S103中,控制单元405根据层信息来设置移位运算器403的定义移位量的移位参数。在步骤S104中,控制单元405开始输出特征图像的循环,并依次计算输出特征数据。在步骤S105中,控制单元405初始化存储在加法器406中的MAC运算结果以将MAC运算结果设置为零。加法器406具有总和计算功能。

[0054] 在步骤S106中,控制单元405开始输入特征图像的循环并依次处理输入特征数据。在步骤S107中,在控制单元405的控制下,乘法累加器402和移位运算器403分别进行上述MAC运算和移位运算。稍后将描述步骤S107的处理的细节(步骤S115至步骤S117)。

[0055] 在步骤S108中,加法器406将多个MAC运算结果相加以获得“移位的MAC运算结果的总和”。在步骤S109中,控制单元405确定输入特征图像循环的完成。如果所有的输入特征图像的处理已经完成,则处理进入步骤S110。否则,处理返回到步骤S107,并且开始下一个未处理的输入特征图像的处理。在步骤S110中,处理单元407根据下式,基于由加法器406获得的“移位的MAC运算结果的总和”,来计算活化处理结果。

[0056] $$f(x)=\begin{cases}0, & x<0\\ x, & x\geq 0\end{cases} \quad (2)$$

[0057] 在这种情况下,f(•)是活化函数,x是输入数据。尽管在该示例中使用ReLU(整流线性单元)来实现活化函数,但是本发明不限于ReLU,并且活化函数可以由其他非线性函数或量化函数来实现。注意,活化处理结果的位宽将根据需要进行调整。

[0058] 在步骤S111中,处理单元407根据层信息,通过基于活化处理结果执行池化处理来计算活化/池化处理结果。在步骤S112中,处理单元407将在步骤S111中计算出的活化/池化处理结果作为下一层的特征图像存储在数据存储器401中。

[0059] 在步骤S113中,控制单元405确定输出特征图像的循环的完成。如果所有的输出特征图像的处理已经完成,则处理进入步骤S114。否则,处理返回到步骤S105以开始处理未处理的输出特征图像。

[0060] 在步骤S114中,控制单元405确定层的循环的完成。如果所有层的处理已经完成,则根据图1的流程图的处理结束。否则,处理返回到步骤S103,并且开始处理未处理的层。

[0061] <MAC运算和移位运算>

[0062] 将描述步骤S107的MAC运算和移位运算(步骤S115至S117)。在步骤S115中,除了从数据存储器401中读出特征数据并将所读出的特征数据传输至乘法累加器402之外,控制单元405还从系数存储器404中读出滤波系数,并将读出的滤波系数传输至乘法累加器402。滤波系数的数量和传输计数将根据特征数据的位宽而变化。

[0063] 在步骤S116中,乘法累加器402基于特征数据和滤波系数来计算MAC运算结果。在步骤S117中,移位运算器403基于在步骤S103中设置的移位参数所指示的移位量来将在步骤S116中获得的MAC运算结果进行移位。

[0064] <具有不同位宽的情况的详细描述>

[0065] 该实施例可以处理不同位宽的数据。图10A示出了在要处理8位特征数据的情况下乘法累加器402、移位运算器403和加法器406的操作。图10B示出了在要处理2位特征数据的情况下乘法累加器402、移位运算器403和加法器406的操作。

[0066] 在特征数据是8位的情况下,乘法累加器402将8位特征数据1001(值:234)划分为2位数据集(2位数据),如图10A所示。乘法累加器402使用根据该划分而获得的四个2位数据集(值:2、2、2、3)和共享滤波系数来计算四个MAC运算结果,并且移位运算器403基于四个移位参数对四个MAC运算结果进行移位。随后,加法器406将四个移位的MAC运算结果相加以计算一个特征数据集(8位输入特征数据的MAC运算结果)。数据处理单元305可以以这种方式处理一个8位输入特征数据集。

[0067] 在特征数据是2位的情况下,乘法累加器402使用四个2位数据集1002(值:2、2、2、3)和四个滤波系数来计算四个MAC运算结果,如图10B所示。移位运算器403基于一个移位参数来将四个MAC运算结果进行移位。由于移位参数为零,因此在移位运算之前和之后,MAC运算结果的状态相同。随后,加法器406通过将四个MAC运算结果相加来计算一个特征数据集(四个2位输入特征数据集的MAC运算结果的总和)。数据处理单元305可以以此方式并行处理四个2位输入特征数据集。

[0068] 令M为输入特征图像的数量,并且1×1为滤波器大小。由于滤波器大小等于一个像素,并且变量x和y的值是常数,因此将使用$I_{i,j}(n)$计算$O_{i,j}(n)$。MAC运算的计算(等式1)可

以简化为下式。

[0069] $$O(n)=\sum_{m=1}^{M}(I(m)\times C(m,n)) \quad (3)$$

[0070] 尽管在滤波器大小大于1×1的情况下,乘法累加器402将计算滤波系数和输入特征数据的各个卷积结果,但是在滤波器大小等于1×1的情况下,乘法累加器402将计算I(m)和C(m,n)的乘积。

[0071] 假设存在两种特征处理对象数据,即位宽为α位的特征处理数据和位宽为β位的特征处理数据。图4所示的乘法累加器402包括用于计算MAC运算结果的P个α位数据MAC运算单元,而移位运算器403包括用于计算移位运算结果的P个α位数据移位运算单元。α、β和P满足以下条件。

[0072] $\beta=\alpha\times P$ (4)

[0073] 在输入特征数据$I'_{(\beta)}$的位宽是β位的情况下,基于等式(6)、(7)和(8)的前提,加法器406的输出由下面的等式(5)表示。第n个输出图像的MAC运算结果O(n)由下式给出。

[0074] $$O(n)=\sum_{m=1}^{M}\sum_{p=1}^{P}\left[\left(I_{(\alpha),p}(m)\times C_p(m,n)\right)\times 2^{S(p)}\right] \quad (5)$$

[0075] 其中,$I_{(\alpha),P}(m)$是α位数据MAC运算单元的输入数据,$C_p(m,n)$是滤波系数,S(p)是移位参数。变量m是α位输入特征图像组(1组=P个图像)的编号(乘法累加器402的处理编号),变量p是MAC运算单元编号和移位运算单元编号,变量n是输出特征图像编号。移位运算由2的幂处理表示。

[0076] 如等式(6)所示,滤波系数$C_p(m,n)$是与第m个β位特征图像对应的滤波系数C'(m,n)。由于共享滤波系数用于α位输入特征图像组,因此变量p可以省略。并行提供给P个MAC运算单元的滤波系数的数量为1,传输计数为1。

[0077] $$C_p(m,n)=\acute{C}(m,n) \quad (6)$$

[0078] 在这种情况下,输入数据$I'_{(\beta)}$被划分成P个α位数据集$I_{(\alpha),P}(m)$。基于MAC运算单元编号p和划分数据的位宽α,通过下式计算移位参数S(p)的值。

[0079] $S(p)=\alpha\times(p-1)$ (7)

[0080] β位输入特征数据$I'_{(\beta)}$由划分的P个α位数据集$I_{(\alpha),P}(m)$表示为下式。

[0081] $$\sum_{p=1}^{P}I_{(\alpha),p}(m)\times 2^{S(p)}=\acute{I}_{(\beta)}(m) \quad (8)$$

[0082] 在这种情况下,将等式(6)、(7)和(8)代入等式(5),得出输出数据O(n)的等式如下。

[0083] $$O(n)=\sum_{m=1}^{M}\sum_{p=1}^{P}\left[\left(I_{(\alpha),p}(m)\times C_p(m,n)\right)\times 2^{\alpha\times(p-1)}\right]=\sum_{m=1}^{M}\left(\acute{I}_{(\beta)}(m)\times\acute{C}(m,n)\right) \quad (9)$$

[0084] 另一方面,在输入特征数据$I'_{(\alpha)}$的位宽为α位的情况下,基于等式(11)、(12)和

(13)的前提,加法器406的输出被表示为(10)。第n个输出图像的MAC运算结果O(n)由下式给出。

[0085] $$O(n)=\sum_{m=1}^{\frac{M}{P}}\sum_{p=1}^{P}\left[\left(I_{(\alpha),p}(m)\times C_p(m,n)\right)\times 2^{S(p)}\right] \quad (10)$$

[0086] 其中$I_{(\alpha),P}(m)$是α位数据MAC运算单元的输入数据,$C_p(m,n)$是滤波系数,S(p)是移位参数。变量m是α位输入特征图像组(1组=P个图像)的编号(乘法累加器402的处理编号),变量p是MAC运算单元编号和移位运算单元编号,变量n是输出特征图像编号。移位运算由2的幂处理表示。

[0087] 滤波系数$C_p(m,n)$是与第{(m-1)×P+p}个α位特征图像相对应的滤波系数C′((m-1)×P+p,n)。由于滤波系数根据MAC运算单元编号p而不同,因此要并行提供给P个MAC运算单元的滤波系数的数量为P,并且传输计数为P。

[0088] $$C_p(m,n)=\acute{C}((m-1)\times P+p,n) \quad (11)$$

[0089] 输入特征数据$I'_{(\alpha)}$成为α位数据MAC运算单元的输入数据$I_{(\alpha),P}(m)$,并且移位参数S(p)的值始终为0,如下式所示。

[0090] S(p)=0 (12)

[0091] 尽管将P个α位输入特征数据集$I'_{(\alpha)}$直接输入到MAC运算单元,但是P个输入数据集是不同特征图像的特征数据。特征图像编号由MAC运算单元编号p、移位运算单元的数量P和乘法累加器402的处理编号m来表示,如以下等式(13)所示。

[0092] $$I_{(\alpha),p}(m)=\acute{I}_{(\alpha)}((m-1)\times P+p) \quad (13)$$

[0093] 将等式(11)、(12)和(13)代入等式(10)可得出输出数据O(n)的等式如下。

[0094] $$O(n)=\sum_{m=1}^{\frac{M}{P}}\sum_{p=1}^{P}\left[\acute{I}_{(\alpha),p}((m-1)\times P+p)\times\acute{C}((m-1)\times P+p,n)\right] \quad (14)$$

[0095] 通过改变移位参数S(p)的值和滤波系数的数量,可以通过使用相同的运算器(乘法累加器402、移位运算器403和加法器406),来处理位宽为α位的特征数据$I'_{(\alpha)}$和位宽为β位的特征数据$I'_{(\beta)}$。

[0096] <具有不同位宽的情况的处理示例>

[0097] 图5和图10A和图10B示出了当P=4、β=8并且α=2时的布置的示例。乘法累加器402的输入数据的位宽是2位,移位运算器403的输入数据的位宽是6位,加法器406的输入数据的位宽是12位。

[0098] 图6A至图6C示出了在通过使用图5所示的硬件布置来处理图2所示的处理对象网络的情况下的处理时间的示例。图6A和图10A示出了当层1(8位数据,输入特征图像的数量M=4)被处理时的示例。特征图像(1,1)的特征数据$I'_{(8)}(1)$是8位,并且通过基于等式(8)将特征数据划分为四份而获得的四个数据集$I_{(2),1}(1)$至$I_{(2),4}(1)$被输入到乘法累加器402。通过使用输入的特征数据集、移位参数和滤波系数C(m,n)来计算移位运算结果,将计算出的

移位运算结果输入到加法器406,并将初始值零与所获得的结果相加。计算结果被设置为移位运算结果并被加法器406保持。该处理的持续时间是1ms。

[0099] 特征图像(1,2)的特征数据$I'_{(8)}(2)$是8位,并且通过基于等式(8)将特征数据划分为四份而获得的四个数据集$I_{(2),1}(2)$至$I_{(2),4}(2)$被输入到乘法累加器402。通过使用输入的特征数据集、移位参数和滤波系数C(m,n)来计算移位运算结果,计算出的移位运算结果被输入到加法器406并与先前的结果相加。此处理的持续时间为1ms。

[0100] 以与特征图像(1,2)相似的方式依次处理特征图像(1,3)和(1,4),累加移位运算结果,并计算相加结果。该过程的持续时间为2ms。最后,经由处理单元407输出特征图像(2,1)的特征数据。四个特征图像的处理时间为4ms。

[0101] 图6B和图10B示出了当层2(2位数据,输入特征图像的数量M=4)被处理时的示例。特征图像(2,1)至(2,4)的各特征数据集$I'_{(2)}(1)$至$I'_{(2)}(4)$为2位,四个数据集$I_{(2),1}(1)$至$I_{(2),4}(1)$基于等式(13)被并行地输入到乘法累加器402。通过使用输入特征数据、移位参数和滤波系数Cp(m,n)来计算移位运算结果,所获得的结果被输入到加法器406并与初始值零相加,并且计算结果成为移位运算结果。最后,经由处理单元407输出特征图像(3,1)的特征数据。四个特征图像的处理时间为1ms。

[0102] 如图6A和图6B以及图10A和图10B所示,当输入特征数据是8位时,每个输出数据的处理时间是4ms,而当输入特征数据是2位时,每个输出数据的处理时间是1ms。通过公共的数据处理单元305可以高效率地处理不同位宽的数据。

[0103] [第二实施例]

[0104] 下面将描述与第一实施例的不同之处。以下未特别提及的事项与第一实施例相似。

[0105] <移位运算和MAC运算的顺序>

[0106] 第一实施例描述了在MAC运算之后进行移位运算的示例。然而,即使MAC运算和移位运算的顺序被切换,也可以获得相同的处理结果。当MAC运算和移位运算的顺序被切换时,图1的流程图的一部分将改变。步骤S107改变为图9中的步骤S901至S903。

[0107] 图7示出了根据该实施例的数据处理单元305的布置的示例。移位运算器701基于移位参数来将在数据存储器401中存储的特征数据进行移位,并且乘法累加器702基于移位的特征数据和滤波系数来计算MAC运算结果。

[0108] <移位运算和MAC运算>

[0109] 将描述在步骤S107中进行的MAC运算和移位运算(步骤S901至S903)。在步骤S901中,控制单元704从数据存储器401中读出特征数据,并从系数存储器703中读出滤波系数。在步骤S902中,移位运算器701基于在步骤S103中设置的移位参数来移位特征数据。在步骤S903中,乘法累加器702基于移位的特征数据和滤波系数来计算MAC运算结果。

[0110] <不同位宽的情况的详细说明>

[0111] 在该实施例中,移位运算器701包括用于计算移位运算结果的P个α位数据移位运算单元,并且乘法累加器702包括用于计算MAC运算结果的P个α位数据MAC运算单元。乘法累加器702的输出由下面的等式(15)表示,并且等效于等式(5)中所示的移位运算器403的输出。

[0112] $$O(n)=\sum_{m=1}^{M}\sum_{p=1}^{P}\left[\left(I_{(\alpha),p}(m)\times 2^{S(p)}\right)\times C_p(m,n)\right] \quad (15)$$

[0113] 图8示出了P＝4、β＝8并且α＝2的情况的示例。移位运算器701的输入数据的位宽是2位,乘法累加器702的输入数据的位宽是8位,加法器406的输入数据的位宽是12位。由于移位运算器701的电路规模和乘法累加器702的电路规模因位宽不同而不同,因此可以通过切换移位运算器701(移位运算器403)和乘法累加器702(乘法累加器402)的顺序来减小整体电路规模。

[0114] [第三实施例]

[0115] 第一和第二实施例描述了如下示例:输入特征数据的位宽是α位(各个MAC运算单元的位宽)和β位(各个MAC运算单元的位宽与MAC运算单元的数量的乘积)。然而,本发明不限于这些,并且可以使用α和β之外的位宽。

[0116] <输入特征数据的位宽为γ位的情况>

[0117] 在该实施例中,可以处理位宽为γ位的输入特征数据。图10C示出了特征数据是4位的示例。在特征数据是4位的情况下,乘法累加器402将两个4位特征数据集1003(值:10、14)中的各个划分为2位,如图10C所示。乘法累加器402使用通过该划分而获得的四个2位数据集(值:2、2、2、3)和两个滤波系数来计算四个MAC运算结果。移位运算器403基于两个移位参数来移位四个MAC运算结果。加法器406将四个移位的MAC运算结果相加以计算一个特征数据集(两个4位输入特征数据集的MAC运算结果的总和)。以这种方式,数据处理单元305可以并行处理两个4位输入特征数据集。γ是输入特征数据的位宽,γ的值与β的值不同。α、β和P的定义与第一实施例中的相同,并且γ、α和P'满足以下条件。

[0118] $$\gamma=\alpha\times\acute{P} \quad (16)$$

[0119] 其中γ小于β,并且P是P'的倍数。在输入特征数据$I'_{(\gamma)}$的位宽是γ位的情况下,基于等式(18)、(19)和(20)的前提,加法器406的输出数据O(n)由下面的等式(17)表示。第n个输出特征图像的MAC运算结果O(n)由下式给出。

[0120] $$O(n)=\sum_{m=1}^{\frac{M\acute{P}}{P}}\sum_{q=1}^{\frac{P}{\acute{P}}}\sum_{p=1}^{\acute{P}}\left[\left(I_{(\alpha),p,q}(m)\times C_{p,q}(m,n)\right)\times 2^{S(p,q)}\right] \quad (17)$$

[0121] 其中$I_{(\alpha),p}(m)$是α位数据MAC运算单元的输入数据,$C_p(m,n)$是滤波系数,S(p)是移位参数。变量m是α位输入特征图像组(1组＝P个图像)的编号(乘法累加器402的处理编号)。MAC运算单元被划分为P/P'个集合,移位运算单元被划分为P/P'个集合,并且变量q是MAC运算单元的集合编号。变量p是集合中的MAC运算单元编号和移位运算单元编号,变量n是输出特征图像编号。移位运算由2的幂处理表示。

[0122] 滤波系数$C_{p,q}(m,n)$是与第{(m-1)×P/P'+q}个γ位特征图像相对应的滤波系数C'((m-1)×P/P'+q,n)。基于MAC运算单元的集合编号q来计算滤波系数。由于一部分滤波系数是共享的,因此要并行提供给P个MAC运算单元的滤波系数的数量为P/P',传输计数为P/P'。

[0123] $$C_{p,q}(m,n)=\acute{C}\left((m-1)\times\frac{P}{\acute{P}}+q,n\right) \quad (18)$$

[0124] 在这种情况下,输入特征数据I'$_{(\gamma)}$被划分为P'个α位数据集I$_{(\alpha),p}$(m)。基于MAC运算单元的位宽α和MAC运算单元编号p来计算移位参数S(.)。

[0125] S(p,q)=α×(p-1) (19)

[0126] γ位输入特征数据I'$_{(\gamma)}$i由划分的P'个α位数据集I$_{(\alpha),p,q}$(m)表示。

[0127] $$\sum_{p=1}^{\acute{P}}\left(I_{(\alpha),p,q}(m)\times 2^{S(p,q)}\right)=\acute{I}_{(\gamma)}\left((m-1)\times\frac{P}{\acute{P}}+q\right) \quad (20)$$

[0128] 将等式(18)、(19)和(20)代入等式(17)可得出输出数据O(n)的等式如下。

[0129] $$O(n)=\sum_{m=1}^{\frac{M\acute{P}}{P}}\sum_{q=1}^{\frac{P}{\acute{P}}}\sum_{p=1}^{\acute{P}}\left[\left(I_{(\alpha),p,q}(m)\times C_{p,q}(m,n)\right)\times 2^{\alpha\times(p-1)}\right]$$

$$=\sum_{m=1}^{\frac{M\acute{P}}{P}}\sum_{q=1}^{\frac{P}{\acute{P}}}\left[\acute{I}_{(\gamma)}\left((m-1)\times\frac{P}{\acute{P}}+q\right)\times\acute{C}\left((m-1)\times\frac{P}{\acute{P}}+q,n\right)\right] \quad (21)$$

[0130] 通过设置移位参数S(p,q)的值和滤波系数的数量,可以通过使用与第一实施例中相同的运算器(乘法累加器402、移位运算器403和加法器406)来处理位宽为γ位的特征数据I'$_{(\gamma)}$。

[0131] <具有不同位宽的情况的处理示例>

[0132] 图5和图10C示出了当P=4、β=8且α=2时的布置的示例。图6A至图6C示出了当通过使用图5所示的硬件布置来处理图2所示的处理对象网络时的处理时间的示例。

[0133] 图6C和图10C示出了P'=2、γ=4并且层3(4位数据,输入特征图像的数量M=4)被处理的情况的示例。特征图像(3,1)和(3,2)的特征数据I'$_{(4),}$(1)和I'$_{(4),}$(2)中的各个是4位,并且将基于等式(20)划分的四个数据集I$_{(2),1}$(1)至I$_{(2),4}$(1)输入到乘法累加器402。通过使用输入特征数据、移位参数和滤波系数C(m,n)计算移位运算结果,计算出的移位运算结果被输入到加法器406,并且将初始值零与计算出的结果相加。计算结果变为移位运算结果,并由加法器406保持。该处理的持续时间是1ms。

[0134] 特征图像(3,3)和(3,4)的特征数据I'$_{(4),}$(3)和I'$_{(4),}$(4)中的各个是4位,并且将基于等式(20)划分的四个数据集I$_{(2),1}$(2)至I$_{(2),4}$(2)输入到乘法累加器402。通过使用输入特征数据、移位参数和滤波系数C(m,n)来计算移位运算结果,将计算出的移位运算结果输入到加法器406,并将该结果与先前的结果相加。该操作的持续时间为1ms。最后,经由处理单元407输出特征图像(4,1)的特征数据。四个特征图像的处理时间为2ms。

[0135] 以这种方式,该实施例的优点在于高度灵活,这是因为可以处理除了位宽为α位(各个MAC运算单元的位宽)或β位(各个MAC运算单元的位宽α与MAC运算单元的数量P的乘积)的数据之外的特征数据。

[0136] [第四实施例]

[0137] 尽管第一实施例描述了由处理单元407执行活化处理的示例,但是活化处理的执行不限于处理单元407,并且可以被设置为使得其他设备(例如CPU 306)将执行活化处理。这也类似地适用于其他处理操作,并且以上实施例仅示出了各种处理的主体的示例,并且可以使用与以上实施例中描述的主体不同的主体。

[0138] 另外,在第一实施例中,根据层信息执行活化/池化处理。然而,根据情况,可以省略活化/池化处理。

[0139] 此外,尽管第一至第三实施例描述了滤波器大小(各个滤波器的高度和宽度)是1×1的情况,但是滤波器大小不限于1×1,并且可以是其他大小。在以上实施例的描述中使用的数值仅是用于进行更具体说明的示例,并且不旨在将要使用的数值限制为在以上实施例中描述的数值。

[0140] 在滤波器大小较小的情况下,具有的优点在于,可以使用于保持滤波系数的存储器(系数存储器404或703)的容量较小。被设置为滤波器宽度和滤波器高度的最小值为1。

[0141] 此外,第一至第三实施例将输入特征图像的数量设置为M,并且将输出特征图像的数量设置为N。然而,适用于M和N的数值不限于特定数值。以这种方式,适用于上述各种变量的数值不限于特定数值。

[0142] 此外,尽管在第一至第三实施例中将滤波系数保持在系数存储器404或703中并将特征数据保持在数据存储器401中,但是用于保持滤波系数和特征数据的存储器不限于特定存储器。例如,滤波系数和特征数据可以保持在乘法累加器402或702中包括的存储器中,或者可以保持在RAM 308中。

[0143] 另外,各个滤波系数的位宽不限于特定的位宽。此外,尽管在第一至第三实施例中将CNN用作处理对象网络,但是处理对象网络不限于CNN,并且可以是多个其他种类的层被分层连接到的网络,例如RNN、MLP(多层感知器)等。

[0144] 其他实施例

[0145] 还可以通过读出并执行记录在存储介质(也可更完整地称为“非临时性计算机可读存储介质”)上的计算机可执行指令(例如,一个或更多个程序)以执行上述实施例中的一个或更多个的功能、并且/或者包括用于执行上述实施例中的一个或更多个的功能的一个或更多个电路(例如,专用集成电路(ASIC))的系统或装置的计算机,来实现本发明的实施例,并且,可以利用通过由所述系统或装置的所述计算机例如读出并执行来自所述存储介质的所述计算机可执行指令以执行上述实施例中的一个或更多个的功能、并且/或者控制所述一个或更多个电路执行上述实施例中的一个或更多个的功能的方法,来实现本发明的实施例。所述计算机可以包括一个或更多个处理器(例如,中央处理单元(CPU),微处理单元(MPU)),并且可以包括分开的计算机或分开的处理器的网络,以读出并执行所述计算机可执行指令。所述计算机可执行指令可以例如从网络或所述存储介质被提供给计算机。所述存储介质可以包括例如硬盘、随机存取存储器(RAM)、只读存储器(ROM)、分布式计算系统的存储器、光盘(诸如压缩光盘(CD)、数字通用光盘(DVD)或蓝光光盘(BD)™)、闪存设备以及存储卡等中的一个或更多个。

[0146] 本发明的实施例还可以通过如下的方法来实现,即,通过网络或者各种存储介质将执行上述实施例的功能的软件(程序)提供给系统或装置,该系统或装置的计算机或是中央处理单元(CPU)、微处理单元(MPU)读出并执行程序的方法。

[0147] 虽然参照示例性实施例对本发明进行了描述,但是应当理解,本发明并不限于所公开的示例性实施例。应当对所附权利要求的范围给予最宽的解释,以使其涵盖所有这些变型例以及等同的结构和功能。

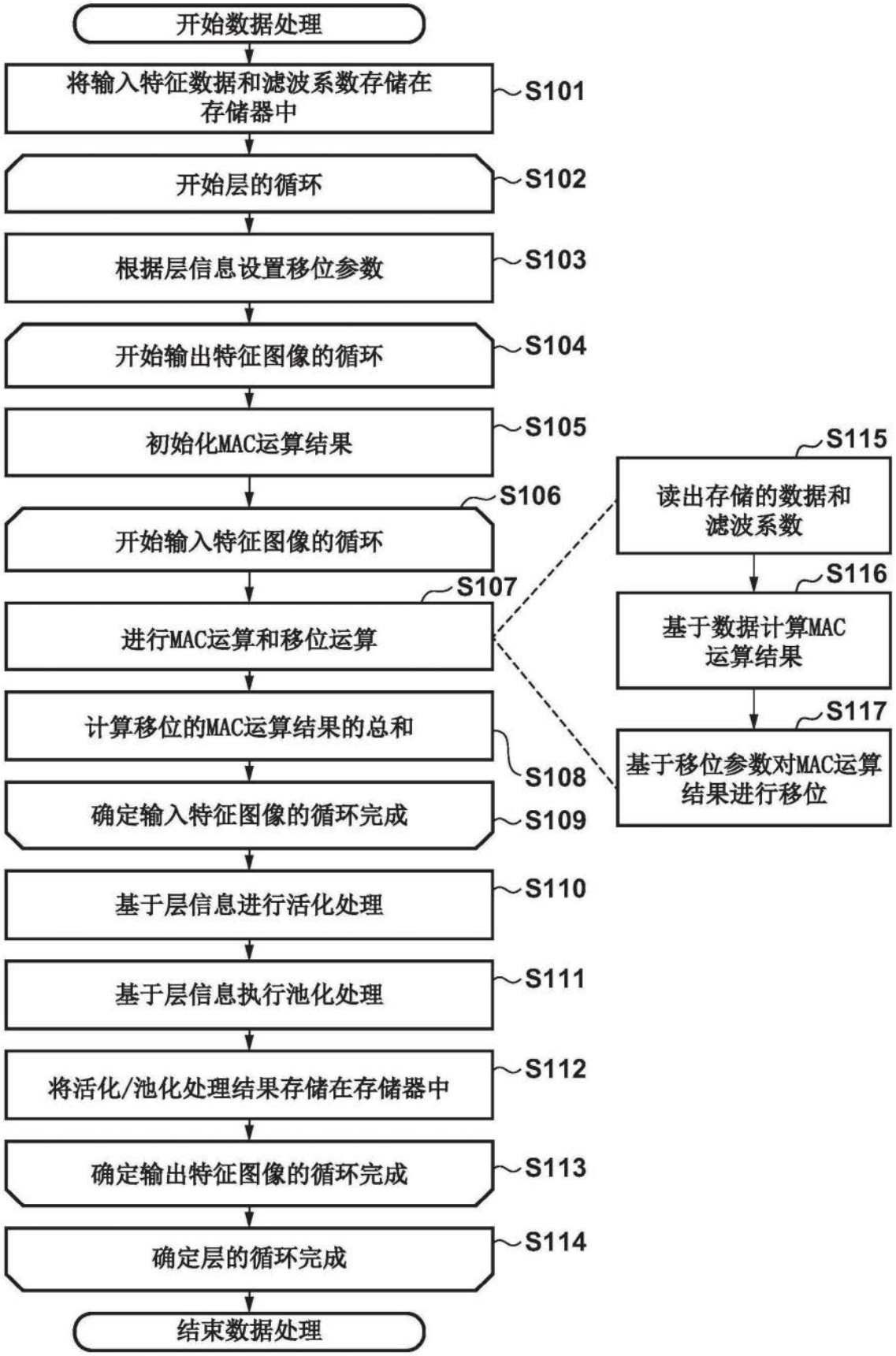

开始数据处理
将输入特征数据和滤波系数存储在存储器中
S101
开始层的循环
S102
根据层信息设置移位参数
S103
开始输出特征图像的循环
S104
初始化MAC运算结果
S105
开始输入特征图像的循环
S106
进行MAC运算和移位运算
S107
计算移位的MAC运算结果的总和
S108
确定输入特征图像的循环完成
S109
基于层信息进行活化处理
S110
基于层信息执行池化处理
S111
将活化/池化处理结果存储在存储器中
S112
确定输出特征图像的循环完成
S113
确定层的循环完成
S114
结束数据处理
读出存储的数据和滤波系数
S115
基于数据计算MAC运算结果
S116
基于移位参数对MAC运算结果进行移位
S117

图1

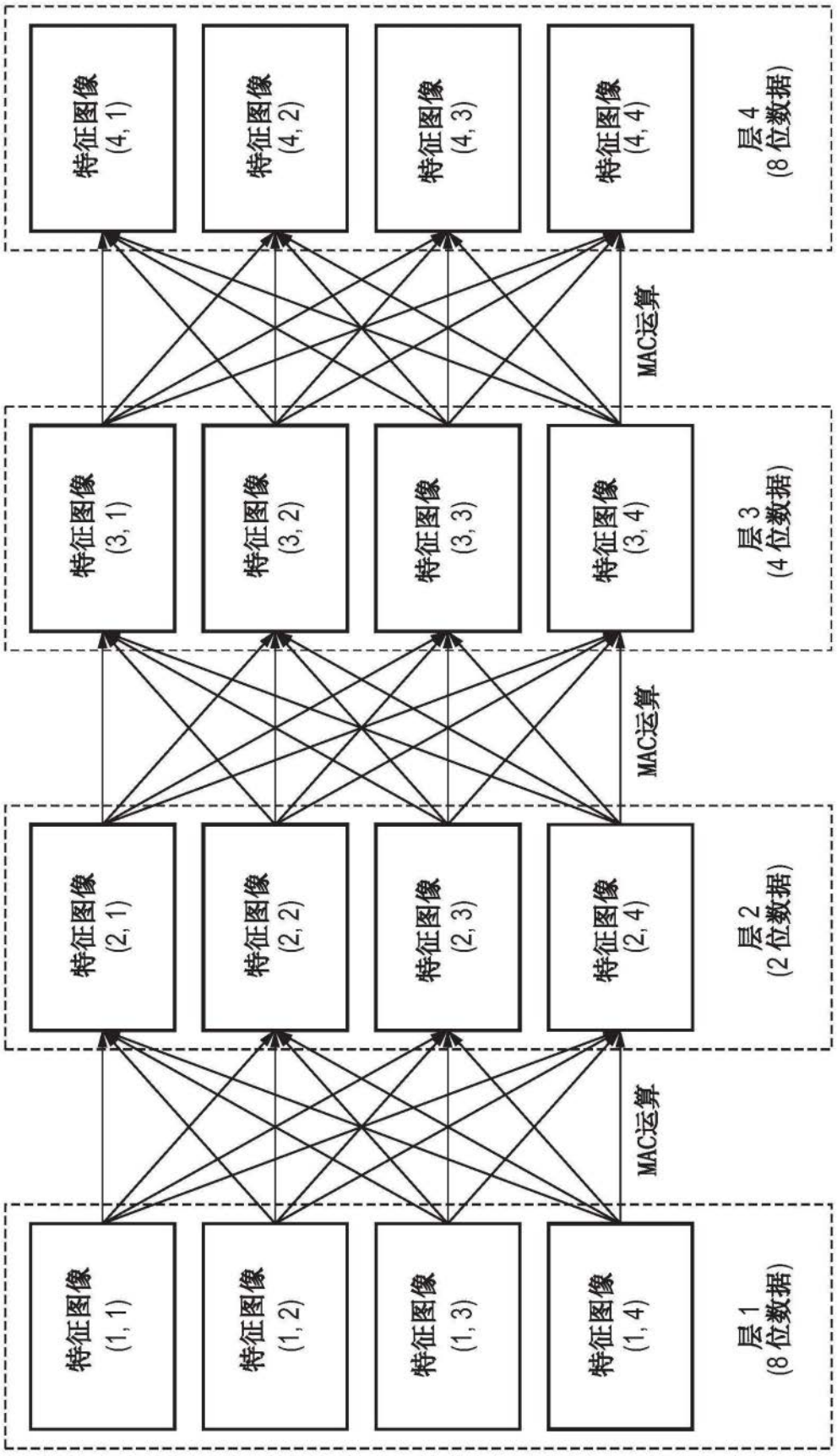

特征图像 (4, 1)
特征图像 (4, 2)
特征图像 (4, 3)
特征图像 (4, 4)
层 4 (8 位数据)
MAC运算
特征图像 (3, 1)
特征图像 (3, 2)
特征图像 (3, 3)
特征图像 (3, 4)
层 3 (4 位数据)
MAC运算
特征图像 (2, 1)
特征图像 (2, 2)
特征图像 (2, 3)
特征图像 (2, 4)
层 2 (2 位数据)
MAC运算
特征图像 (1, 1)
特征图像 (1, 2)
特征图像 (1, 3)
特征图像 (1, 4)
层 1 (8 位数据)

图2

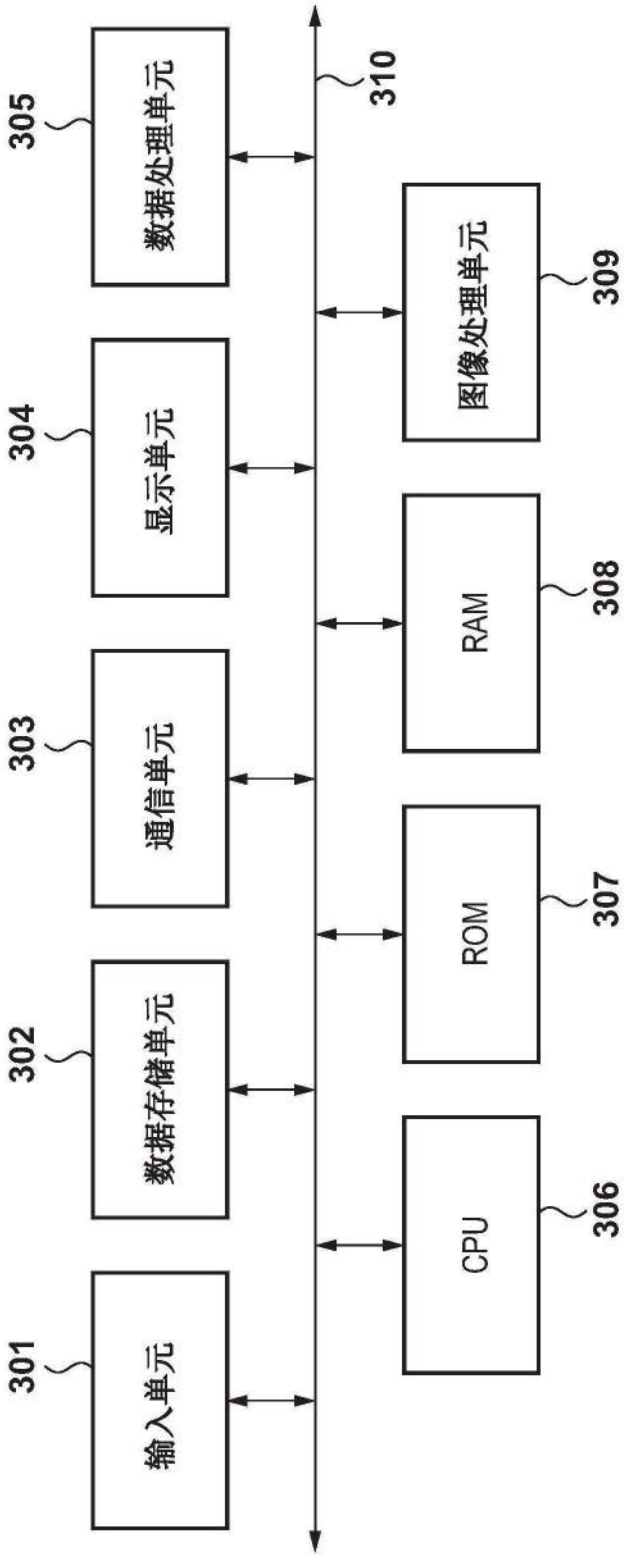

301
输入单元
302
数据存储单元
303
通信单元
304
显示单元
305
数据处理单元
306
CPU
307
ROM
308
RAM
309
图像处理单元
310

图3

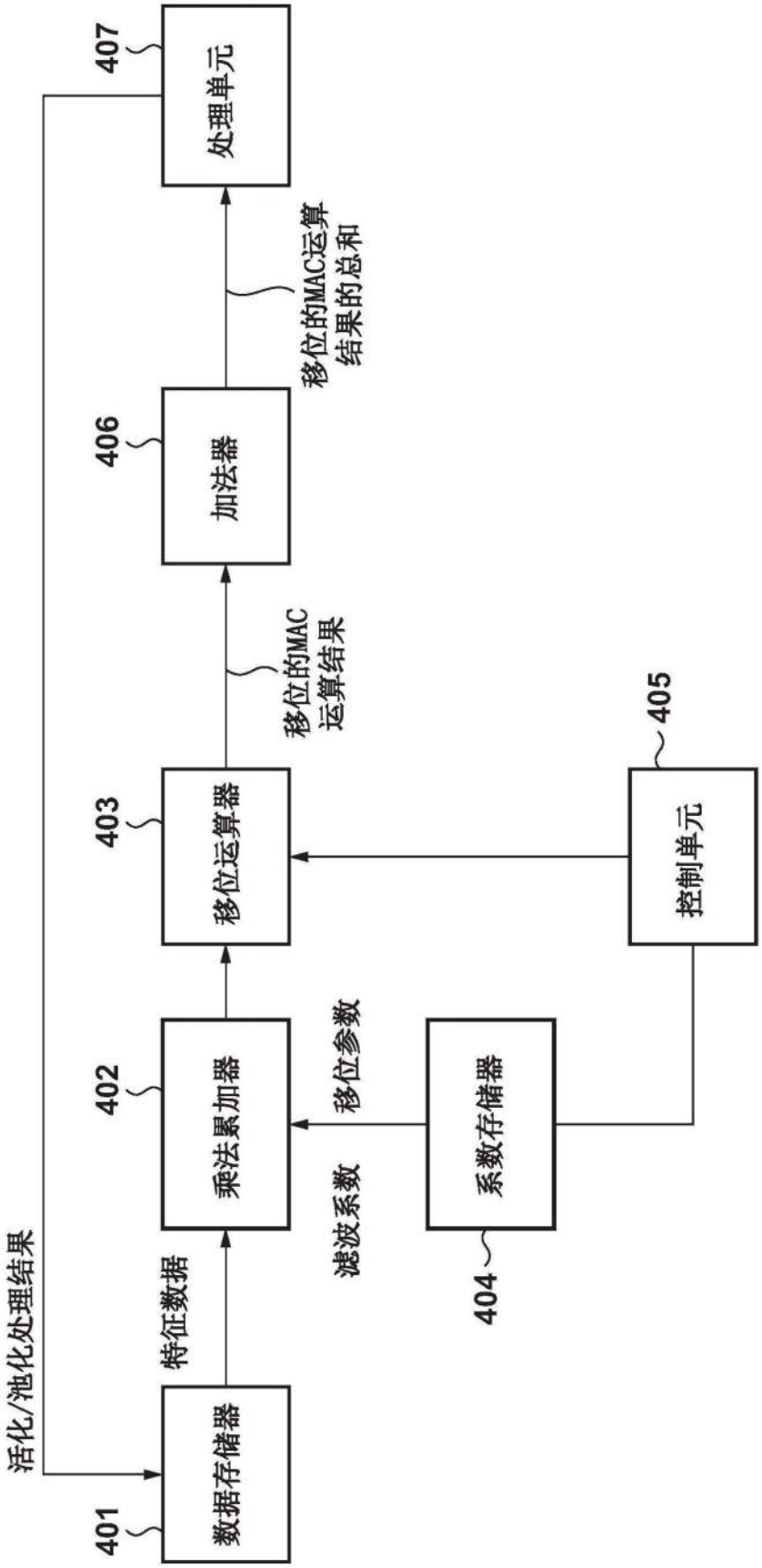

401
数据存储器
402
乘法累加器
403
移位运算器
404
系数存储器
405
控制单元
406
加法器
407
处理单元
特征数据
滤波系数
移位参数
移位的MAC
运算结果
移位的MAC运算
结果的总和
活化/池化处理结果

图4

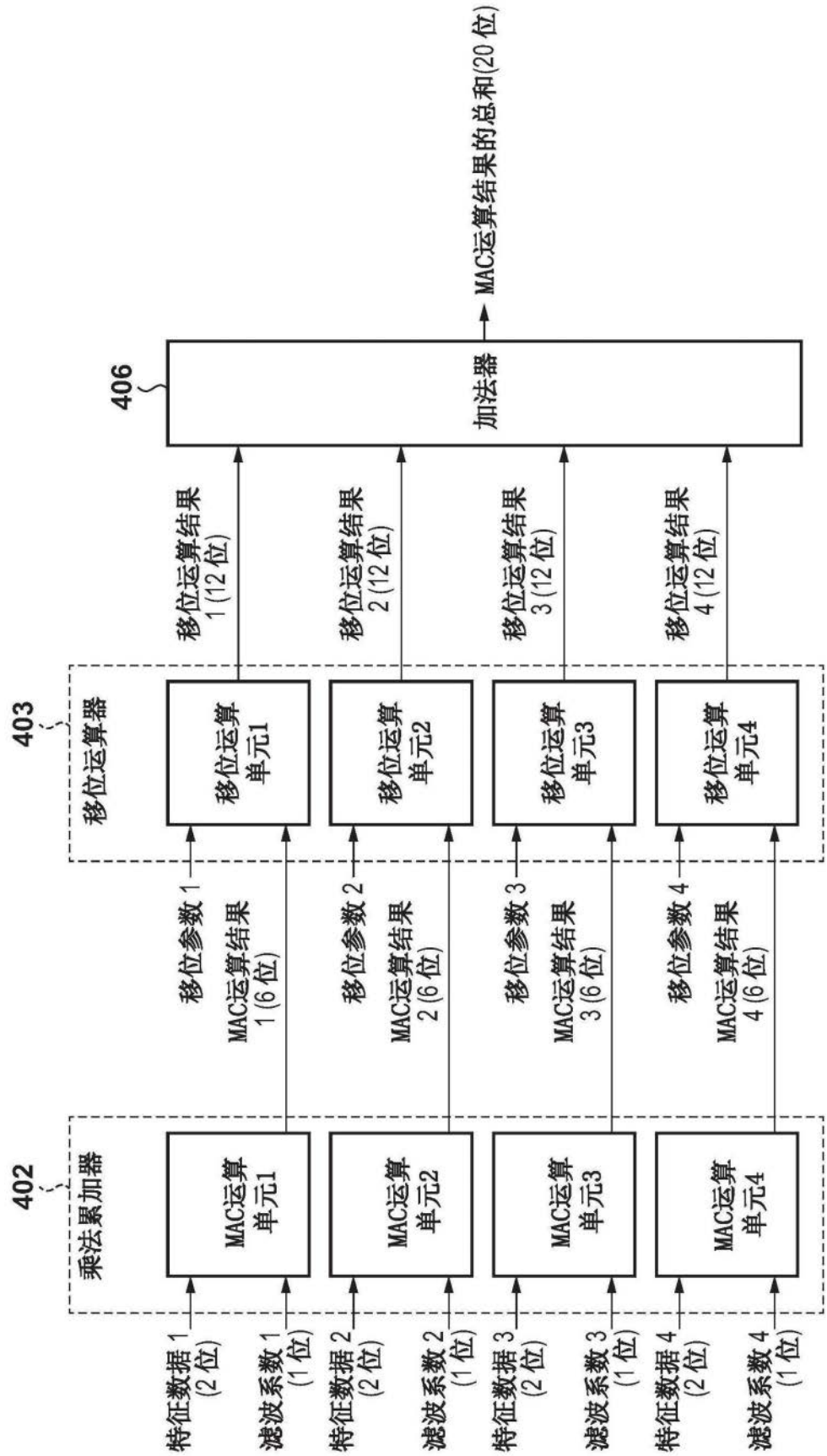

MAC运算结果的总和(20 位)
406
加法器
移位运算结果 1 (12 位)
移位运算结果 2 (12 位)
移位运算结果 3 (12 位)
移位运算结果 4 (12 位)
403
移位运算器
移位运算单元1
移位运算单元2
移位运算单元3
移位运算单元4
移位参数 1
MAC运算结果 1 (6 位)
移位参数 2
MAC运算结果 2 (6 位)
移位参数 3
MAC运算结果 3 (6 位)
移位参数 4
MAC运算结果 4 (6 位)
402
乘法累加器
MAC运算单元1
MAC运算单元2
MAC运算单元3
MAC运算单元4
特征数据 1 (2 位)
滤波系数 1 (1 位)
特征数据 2 (2 位)
滤波系数 2 (1 位)
特征数据 3 (2 位)
滤波系数 3 (1 位)
特征数据 4 (2 位)
滤波系数 4 (1 位)

图5

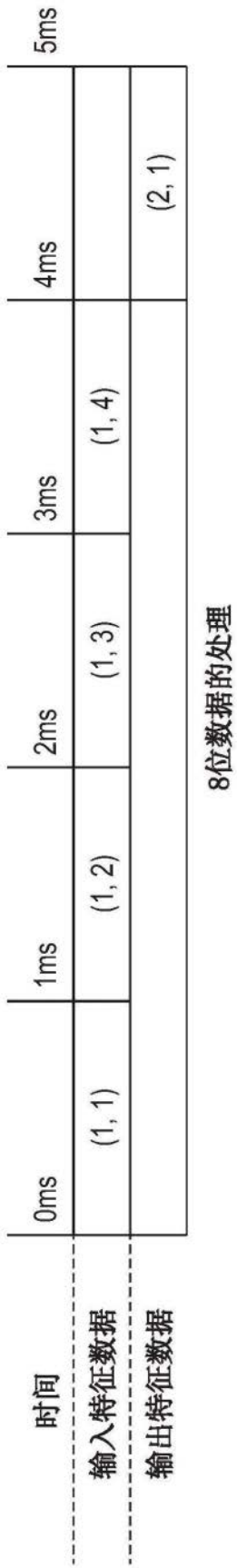


图6A

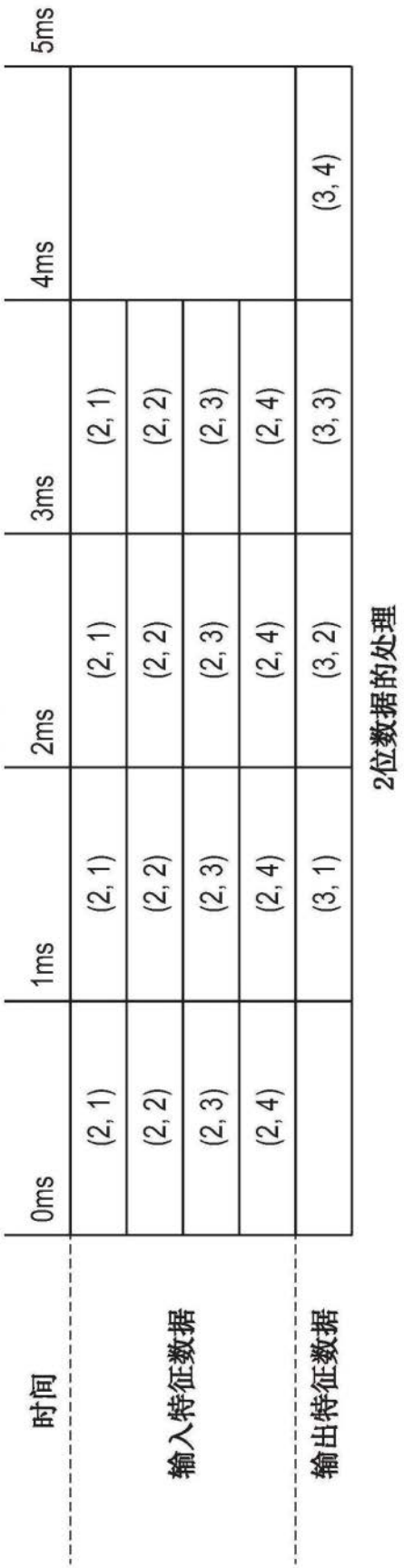

时间
0ms
1ms
2ms
3ms
4ms
5ms
输入特征数据
(2, 1)
(2, 2)
(2, 3)
(2, 4)
(2, 1)
(2, 2)
(2, 3)
(2, 4)
(3, 1)
(2, 1)
(2, 2)
(2, 3)
(2, 4)
(3, 2)
(2, 1)
(2, 2)
(2, 3)
(2, 4)
(3, 3)
(3, 4)
输出特征数据
2位数据的处理

图6B

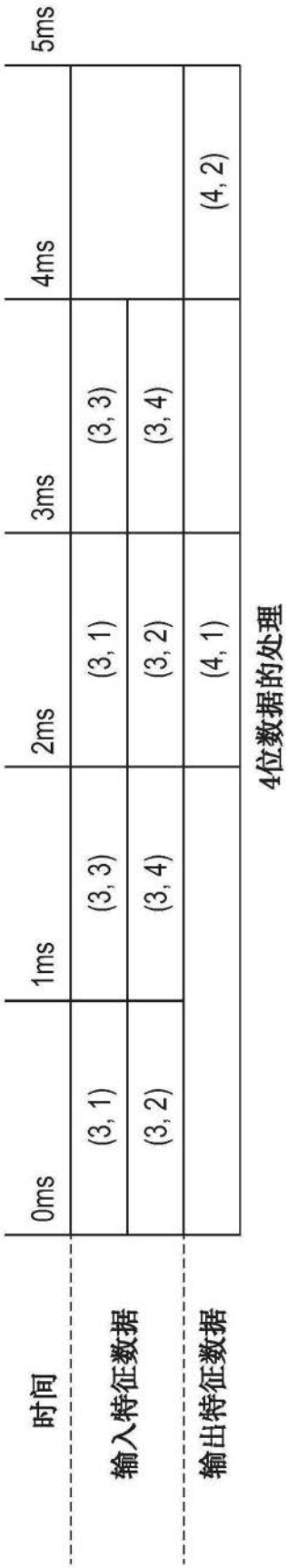

时间
0ms
1ms
2ms
3ms
4ms
5ms
输入特征数据
(3, 1)
(3, 2)
(3, 3)
(3, 4)
(3, 1)
(3, 2)
(3, 3)
(3, 4)
输出特征数据
(4, 1)
(4, 2)
4位数据的处理

图6C

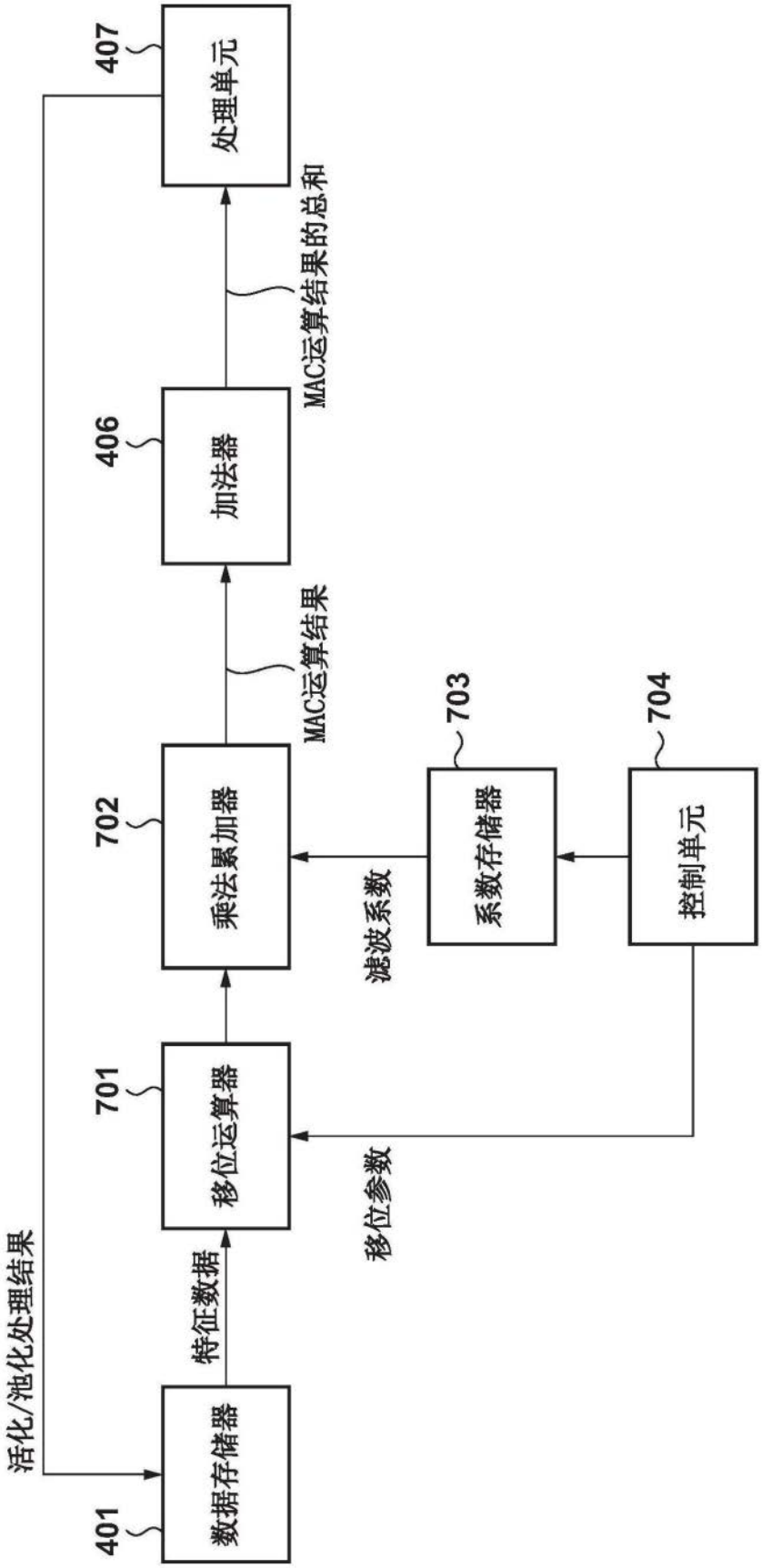

401
数据存储器
特征数据
701
移位运算器
移位参数
702
乘法累加器
滤波系数
703
系数存储器
704
控制单元
MAC运算结果
406
加法器
MAC运算结果的总和
407
处理单元
活化/池化处理结果

图7

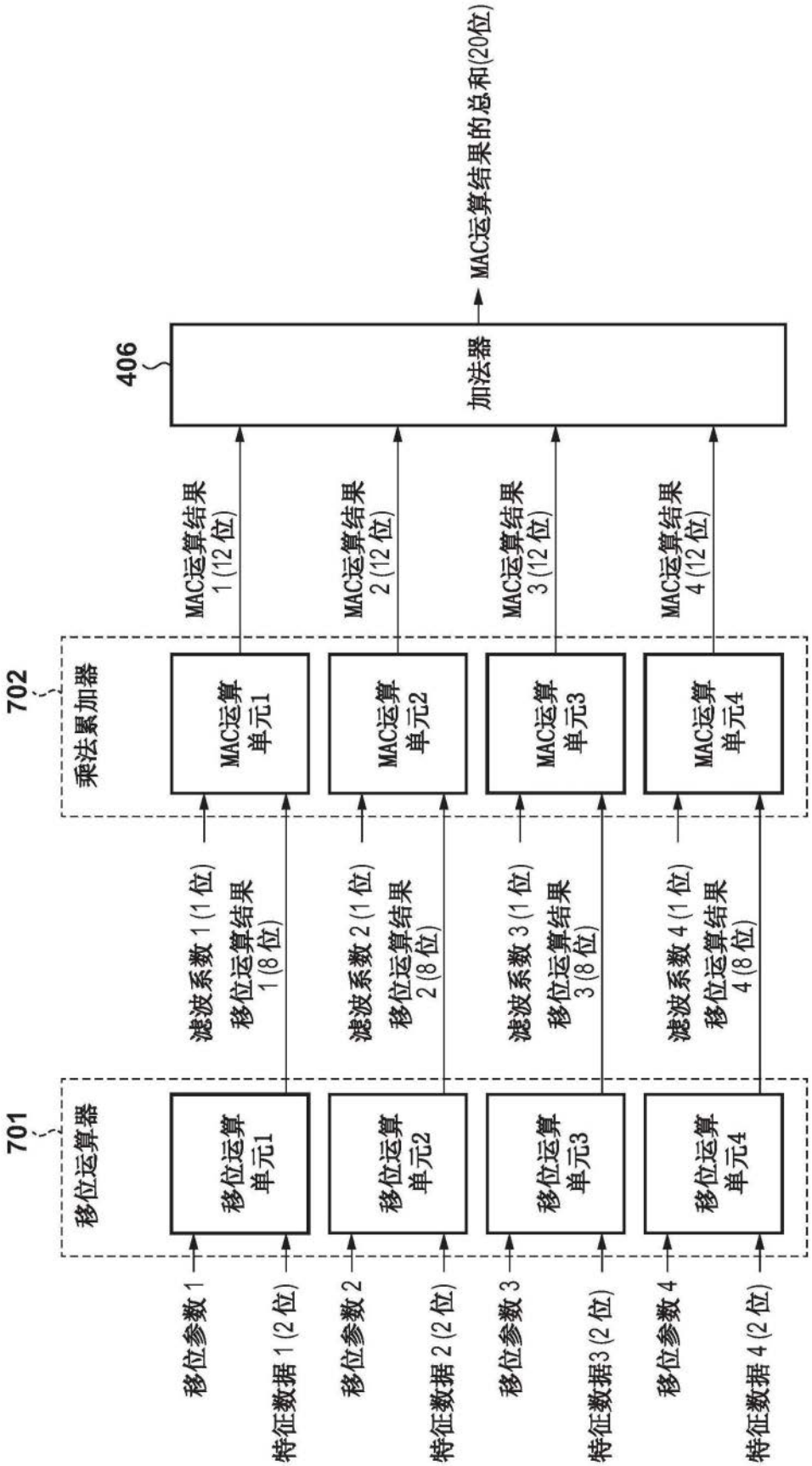

MAC运算结果的总和(20位)
406
加法器
MAC运算结果 1 (12 位)
MAC运算结果 2 (12 位)
MAC运算结果 3 (12 位)
MAC运算结果 4 (12 位)
702
乘法累加器
MAC运算单元1
MAC运算单元2
MAC运算单元3
MAC运算单元4
滤波系数 1 (1 位)
移位运算结果 1 (8 位)
滤波系数 2 (1 位)
移位运算结果 2 (8 位)
滤波系数 3 (1 位)
移位运算结果 3 (8 位)
滤波系数 4 (1 位)
移位运算结果 4 (8 位)
701
移位运算器
移位运算单元1
移位运算单元2
移位运算单元3
移位运算单元4
移位参数 1
特征数据 1 (2 位)
移位参数 2
特征数据 2 (2 位)
移位参数 3
特征数据3 (2 位)
移位参数 4
特征数据 4 (2 位)

图8

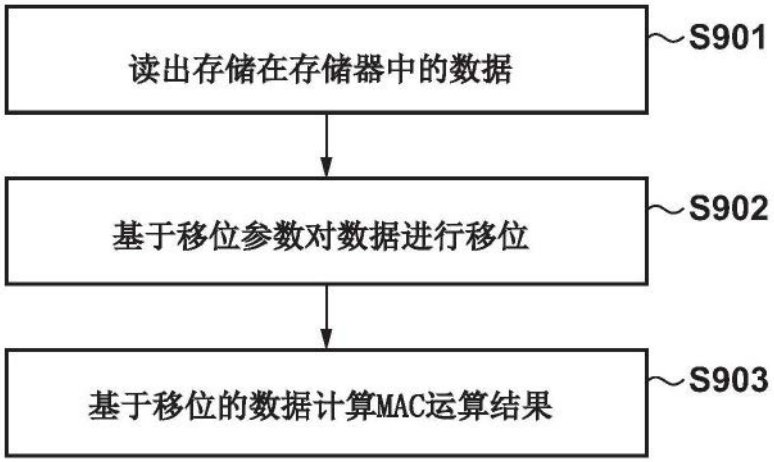

读出存储在存储器中的数据
S901
基于移位参数对数据进行移位
S902
基于移位的数据计算MAC运算结果
S903

图9

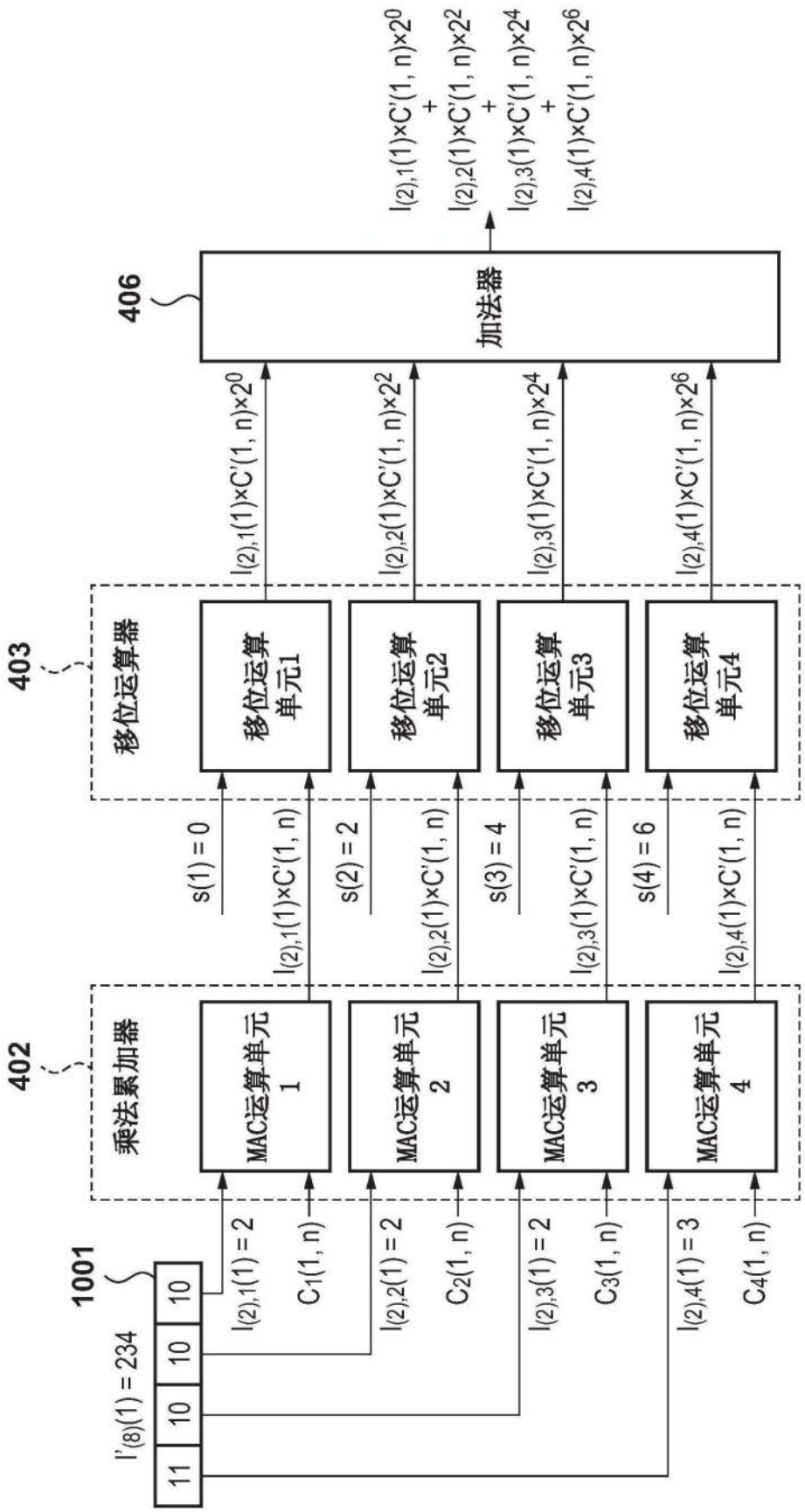

406
加法器
$I_{(2),1}(1)\times C'(1,n)\times 2^0$ + $I_{(2),2}(1)\times C'(1,n)\times 2^2$ + $I_{(2),3}(1)\times C'(1,n)\times 2^4$ + $I_{(2),4}(1)\times C'(1,n)\times 2^6$
403
移位运算器
移位运算单元1
移位运算单元2
移位运算单元3
移位运算单元4
$I_{(2),1}(1)\times C'(1,n)\times 2^0$
$I_{(2),2}(1)\times C'(1,n)\times 2^2$
$I_{(2),3}(1)\times C'(1,n)\times 2^4$
$I_{(2),4}(1)\times C'(1,n)\times 2^6$
s(1) = 0
s(2) = 2
s(3) = 4
s(4) = 6
402
乘法累加器
MAC运算单元 1
MAC运算单元 2
MAC运算单元 3
MAC运算单元 4
$I_{(2),1}(1)\times C'(1,n)$
$I_{(2),2}(1)\times C'(1,n)$
$I_{(2),3}(1)\times C'(1,n)$
$I_{(2),4}(1)\times C'(1,n)$
1001
$I'_{(8)}(1) = 234$
11
10
10
10
$I_{(2),1}(1) = 2$
$I_{(2),2}(1) = 2$
$I_{(2),3}(1) = 2$
$I_{(2),4}(1) = 3$
$C_1(1,n)$
$C_2(1,n)$
$C_3(1,n)$
$C_4(1,n)$

图10A

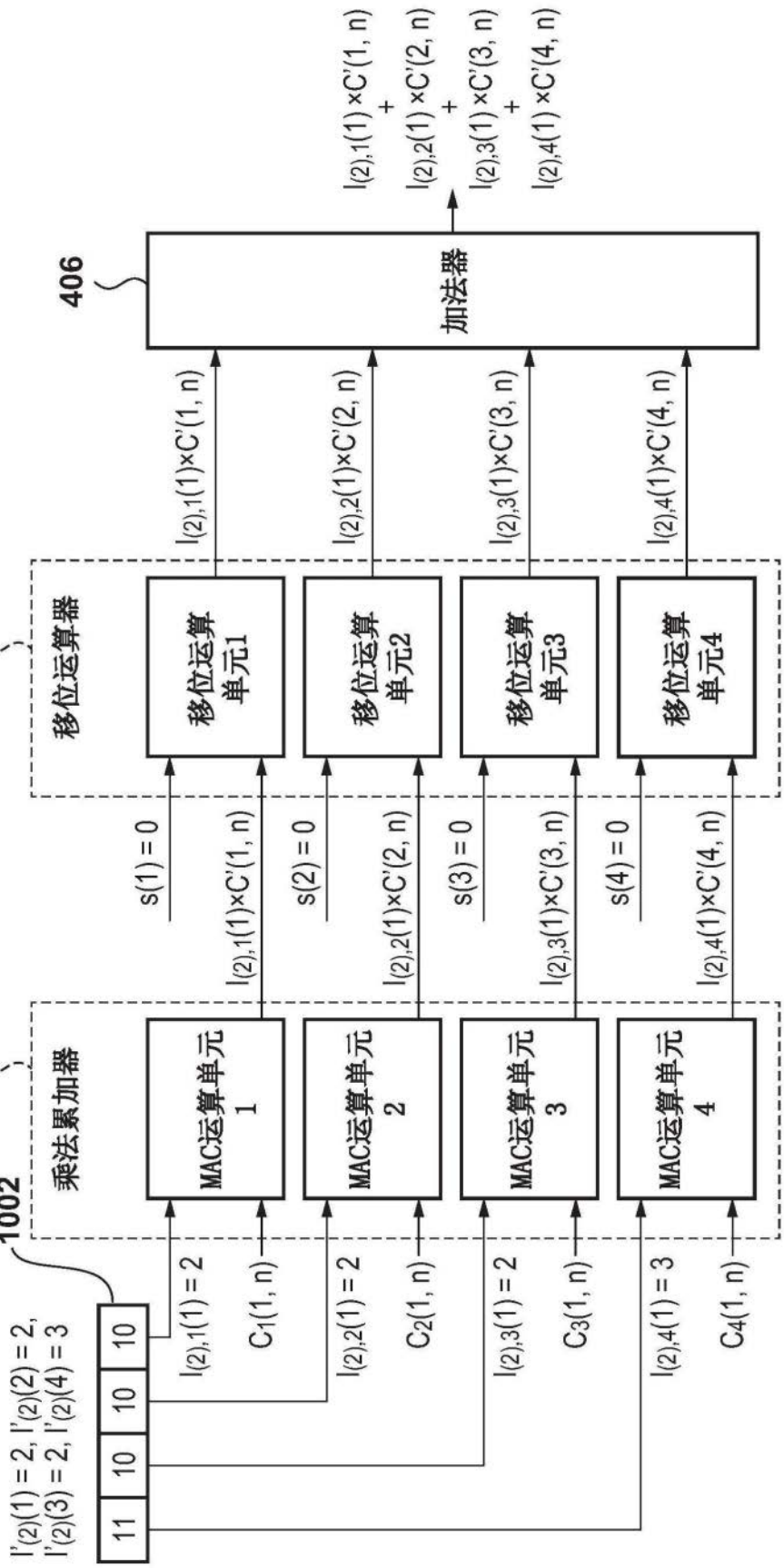

406
加法器
$I_{(2),1}(1)\times C'(1,n) + I_{(2),2}(1)\times C'(2,n) + I_{(2),3}(1)\times C'(3,n) + I_{(2),4}(1)\times C'(4,n)$
403
移位运算器
移位运算单元1
移位运算单元2
移位运算单元3
移位运算单元4
$I_{(2),1}(1)\times C'(1,n)$
$I_{(2),2}(1)\times C'(2,n)$
$I_{(2),3}(1)\times C'(3,n)$
$I_{(2),4}(1)\times C'(4,n)$
s(1) = 0
s(2) = 0
s(3) = 0
s(4) = 0
402
乘法累加器
MAC运算单元1
MAC运算单元2
MAC运算单元3
MAC运算单元4
1002
$I'_{(2)}(1) = 2, I'_{(2)}(2) = 2,$
$I'_{(2)}(3) = 2, I'_{(2)}(4) = 3$
11
10
10
10
$I_{(2),1}(1) = 2$
$I_{(2),2}(1) = 2$
$I_{(2),3}(1) = 2$
$I_{(2),4}(1) = 3$
$C_1(1,n)$
$C_2(1,n)$
$C_3(1,n)$
$C_4(1,n)$

图10B

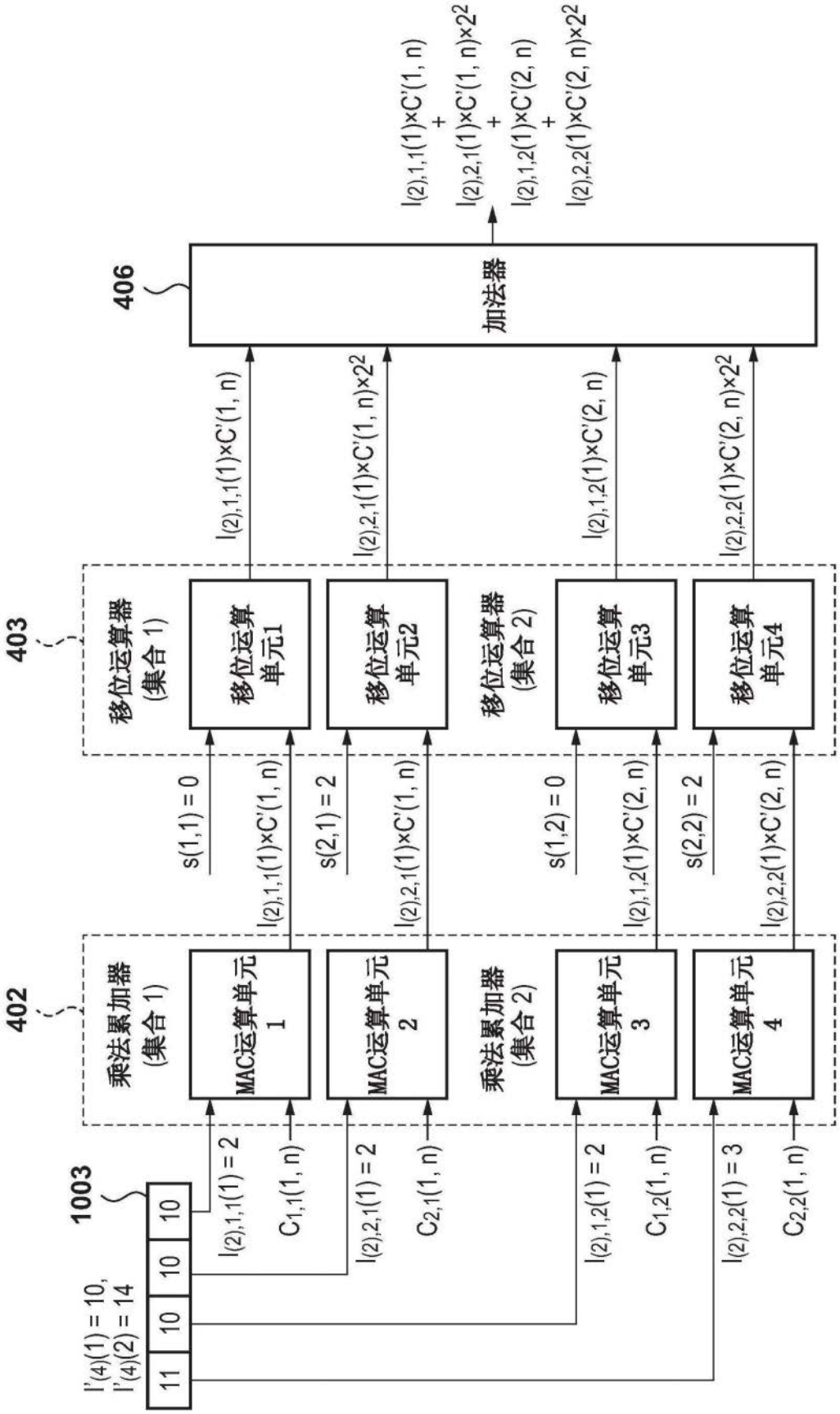

406
加法器
$I_{(2),1,1}(1)\times C'(1,n) + I_{(2),2,1}(1)\times C'(1,n)\times 2^2 + I_{(2),1,2}(1)\times C'(2,n) + I_{(2),2,2}(1)\times C'(2,n)\times 2^2$
403
移位运算器(集合 1)
移位运算单元1
移位运算单元2
移位运算器(集合 2)
移位运算单元3
移位运算单元4
$I_{(2),1,1}(1)\times C'(1,n)$
$I_{(2),2,1}(1)\times C'(1,n)\times 2^2$
$I_{(2),1,2}(1)\times C'(2,n)$
$I_{(2),2,2}(1)\times C'(2,n)\times 2^2$
s(1,1) = 0
s(2,1) = 2
s(1,2) = 0
s(2,2) = 2
402
乘法累加器(集合 1)
MAC运算单元1
MAC运算单元2
乘法累加器(集合 2)
MAC运算单元3
MAC运算单元4
1003
11
10
10
10
$I'_{(4)}(1) = 10,$
$I'_{(4)}(2) = 14$
$I_{(2),1,1}(1) = 2$
$I_{(2),2,1}(1) = 2$
$I_{(2),1,2}(1) = 2$
$I_{(2),2,2}(1) = 3$
$C_{1,1}(1,n)$
$C_{2,1}(1,n)$
$C_{1,2}(1,n)$
$C_{2,2}(1,n)$

图10C