(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 108073983 B
(45)授权公告日 2022.04.26

(21)申请号 201710909648.4
(22)申请日 2017.09.29
(65)同一申请的已公布的文献号
申请公布号 CN 108073983 A
(43)申请公布日 2018.05.25
(30)优先权数据
15/348,199 2016.11.10 US
15/467,382 2017.03.23 US
(73)专利权人 谷歌有限责任公司
地址 美国加利福尼亚州
(72)发明人 雷吉纳尔德·克利福德·扬
威廉·约翰·格兰德
(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
代理人 周亚荣 安翔
(51)Int.Cl.
G06N 3/063(2006.01)
G06N 3/04(2006.01)
(56)对比文件
CN 105488565 A,2016.04.13
CN 105426517 A,2016.03.23
CN 105589938 A,2016.05.18
CN 205621018 U,2016.10.05
CN 104035751 A,2014.09.10
CN 105678379 A,2016.06.15
CN 105320495 A,2016.02.10
CN 104915322 A,2015.09.16
CN 105681628 A,2016.06.15
WO 2016030230 A1,2016.03.03
US 2011119467 A1,2011.05.19
US 2015261702 A1,2015.09.17
US 2016283841 A1,2016.09.29 (续)
审查员 向奎

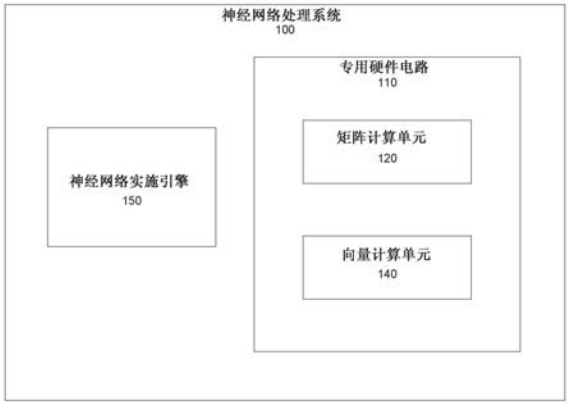
权利要求书3页 说明书16页 附图10页

(54)发明名称

在硬件中执行核心跨越

(57)摘要

本申请涉及在硬件中执行核心跨越的方法、系统和存储介质。该方法用于接收请求以在硬件电路上处理神经网络,所述神经网络包括具有大于1的步长的第一卷积神经网络层,和作为响应产生指令,所述指令引起所述硬件电路在处理输入张量期间通过执行下列操作产生等效于所述第一卷积神经网络层的输出的层输出张量,所述操作包括:使用具有等于1否则等效于所述第一卷积神经网络层的步长的第二卷积神经网络层处理所述输入张量以产生第一张量;将如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长则本就不会被生成的所述第一张量的元素归零以生成第二张量;以及在所述第二张量上执行最大池化以生成所述层输出张量。



CN 108073983 B

[转续页]

[接上页]

(56)对比文件

Chen Zhang 等.Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks.《FPGA 15:Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays》.2015,第161-170页.

Clement Farabet 等.CNP: An FPGA-based processor for Convolutional Networks.《2009 International Conference on Field Programmable Logic and Applications》.2009,第32-37页.

1.一种计算机实现的方法,包括:

接收在专用硬件电路上实现卷积神经网络的请求,并且通过使所述专用硬件电路执行指令来接收和处理神经网络输入,所述神经网络包括具有大于1的步长的第一卷积神经网络层,所述专用硬件电路是用于执行神经网络计算的集成电路并且包括矩阵计算单元和包括池化电路的向量计算单元,所述矩阵计算单元适合于执行向量矩阵乘法,所述池化电路适合于在所述矩阵计算单元的输出上执行池化;以及

响应于接收到所述请求,产生指令,所述指令在被所述专用硬件电路执行时使所述专用硬件电路在所述神经网络处理输入张量期间通过执行下列操作产生等效于所述第一卷积神经网络层的输出的层输出张量:

所述矩阵计算单元使用第二卷积神经网络层来处理对所述第一卷积神经网络层的所述输入张量以产生第一张量,所述第二卷积神经网络层具有等于1的步长但除此以外等效于所述第一卷积神经网络层;

所述向量计算单元将如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长则本就不会被生成的所述第一张量的元素归零以生成第二张量;以及

所述向量计算单元的所述池化电路在所述第二张量上执行最大池化以生成所述层输出张量,

其中所述向量计算单元将所述第一张量的元素归零包括:

所述向量计算单元执行掩蔽张量和所述第一张量的元素乘法以生成所述第二张量,其中所述掩蔽张量包括(i)在对应于如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长则本就不会被生成的所述第一张量的元素的所述掩蔽张量的每个元素位置处的零,和(ii)在所述掩蔽张量的每个其它元素位置处的1。

2.根据权利要求1所述的方法,其中所述掩蔽张量被存储在可由所述专用硬件电路访问的存储器处。

3.根据权利要求2所述的方法,其中在所述存储器处存储多个掩蔽张量以相应地对应于大于1的多个步长,以及

所述方法进一步包括在所述多个掩蔽张量当中选择与所述第一卷积神经网络层的步长相对应的掩蔽张量。

4.根据权利要求1所述的方法,其中所述向量计算单元的所述池化电路执行最大池化包括对于由所述第一卷积神经网络层的步长限定的所述第二张量的一个或多个窗口中的每个窗口都获得处于所述窗口内的元素的最大值元素。

5.根据权利要求4所述的方法,其中所述第二张量的一个或多个窗口中的每个窗口都为具有对应于所述第一卷积神经网络层的步长的尺寸的矩形窗口,并且包括所述第二张量的不同元素。

6.根据权利要求1所述的方法,其中所述向量计算单元的所述池化电路执行最大池化包括:对于所述第二张量的元素的一个或多个子集中的每个子集都获得该子集的最大值元素。

7.根据权利要求1所述的方法,其中所述输入张量是包括对应于数字图像的像素的元素的所述数字图像的表示。

8.根据权利要求1所述的方法,其中所述输入张量被存储在所述专用硬件电路的统一

的缓冲器中，并且所述第二卷积神经网络层的权重被存储在所述专用硬件电路的动态存储器中，并且其中使用所述第二卷积神经网络层处理对所述第一卷积神经网络层的所述输入张量包括：

将所述输入张量从所述统一的缓冲器发送到所述矩阵计算单元；

将所述第二卷积神经网络层的权重从所述动态存储器发送到所述矩阵计算单元；以及

由所述矩阵计算单元，使用所述第二卷积神经网络层的权重处理所述输入张量以生成所述第一张量。

9.一种系统，包括：

专用硬件电路，所述专用硬件电路是用于执行神经网络计算的集成电路并且包括矩阵计算单元和包括池化电路的向量计算单元，所述矩阵计算单元适合于执行向量矩阵乘法，所述池化电路适合于在所述矩阵计算单元的输出上执行池化；以及

一个或多个存储装置，所述一个或多个存储装置存储指令，所述指令在被所述专用硬件电路执行时使所述专用硬件电路执行根据权利要求1至8中的任一项所述的方法。

10.包括指令的一个或多个计算机可读存储介质，所述指令在被一个或多个计算机执行时使所述一个或多个计算机执行根据权利要求1至8中的任一项所述的方法。

11.一种计算机实现的方法，包括：

接收在专用硬件电路上实现卷积神经网络的请求，并且通过使所述专用硬件电路执行指令来接收和处理神经网络输入，所述神经网络包括具有大于1的步长的第一卷积神经网络层，所述专用硬件电路是用于执行神经网络计算的集成电路并且包括矩阵计算单元和包括池化电路的向量计算单元，所述矩阵计算单元适合于执行向量矩阵乘法，所述池化电路适合于在所述矩阵计算单元的输出上执行池化；以及

响应于接收到所述请求，产生指令，所述指令在被所述专用硬件电路执行时使所述专用硬件电路在所述神经网络处理输入张量期间通过执行下列操作产生等效于所述第一卷积神经网络层的输出的层输出张量：

所述矩阵计算单元使用第二卷积神经网络层来处理对所述第一卷积神经网络层的所述输入张量以产生第一张量，所述第二卷积神经网络层具有等于1的步长但除此以外等效于所述第一卷积神经网络层；

所述向量计算单元将如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长则本就不会被生成的所述第一张量的元素归零以生成第二张量；以及

所述向量计算单元的所述池化电路在所述第二张量上执行最大池化以生成所述层输出张量，

其中所述向量计算单元对所述第一张量的元素归零包括：

所述向量计算单元执行第一掩蔽张量和所述第一张量的元素乘法以生成修改的第一张量，其中所述第一掩蔽张量包括(i)在对应于如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长则本就不会被生成的所述第一张量的元素的所述第一掩蔽张量的每个元素位置处的零，和(ii)在对应于如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长将生成的所述第一张量的元素的所述第一掩蔽张量的每个元素位置处的相应的非零值；以及

所述向量计算单元执行第二掩蔽张量和所述修改的第一张量的元素乘法，其中所述第

二掩蔽张量在对应于如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长则本就会被生成的所述第一张量的元素的每个元素位置处都包括所述第一掩蔽张量的相应的非零值的逆。

12.根据权利要求11所述的方法，其中所述掩蔽张量被存储在可由所述专用硬件电路访问的存储器处。

13.根据权利要求12所述的方法，其中在所述存储器处存储多个掩蔽张量以相应地对应于大于1的多个步长，以及

所述方法进一步包括在所述多个掩蔽张量当中选择与所述第一卷积神经网络层的步长相对应的掩蔽张量。

14.根据权利要求11所述的方法，其中所述向量计算单元的所述池化电路执行最大池化包括对于由所述第一卷积神经网络层的步长限定的所述第二张量的一个或多个窗口中的每个窗口都获得处于所述窗口内的元素的最大值元素。

15.根据权利要求14所述的方法，其中所述第二张量的一个或多个窗口中的每个窗口都为具有对应于所述第一卷积神经网络层的步长的尺寸的矩形窗口，并且包括所述第二张量的不同元素。

16.根据权利要求11所述的方法，其中所述向量计算单元的所述池化电路执行最大池化包括：对于所述第二张量的元素的一个或多个子集中的每个子集都获得该子集的最大值元素。

17.根据权利要求11所述的方法，其中所述输入张量是包括对应于数字图像的像素的元素的所述数字图像的表示。

18.根据权利要求11所述的方法，其中所述输入张量被存储在所述专用硬件电路的统一的缓冲器中，并且所述第二卷积神经网络层的权重被存储在所述专用硬件电路的动态存储器中，并且其中使用所述第二卷积神经网络层处理对所述第一卷积神经网络层的所述输入张量包括：

将所述输入张量从所述统一的缓冲器发送到所述矩阵计算单元；

将所述第二卷积神经网络层的权重从所述动态存储器发送到所述矩阵计算单元；以及

由所述矩阵计算单元，使用所述第二卷积神经网络层的权重处理所述输入张量以生成所述第一张量。

19.一种系统，包括：

专用硬件电路，所述专用硬件电路是用于执行神经网络计算的集成电路并且包括矩阵计算单元和包括池化电路的向量计算单元，所述矩阵计算单元适合于执行向量矩阵乘法，所述池化电路适合于在所述矩阵计算单元的输出上执行池化；以及

一个或多个存储装置，所述一个或多个存储装置存储指令，所述指令在被所述专用硬件电路执行时使所述专用硬件电路执行根据权利要求11至18中的任一项所述的方法。

20.包括指令的一个或多个计算机可读存储介质，所述指令在被一个或多个计算机执行时使所述一个或多个计算机执行根据权利要求11至18中的任一项所述的方法。

在硬件中执行核心跨越

技术领域

[0001] 本说明书涉及计算硬件中的神经网络推理。

背景技术

[0002] 神经网络是采用一层或多层为所接收的输入生成输出，例如分类的机器学习模型。除了输出层之外，一些神经网络还包括一个或多个隐藏层。每个隐藏层的输出都被用作网络中另一层，例如网络的下一个隐藏层或输出层的输入。网络的每个层都根据相应参数集的当前值从所接收的输入生成输出。

发明内容

[0003] 一般而言，本说明书描述了一种计算神经网络推理的专用硬件电路。

[0004] 一般而言，本说明书中所述的主题的一个创新方面能够被体现为用于接收在硬件电路上处理神经网络的请求的方法和系统，该神经网络包括具有大于1的步长的第一卷积神经网络层，并且作为响应产生指令，当被硬件电路执行时，该指令引起硬件电路在神经网络处理输入张量期间通过执行下列操作产生等效于第一卷积神经网络层的输出的层输出张量，这些操作包括使用具有等于1否则等效于第一卷积神经网络层的步长的第二卷积神经网络层处理对第一卷积神经网络层的输入张量以产生第一张量，将如果第二卷积神经网络层具有第一卷积神经网络层的步长则本就不会被生成的第一张量的元素归零以生成第二张量，以及在第二张量上执行最大池化以生成层输出张量。

[0005] 实施能够包括一个或多个下列特征。在一些实施中，将第一张量的元素归零包括将第一张量的元素子集乘以0，并且将该子集中未包括的第一张量的元素乘以1。将第一张量的元素归零包括执行掩蔽张量(masking tensor)和第一张量的元素乘法(element-wise multiplication)以生成第二张量，其中掩蔽张量包括(i)在对应于如果第二卷积神经网络层具有第一卷积神经网络层的步长则本就不会被生成的第一张量的元素的掩蔽张量的每个元素位置处的零，和(ii)在掩蔽张量的每个其它元素位置处的1。在一些实施中，掩蔽张量被存储在可由硬件电路访问的存储器处，并且其中由在硬件电路中所包括的硬件中实施的向量计算单元执行掩蔽张量和第一张量的元素乘法。

[0006] 实施还能够包括一个或多个下列特征。在一些实施中，对第一张量的元素归零包括执行第一掩蔽张量和第一张量的元素乘法以生成修改的第一张量，其中第一掩蔽张量包括(i)在对应于如果第二卷积神经网络层具有第一卷积神经网络层的步长则本就不会被生成的第一张量的元素的掩蔽张量的每个元素位置处的零，和(ii)在对应于如果第二卷积神经网络层具有第一卷积神经网络层的步长则本就会被生成的第一张量的元素的掩蔽张量的每个元素位置处的相应的非零值，并且执行第二掩蔽张量和修改的第一张量的元素乘法，其中第二掩蔽张量在对应于如果第二卷积神经网络层具有第一卷积神经网络层的步长将生成的第一张量的元素的每个元素位置处都包括第一掩蔽张量的相应的非零值的逆。

[0007] 实施还能够包括一个或多个下列特征。在一些实施中，执行最大池化包括对于由

第一卷积神经网络层的步长限定的第二张量的一个或多个窗口中的每个窗口都获得处于该窗口内的元素的最大值元素。第二张量的一个或多个窗口中的每个窗口都为具有对应于卷积神经网络层的步长的尺寸的矩形窗口，并且包括第二张量的不同元素。在一些实施中，执行最大池化包括对于第二张量的元素的一个或多个子集中的每个子集都获得子集的最大值元素。由硬件电路的池化电路执行在第二张量上执行的最大池化。卷积神经网络层是神经网络中的第一神经网络层，并且其中输入张量是包括对应于数字图像的像素的元素的数字图像的表示。

[0008] 实施还能够包括一个或多个下列特征。在一些实施中，输入张量被存储在硬件电路的统一的缓冲器中，并且第二卷积神经网络层的权重被存储在硬件电路的动态存储器中，并且其中使用第二卷积神经网络层处理对第一卷积神经网络层的输入张量包括将输入张量从统一的缓冲器发送到在硬件中实施的硬件电路的矩阵计算单元，将第二卷积神经网络层的权重从动态存储器发送到硬件电路的矩阵计算单元，并由硬件电路的矩阵计算单元，使用第二卷积神经网络层的权重处理输入张量以生成第一张量。

[0009] 本说明书中所述的主题的特殊实施例能够被实施为实现一个或多个下列优点。甚至在硬件电路不能使用具有大于1的步长的卷积神经网络直接地处理输入张量时，也能够由专用硬件电路在硬件中生成对应于具有大于1的步长的卷积神经网络层的输出张量。通过使用专用硬件电路生成适当的输出，能够不将数据传回主机，即不执行至少部分片外计算地执行对具有大于1的步长的神经网络层的处理，即使专用硬件电路直接地支持这种处理也是如此。这允许不改变专用硬件电路的硬件架构地，高效地确定包括具有大于1的步长的卷积层的神经网络的推断。即，避免了在片外，在软件中或两者中执行部分处理导致的处理延迟。

[0010] 本说明书中所述的主题例如也涉及一种在计算神经网络推理时，使用所公开的技术和硬件执行核心跨越的图像识别或分类方法和系统。

[0011] 在附图和下面的描述中提出本说明书的主题的一个或多个实施例的细节。通过说明、附图和权利要求，主题的其它特征、方面和优点将变得显而易见。

附图说明

[0012] 图1示出示例神经网络处理系统。

[0013] 图2是用于对给定神经网络层执行计算的示例方法的流程图。

[0014] 图3示出示例神经网络处理系统。

[0015] 图4示出包括矩阵计算单元的示例架构。

[0016] 图5示出脉动阵列内部的cell的示例架构。

[0017] 图6示出向量计算单元的示例架构。

[0018] 图7示出池化电路的示例架构。

[0019] 图8是指令神经网络处理系统对具有大于1的步长的给定神经网络层执行计算的示例方法的流程图。

[0020] 图9是用于对具有大于1的步长的给定神经网络层执行计算的示例方法的流程图。

[0021] 图10是具有大于1的步长的给定神经网络层的计算示例。各附图中的相同附图标记和名称指示相同元件。

具体实施方式

[0022] 能够使用具有多个层的神经网络来计算推理。例如，给定输入，神经网络能够计算输入的推理。神经网络通过处理经过神经网络层的每个的输入计算该推理。每层都接收输入，并根据该层的权重集处理输入以生成输出。

[0023] 因此，为了从所接收的输入计算推理，神经网络接收输入并通过每个神经网络层对其处理以生成推理，一个神经网络层的输出被作为输入提供给下一个神经网络层。对神经网络层的数据输入，例如对神经网络的输入，或者对神经网络层的按顺序处于该层之下的层的输出能够被称为对该层的激活输入。

[0024] 在一些实施中，神经网络层被按顺序布置。在一些其它实施中，层被作为有向图加以布置。即，任何特殊层都能够接收多个输入、多个输出或者两者。神经网络层也能够被布置成层的输出能够作为输入被发回至上先前层。

[0025] 一些神经网络将来自一个或多个神经网络层的输出池化以生成被用作后续神经网络层的输入的池化的值。在一些实施中，神经网络通过确定输出组的最大值、最小值或平均值并使用最大值、最小值或平均值作为该组的池化输出来池化一组输出。池化输出能够保持一定的空间不变性，使得以各种配置排列的输出都能够被处理以具有相同的推理。池化输出也能够降低在后续神经网络层处接收的输入的维度，同时在池化之前保持输出的期望特性，这能够提高效率而不显著地损害由神经网络生成的推理的质量。

[0026] 一些神经网络包括具有大于1的步长的一个或多个卷积神经网络层。在概念上，对于步长1，卷积神经网络层能够依次将权重集应用于激活输入。即，对于激活输入阵列，权重能够被应用于激活输入的子集，并且被移动一个位置，例如一行或一列到达激活输入的每个其它子集，直到卷积计算完成。对于具有大于1的步长的卷积神经网络层，其中步长是整数，权重能够被应用于激活输入的子集，并且被移动等效于步长的多个位置，例如步长指示的行或列数，到达激活输入的每个其它子集，直到卷积计算完成。

[0027] 本说明书描述了处理神经网络层的专用硬件电路，并且可选地在一个或多个神经网络层的输出上执行池化。专用硬件电路包括能够处理具有步长1的神经网络层的电路。虽然专用硬件电路不直接支持处理具有大于1的步长的神经网络层，但是可控制专用硬件电路以生成等效于具有大于或等于1的步长的神经网络层的输出的输出。因而，所公开技术的一种技术效果和优点是能够以更灵活的方式使用能够处理具有步长1的神经网络层的电路，并且将其用于计算具有大于1的步长的神经网络层的神经网络推理。

[0028] 图1示出了示例神经网络处理系统100。神经网络处理系统100是在能够实施下文所述的系统、部件和技术的一个或多个位置被实施为一个或多个计算机的系统的示例。

[0029] 神经网络处理系统100是使用专用硬件电路110执行神经网络计算的系统。硬件电路110是用于执行神经网络计算的集成电路，并且包括在硬件中执行向量矩阵乘法的矩阵计算单元120。硬件电路110还包括向量计算单元140，其包括用于在矩阵计算单元120的输出上执行池化的池化电路。下面参照图3更详细地描述示例专用硬件电路120。

[0030] 特别地，神经网络处理系统100接收在专用硬件电路110上实施神经网络的请求，在专用硬件电路110上实施神经网络，并且一旦实施了给定的神经网络，就使用专用硬件电路110处理对神经网络的输入来生成神经网络推理。

[0031] 即，神经网络处理系统100能够接收指定用于将被用于处理输入的神经网络的神

经网络架构的请求。神经网络架构定义了神经网络中层的数量和配置，以及具有参数的每个层的参数值。

[0032] 为了在专用硬件电路110上实施神经网络，神经网络处理系统100包括神经网络实施引擎150，该引擎被实施为在一个或多个物理位置中的一个或多个计算机上的一个或多个计算机程序。

[0033] 神经网络实施引擎150产生指令，当被专用硬件电路110执行时，该指令引起硬件电路110执行由神经网络指定的操作以从所接收的神经网络输入生成神经网络输出。

[0034] 一旦指令已经由神经网络实施引擎150生成，并且被提供给硬件电路110，则神经网络处理系统100能够接收神经网络输入，并且能够通过引起硬件电路110执行所生成的指令而使用神经网络来处理神经网络输入。

[0035] 然而，一些神经网络包括一个或多个不兼容的神经网络层。本说明书中使用的术语“不兼容的神经网络层”涉及指定不能由专用硬件电路110在硬件中直接执行的操作的神经网络层。为了在硬件电路110上实现这些神经网络，神经网络实施引擎150生成指令，当被硬件电路110执行时，该指令通过在硬件中执行与神经网络层指定的那些不同，但是导致生成满足不可兼容神经网络层的规范的层输出，例如层输出张量，即与通过直接地执行层指定的操作本就会被生成的输出相同的层输出，而生成用于不可兼容神经网络层的输出。

[0036] 特别地，一些神经网络包括具有大于1的步长的卷积神经网络层。这种神经网络层特征在于利用输入张量非顺序地处理的一个或者多个核心。例如，当以步长1执行核心跨越时，核心被按顺序地应用于输入张量的元素。然而，当以步长2执行核心跨越时，神经网络层的核心被移位，使得核心的特殊元素被应用于输入张量的每个其它元素以生成输出张量。然后，输出张量能够被另一神经网络层用作输入。

[0037] 由于在硬件电路110上执行矩阵运算的主要硬件单元是矩阵计算单元120，因此集成电路不能直接计算具有大于1的步长的神经网络层。为了实施包括具有大于1的步长的层的神经网络，神经网络实施引擎150产生指令，当在由神经网络处理神经网络输入期间由专用硬件电路110执行时，该指令引起硬件电路110在硬件中执行其它操作，以使用特征化池化电路的矩阵计算单元120和向量计算单元140生成满足具有大于1的步长的神经网络层的规范的输出张量。下文参考图7-10更详细地描述这些指令和其它操作。

[0038] 图2是使用专用硬件电路对给定神经网络层执行计算的示例过程200的流程图。为了方便，将关于具有执行方法200的一个或多个电路的系统来描述方法200。能够对神经网络层的每个都执行方法200，以便从所接收的输入计算推理。

[0039] 系统接收权重输入集(步骤202)，和给定层的激活输入集(步骤204)。能够分别从专用硬件电路的动态存储器和统一的缓冲器接收权重输入集和激活输入集。在一些实施中，能够从统一的缓冲器接收权重输入集和激活输入集两者。

[0040] 系统使用专用硬件电路的矩阵乘法单元从权重输入和激活输入生成累积的值(步骤206)。在一些实施中，累积的值是权重输入集和激活输入集的点积。即，对于作为层中所有权重的子集的一组权重，系统能够将每个权重输入都与每个激活输入相乘，并将乘积加在一起以形成累积的值。然后，系统能够计算权重的其它集合与激活输入的其它集合的点积。在一些实施中，专用硬件电路可与特殊神经网络层的补偿无关地，即与神经网络层是否具有为1的步长或大于1的步长无关地类似地执行这些操作。能够执行对来自矩阵乘法单元

的输出的后续处理，以生成等效于如果以大于1的指定步长处理神经网络层则本就会被生成的输出的输出。

[0041] 系统能够使用专用硬件电路的向量计算单元从累积值生成层输出(步骤208)。在一些实施中，向量计算单元对累积的值应用激活函数，这将在下文参考图5进一步描述。该层的输出能够被存储在统一的缓冲器中，用作神经网络中的后续层的输入，或者能够用于确定推理。在一些实施中，神经网络层能够指定大于1的步长，并且系统可对累积值执行附加处理，以获得等效于具有大于1的步长的神经网络层的输出的层输出。当所接收的输入已经通过神经网络层的每个进行处理以生成所接收的输入的推理时，系统完成对神经网络的处理。

[0042] 图3示出用于执行神经网络计算的示例专用硬件电路300。系统300包括主机接口302。主机接口302能够接收包括用于神经网络计算的参数的指令。这些参数能够包括以下一个或多个参数：应该处理多少层、模型的每个层的相应权重输入集、初始激活输入集，即待从其中计算推理的对神经网络的输入、每层的相应输入和输出大小、用于神经网络计算的步长值，以及待处理的层的类型，例如卷积层或完全连接的层。

[0043] 主机接口302能够将指令发送到定序器306，定序器306将指令转换成控制电路执行神经网络计算的低电平控制信号。在一些实施中，控制信号调节电路中的数据流，例如，权重输入集和激活输入集如何流经电路。定序器306能够将控制信号发送到统一的缓冲器308、矩阵计算单元312和向量计算单元314。在一些实施中，定序器306也向直接存储器访问引擎304和动态存储器310发送控制信号。在一些实施中，定序器306是生成控制信号的处理器。定序器306能够在适当时间使用控制信号的定时，以将控制信号发送至电路300的每个部件。在一些其它实施中，主机接口302从外部处理器传入控制信号。

[0044] 主机接口302能够向直接存储器访问引擎304发送权重输入集和初始激活输入集。直接存储器访问引擎304能够在统一的缓冲器308处存储激活输入集。在一些实施中，直接存储器访问将权重集存储到能够为存储器单元的动态存储器310。在一些实施中，动态存储器310位于电路之外。

[0045] 统一的缓冲器308是存储器缓冲器。它能够被用于存储来自直接存储器访问引擎304的激活输入集和向量计算单元314的输出。下面将参考图6更详细地描述向量计算单元314。直接存储器访问引擎304也能够从统一的缓冲器308读取向量计算单元314的输出。

[0046] 动态存储器310和统一的缓冲器308能够分别向矩阵计算单元312发送权重输入集和激活输入集。在一些实施中，矩阵计算单元312是二维脉动阵列。矩阵计算单元312也能够为能够执行数学运算，例如乘法和加法的一维脉动阵列或其它电路。在一些实施中，矩阵计算单元312是通用矩阵处理器。

[0047] 矩阵计算单元312能够处理权重输入和激活输入，并向向量计算单元314提供输出的向量。在一些实施中，矩阵计算单元312将输出的向量发送到统一的缓冲器308，统一的缓冲器308将输出的向量发送至向量计算单元314。向量计算单元314能够处理输出的向量并将经处理的输出的向量存储到统一的缓冲器308。对于具有大于1步长的神经网络层，向量计算单元314能够处理输出的向量以生成等效于具有大于1的步长的神经网络层的输出的层输出张量，并且能够将层输出张量存储在统一的缓冲器308处。经处理的输出的向量能够被用作矩阵计算单元312的激活输入，例如用于神经网络中的后续层。下面将分别参考图4

和图6更详细地描述矩阵计算单元312和向量计算单元314。

[0048] 图4示出了包括矩阵计算单元的示例架构400。矩阵计算单元是二维脉动阵列406。阵列406包括多个cell 404。在一些实施中，脉动阵列406的第一维度420对应于cell的列，并且脉动阵列406的第二维度422对应于cell的行。脉动阵列能够具有比列更多的行、比行更多的列或相等数量的列和行。

[0049] 在所示示例中，值加载器402将激活输入发送到阵列406的行，并且权重提取器接口408将权重输入发送到阵列406的列。然而，在一些其它实施中，激活输入被传送到列，并且权重输入被传送到阵列406的行。

[0050] 值加载器402能够从统一的缓冲器，例如图3的统一的缓冲器308接收激活输入。每个值加载器都能够将相应的激活输入发送到阵列406的不同最左边cell。例如，值加载器412能够向cell 414发送激活输入。

[0051] 权重提取器接口408能够从存储器单元，例如，图3的动态存储器310接收权重输入。权重提取器接口408能够将相应的权重输入发送到阵列406的不同最顶部cell。例如，权重提取器接口408能够向cell414和416发送权重输入。权重提取器接口408还能够从存储器单元，例如，动态存储器310接收多个权重，并且并行地将多个权重发送到阵列406的不同最顶部cell。例如，权重提取器接口408可同时向cell414和416发送不同的权重。

[0052] 在一些实施中，主机接口，例如图3的主机接口302贯穿阵列406沿着一个维度例如向右移动激活输入，同时贯穿阵列406沿着另一个维度，例如向底部移动权重输入。例如，在一个时钟周期内，cell 414处的激活输入能够移动到至cell 414右侧的cell 416中的激活寄存器。类似地，cell 416处的权重输入能够移动到cell 414之下的cell 418处的权重寄存器。

[0053] 在每个时钟周期，每个cell都能够处理给定的权重输入、给定的激活输入和来自相邻cell的累积的输出以生成累积的输出。累积的输出也能够作为给定权重输入沿相同维度传递给相邻cell。每个cell也可处理给定的权重输入和给定的激活输入以生成输出，不处理来自相邻cell的累积的输出。输出能够作为给定权重输入和未累积的输出，沿相同维度传递至相邻cell。下面参考图5进一步描述个别cell。

[0054] 在一些实施中，能够将单位矩阵，即在主对角线上为1其余为0的矩阵传递到阵列406，由此将值装载器402处的输入传递到累加器410而不进行修改。这可用于执行两个输入的元素乘法，其中在累加器处的第一输出能够被表示为output＝MatMul(input1,identity)，其中MatMul是用于矩阵计算单元执行矩阵乘法的指令，并且对应于元素乘法结果的第二输出被表示为output*＝MatMul(input2,identity)。为了执行*＝运算，即运算output＝output*MatMul(input2,identity)，架构400可包括用于执行+＝或*＝计算的部件。用于执行+＝或*＝运算的部件可位于累加器410之前，即，在cell 404的最后一行之后。在一些实施中，图3的向量计算单元314可包括用于执行+＝或*＝运算的部件，即，其中向量计算单元314执行output*＝MatMul(input2,identity)运算以执行元素乘法。

[0055] 累积的输出能够沿着与权重输入相同的列，例如朝向阵列406中的列的底部传递。在一些实施中，在每列的底部，阵列406都能够包括累加器单元410，其当利用具有比行更多激活输入的层执行计算时，存储并且累加来自每一列的每个累积的输出。在一些实施中，每个累加器单元都存储多个并行的累积。累加器单元410能够累积每个累积的输出以生成最

终累积的值。最终累积的值能够被传送到向量计算单元，例如图6的向量计算单元。在一些其它实施中，当利用具有比行更少的激活输入的层处理层时，累加器单元410将累积的值传递到向量计算单元，不执行任何累加。

[0056] 图5示出脉动阵列内部的cell，例如图4的脉动阵列406的cell414、416或418其中之一的示例架构500。

[0057] cell能够包括存储激活输入的激活寄存器506。激活寄存器能够取决于cell在脉动阵列内的位置，从左侧相邻cell，即位于给定cell左侧的相邻cell，或者从统一的缓冲器接收激活输入。cell能够包括存储权重输入的权重寄存器502。能够取决于脉动阵列内的cell的位置，从顶部相邻cell或从权重提取器接口传送权重输入。cell也能够包括寄存器504中的和。寄存器504中的和能够存储来自顶部相邻cell的累积的值。乘法电路508能够用于将来自权重寄存器502的权重输入与来自激活寄存器506的激活输入相乘。乘法电路508能够将乘积输出到求和电路510。

[0058] 求和电路510能够将乘积和来自寄存器504中的和的累积的值相加以生成新的累积的值。然后，求和电路510能够将新的累积的值发送到位于底部相邻cell中的寄存器中的另一个和。新的累积的值能够被用作底部相邻单元中的求和的操作数。求和电路510也能够接收来自寄存器504中的和的值，并将该值从寄存器504中的和发送到底部相邻cell，不将寄存器504中的和与来自乘法电路508的乘积相加。

[0059] cell也能够将权重输入和激活输入移动到相邻的cell进行处理。例如，权重路径寄存器512能够将权重输入发送到底部相邻cell中的另一个权重寄存器。激活寄存器506能够将激活输入发送到右侧相邻cell中的另一激活寄存器。因此，权重输入和激活输入两者能够在随后的时钟周期被阵列中的其它cell重新使用。

[0060] 在一些实施中，cell也包括控制寄存器。控制寄存器能够存储控制信号，该控制信号确定cell是否应将权重输入或激活输入移动到相邻cell。在一些实施中，移动权重输入或激活输入需要一个或多个时钟周期。控制信号也能够确定激活输入或权重输入是否被传送到乘法电路508，或者能够确定乘法电路508是否对激活和权重输入进行操作。控制信号也能够被传递到一个或多个相邻的cell，例如使用电线。

[0061] 在一些实施中，权重被预先移入权重路径寄存器512中。权重路径寄存器512能够例如从顶部相邻cell接收权重输入，并且基于控制信号将权重输入传送到权重寄存器502。权重寄存器502能够静态地存储权重输入，使得随着激活输入在多个时钟周期上例如通过激活寄存器506被传送到cell，权重输入保持在cell内并且不被传送到相邻cell。因此，权重输入能够例如使用乘法电路508被应用于多个激活输入，并且相应的累积的值能够被传送到相邻cell。

[0062] 图6示出了向量计算单元602的示例架构600。向量计算单元602能够从矩阵计算单元，例如，参考图3所述的矩阵计算单元312，或图4的矩阵计算单元的累加器410接收累积的值的向量。

[0063] 向量计算单元602能够在激活单元604处处理累积的值的向量。在一些实施中，激活单元包括将非线性函数应用于每个累积的值以生成激活值的电路。例如，非线性函数能够为tanh(x)，其中x是累积的值。

[0064] 可选地，向量计算单元602能够使用池化电路608池化值，例如激活值。池化电路

608能够将聚合函数应用于一个或多个值以生成池化的值。在一些实施中，聚合函数是返回值的最大值，最小值或平均值或者值的子集的函数。

[0065] 控制信号610能够例如由图3的定序器306传送，并且能够调节向量计算单元602如何处理累积的值的向量。即，控制信号610能够调节激活值是否被池化，其中激活值被存储在例如统一的缓冲器308中，或者否则能够调节激活值的处理。控制信号610也能够指定激活或池化函数，以及用于处理激活值或池化值，例如步长值的其它参数。

[0066] 向量计算单元602能够向统一的缓冲器，例如图3的统一的缓冲器308发送值，例如激活值或池化的值。在一些实施中，池化电路608接收激活值或池化的值，并将激活值或池化的值存储在统一的缓冲器中。

[0067] 图7示出了用于池化电路的示例架构700。池化电路能够将聚合函数应用于一个或多个激活的值以产生池化的值。作为例示，架构700能够执行激活的值的4×4集合的池化。虽然图7中所示的池化具有正方形区域，即4×4，但矩形区域也是可能的。例如，如果区域具有n×m的窗口，则架构700能够具有n*m个寄存器，即n列和m行。

[0068] 池化电路架构700能够来值的向量，例如来自图6的激活电路604接收一系列元素。例如，该序列能够代表图像的8×8部分的像素，并且池化电路架构700能够从8×8部分的4×4子集池化多个值。在一些实施中，池化的值被附加到由池化电路架构700一次计算出的序列。在一些实施中，神经网络处理器包括多个并行池化电路。在每个时钟周期中，每个池化电路都能够从来自激活电路604的值的向量接收相应的元素。每个池化电路都能够将从激活电路604接收到的元素解释为以光栅顺序到达的二维图像。

[0069] 池化电路能够包括一系列寄存器和存储器单元。每个寄存器都能够向聚合电路706发送输出，聚合电路706跨寄存器内存储的值应用聚合函数。聚合函数能够从一组值返回最小值、最大值或平均值。

[0070] 第一值能够被发送至寄存器702并存储在寄存器702中。在随后的时钟周期中，第一值能够移动到之后的寄存器708并被存储在存储器704中，并且能够将第二值发送到寄存器702并存储在寄存器702内。

[0071] 在四个时钟周期之后，四个值被存储在前四个寄存器702、708-712内。在一些实施中，存储器单元704按先进先出(FIFO)操作。每个存储器单元能够存储达八个值。在存储器单元704包含完整像素行之后，存储器单元704能够向寄存器714发送值。

[0072] 在任何给定的时间点，聚合电路706能够访问每个寄存器的值。寄存器中的值应代表图像的4×4部分的值。

[0073] 池化电路能够通过使用聚合电路706，从访问的值，例如最大值，最小值或平均值生成池化的值。池化的值能够被发送到统一的缓冲器，例如图3的统一的缓冲器308。

[0074] 在生成第一池化的值之后，池化电路能够通过将值移动通过每个寄存器来继续生成池化的值，使得新值被存储在寄存器中并且能够被聚合电路706池化。例如，在架构700中，池化电路能够在4个更多的时钟周期上移动多个值，由此将存储器单元中的值移入寄存器。在一些实施中，池化电路移动新值，直到新值被存储在最后一个最顶部寄存器，例如寄存器716中。

[0075] 然后，聚合电路706能够池化被存储在寄存器中的新值。对新值池化的结果能够被存储在统一的缓冲器中。

[0076] 图8是用于对具有大于1的步长的神经网络的给定卷积层执行计算的示例过程800的流程图。通常，过程700由包括专用硬件电路的一个或多个计算机的系统执行。在一些实施中，示例过程800能够由图1的系统执行。

[0077] 系统接收在专用硬件电路上实施神经网络的请求(步骤802)。特别地，神经网络包括具有大于1的步长的卷积神经网络层。该请求还可指定用于实施神经网络的其它参数，诸如使用神经网络进行处理的输入、存储由神经网络生成的输出张量的位置或其它参数。

[0078] 系统基于将在处理具有大于1的步长的神经网络层中使用的请求生成掩蔽张量(步骤804)。例如，基于接收到实施神经网络的请求和指定对神经网络的输入的信息，系统生成用于处理具有大于1的步长的神经网络层的掩蔽张量。

[0079] 可基于指定输入的尺寸或对具有大于1的步长的神经网络层的输入张量的预期大小确定掩蔽张量的大小。可基于步长大于1的神经网络层的指定步长确定掩蔽张量中包括的值。例如，如果神经网络层具有指定步长4，则掩蔽张量的每个第四个元素可被设置为1，而掩蔽张量的所有其它项都可被设为零。在一些实施中，神经网络可包括具有大于1的步长的多个层，并且系统可对具有大于1的步长的每一层都生成相应的掩蔽张量。另外，在一些实施中，系统例如可在存储器中存储掩蔽矩阵或掩蔽矩阵部件库，并且可基于使用库选择或生成掩蔽矩阵。

[0080] 系统产生指令，当被专用硬件电路110执行时，这些指令引起专用硬件电路110在由神经网络处理输入张量期间使用掩蔽张量生成层输出张量，该层输出张量等效于具有大于1的步长的卷积神经网络层的输出(步骤806)。例如，响应于请求，神经网络实施引擎150能够生成指令，这些指令指导或控制专用硬件电路110生成输出张量，即输出向量，该输出张量等效于如果专用硬件电路110使用具有大于1的步长的卷积神经网络层处理输入张量。

[0081] 系统将指令和掩蔽张量发送到专用硬件电路110(步骤808)。例如，神经网络实施引擎150能够向专用硬件电路110提供指令，并且专用硬件电路110例如能够在图3的主机接口302处接收指令。神经网络实施引擎150也可提供用于神经网络计算的，也能够被主机接口302接收的其它指令和/或参数。

[0082] 图9是用于计算具有大于1的步长的神经网络计算层的示例过程900的流程图。例如，过程900能够由图1的专用硬件电路110基于从神经网络实施引擎150接收的指令执行。

[0083] 例如，一旦接收到用于实施具有大于1的步长的神经网络层的指令，则主机接口302能够将指令发送到图3的定序器306，并且定序器306能够将指令转换成控制图3的专用硬件电路300，以执行神经网络计算的低电平控制信号。

[0084] 基于所接收的指令，专用硬件电路300使用具有步长1的第二卷积神经网络层处理对卷积神经网络层的输入张量(步骤902)。例如，从接收到的指令生成的控制信号控制专用硬件电路300，以使用具有等效于1否则等效于卷积神经网络层的步长的第二卷积神经网络层处理输入张量，例如存储在统一的缓冲器308中的先前层的输出，或被指定或提供给专用硬件电路300的神经网络的输入，以生成卷积的张量。

[0085] 为了使用第二卷积神经网络层处理输入张量，控制信号可控制统一的缓冲器308以向图3的矩阵计算单元312提供输入张量，即可对应于神经网络的输入或先前神经网络的输出的激活输入。控制信号也可指令图3的直接存储器访问引擎304和/或动态存储器310向对应于具有步长1，即单一步长，否则等效于具有大于1的步长的神经网络层的第二神经网

络层的矩阵计算单元312提供权重。

[0086] 定序器306还可进一步生成指令，这些指令控制矩阵计算单元312使用权重，例如使用关于图3所述的过程处理输入张量。在一些实施中，矩阵计算单元312使用于2015年9月3日提交的美国专利申请No.14/844,738中所述的技术执行卷积，其公开内容通过引用整体并入本文。

[0087] 矩阵计算单元312基于控制信号执行计算，并向卷积计算单元314输出卷积的张量。例如，矩阵计算单元312向向量计算单元314发送矩阵计算单元312生成的输出的向量。输出的向量可基于使用对应于具有步长1，其否则等效于具有大于1的步长的神经网络层的神经网络层的权重来处理输入张量而加以确定。向量计算单元314能够在统一的缓冲器308处存储卷积的张量。

[0088] 在通过具有步长1的卷积神经网络层处理激活输入以生成卷积的张量之后，专用硬件电路300将如果第二卷积神经网络层具有包括步长大于1的卷积神经网络层的步长则本就不会被生成的元素归零(步骤904)。将元素归零通常涉及以零替换元素的当前值。消除，即使得值归零可通过执行卷积的张量与掩蔽张量，即由神经网络处理引擎150生成并且被发送至专用神经网络的掩蔽张量的元素乘法来实现。

[0089] 为了消除如果输入张量已经被具有指定步长的卷积神经网络层处理则本就不会被生成的卷积的张量的那些值，定序器306能够发送控制信号以控制矩阵乘法单元312，以执行卷积的张量与掩蔽张量的元素乘法。卷积的张量可被基于来自定序器306的其它控制信号从统一的缓冲器308发送到矩阵乘法单元312，并且掩蔽张量可基于从定序器306至直接存储器访问引擎304或动态存储器310的控制信号，即，在掩蔽张量已经被专用硬件电路300接收并存储在动态存储器310之后，被发送到矩阵计算单元312。

[0090] 通常，如关于图8所述，掩蔽张量是向量，其包括，处于对应于通过以具有大于1的步长的卷积神经网络层处理输入张量则会被生成的元素的元素位置中的单一值元素，即为1的值，并且包括在所有其它位置，即对应于利用具有大于1的步长的卷积神经网络层处理激活值则不会被生成的元素的位置中的零值元素。

[0091] 掩蔽张量例如可被存储在动态存储器310处，并且定序器306可发送控制信号，将掩蔽张量从动态存储器310发送至矩阵计算单元312。例如，提供给专用硬件电路300的指令可识别，例如提供掩蔽张量在动态存储器310中的位置，或可包括限定然后被存储在动态存储器310处的掩蔽张量的数据，并且定序器306可发送引起被存储在动态存储器310中的位置处的掩蔽张量被发送到矩阵计算单元312的控制信号。另外，定序器306可提供控制信号，从而引起被存储在统一的缓冲器308处的卷积的张量被提供给矩阵计算单元312。然后，矩阵计算单元312执行卷积的张量和掩蔽张量的元素乘法以生成修改的卷积的张量。修改的卷积的张量能够被向量计算单元314从矩阵计算单元312接收。向量计算单元314可以可选地在统一的缓冲器308中存储修改的卷积的张量。

[0092] 由于与掩蔽张量的元素乘法，修改的卷积的张量包括如果使用具有大于1的指定步长的神经网络层处理输入张量则本就会被输出的值。修改的卷积的张量包括在对应于使用如果利用具有指定步长的卷积神经网络处理输入张量则本就不会被输出的步长为1的卷积神经网络层在输入张量的计算中输出的值的位置中的零。在其它实施中，可采用其它对卷积的张量的元素归零的方法。例如，可在统一的缓冲器308中，或者以修改形式在另一存

储器中重写卷积的矩阵，其中对应于使用具有指定步长的卷积神经网络在输入张量的计算中输出的值的元素不变，而其它元素被写为零。

[0093] 向量计算单元314接收修改的卷积的张量并对修改的卷积的张量进行最大池化，以生成用于具有大于1的步长的卷积神经网络层的层输出张量(步骤906)。例如，向量计算单元314可从矩阵计算单元312接收修改的卷积的张量，并且使用池化电路608可在修改的卷积的张量上执行最大池化。最大池化是接收一组数据，并且对于数据的一个或多个子集中的每一个都输出子集中的元素的最大值的操作。在修改的卷积的张量上执行最大池化导致下列张量，其对修改的卷积的张量的多个元素子集中的每一个都包括子集的最大值。向量计算单元314可对基于卷积神经网络层的指定步长确定的修改的卷积的张量的窗口执行最大池化。例如，对于步长2，池化电路608将使用2×2窗口执行最大池化，以生成包括来自每个2×2窗口的最大值元素的层输出张量。对于具有步长为4的神经网络层，池化电路608将使用4×4窗口执行最大池化，以生成包括来自每个4×4窗口的最大值元素的层输出张量。最大池化操作的结果由向量计算单元314存储在统一的缓冲器308处，其中结果是输出张量，其等效于如果专用硬件电路300已经使用具有大于1的步长的神经网络层处理了输入张量则本就会被生成的输出。可使用层输出张量执行神经网络的后续层的处理，以最终获得神经网络的推理。

[0094] 图10示出了用于具有大于1的步长的给定神经网络层的计算的示例。可使用图7的过程和图2的专用硬件电路300执行图10的示例。作为例示，图10的示例将具有步长4的卷积神经网络层应用于8×8激活值阵列。卷积神经网络层可具有将被应用于8×8激活值阵列的权重的4×4核心。激活值可表示输入到神经网络的图像的8×8部分，即对应于图像的8×8部分的值序列。可替选地，激活值的8×8阵列可表示另一个输入张量的8×8部分，例如对应于神经网络的先前层的输出的输入张量。

[0095] 在图10的部分(a)中，使用具有步长1，其否则等效于具有步长大于1的卷积神经网络层的卷积神经网络层处理8×8输入张量。因而，部分(a)中所示的权重的4×4核心可以首先被应用于与输入张量的前四行和前四列对应的输入张量的元素(值未示出)。该过程的结果可以是结果卷积的张量中的第一元素，即图10的部分(a)所示的结果卷积的张量的元素“a”。

[0096] 由于输入张量的处理是使用具有步长1而不是指定的步长4的卷积神经网络层执行的，所以可将部分(a)中所示的4×4权重集应用于对应于激活值阵列的前四行的输入张量，以及输入张量的第二至第五列的元素(值未示出)。处理的结果是卷积的张量的第二元素，即图10的部分(a)中所示的卷积结果的元素“b”。可通过使用步长1向激活值阵列应用4×4权重集，即通过在列和行方向两者中增量地向激活值阵列应用4×4权重集重复该过程。处理导致图10的部分(a)中所示的88卷积的张量。

[0097] 如图9的部分(b)所示，然后在卷积的张量和掩蔽张量之间执行元素乘法，以获得修改的卷积的张量。基于输入张量的大小或卷积的张量的大小确定掩蔽张量的大小，由于使用具有步长1的卷积神经网络层在图10的部分(a)处的处理，这些大小通常为相等的。掩蔽张量在对应于如果使用具有指定步长的卷积神经网络层处理输入张量则本就会被生成的值的位置处包括单一值，即1。然后，通常掩蔽张量中的单一值项的位置取决于卷积神经网络层的指定步长。在图10的示例中，由于卷积神经网络层具有步长4，所以掩蔽张量将在

列和方向两者上的每个第四位置处都将包括单一值。掩蔽张量的其他项被分配有零值，使得卷积的张量和掩蔽张量的元素乘法将导致将如果输入张量被具有指定步长的卷积神经网络处理则本就不会被生成的所有值归零。

[0098] 执行卷积的张量和掩蔽张量的元素乘法以生成修改的卷积的张量。如图10所示，在元素乘法之后，保持卷积的张量的每第四个元素，并且由于与掩蔽矩阵的相应零值元素相乘，卷积的张量的其余元素变为零。因而，在8×8卷积的张量的元素中，只有四个元素保持非零。

[0099] 在一些实施中，能够通过首先将卷积的张量的元素乘以非单一因子，然后将那些元素乘以第二非单一因子获得类似的结果。例如，掩蔽张量可在对应于使用具有指定步长的卷积神经网络层处理输入张量则本就会被生成的值的位置处包括多个2(或另一个值)。因而，按上述示例，卷积的张量和掩蔽张量的元素乘法生成修改的卷积的张量，其中卷积的张量的每第四个元素被加倍，并且其余元素为零。之后可执行对修改的卷积的张量乘以二分之一的纯量乘法(或其他值的倒数)。可替选地，可执行修改的卷积的张量与第二掩蔽张量的元素乘法，其中第二掩蔽张量在对应于如果使用具有指定步长的卷积神经网络层处理输入张量则本就会被生成的值的位置处包括二分之一的值。

[0100] 之后对图10的部分(c)中的修改的卷积结果阵列执行最大池化。最大池化的结果等效于如果输入张量已经由步长4的卷积神经网络层处理则本就会获得的结果。使用图6的过程，对修改的卷积的张量执行最大池化，以识别修改的卷积的张量的每个4×4窗口的最大值。然后将最大池化的结果存储为具有步长4的卷积神经网络层的输出张量。因为输入张量是8×8阵列，所以具有步长4的神经网络层进行的处理导致2×2输出阵列。2×2输出阵列可被例如，以光栅顺序存储在图2的统一的缓冲器308中。2×2输出阵列的值可作为输入提供给神经网络的后续层。

[0101] 本说明书中所述的主题和功能操作的实施例能够以数字电子电路、有形计算机软件或固件、计算机硬件实施，包括本说明书中公开的结构和它们的结构等效物，或者以一个或多个这些装置的组合实施。本说明书中所述的主题的实施例能够被实施为一个或多个计算机程序，即，被编码在有形非暂态程序载体上的一个或多个计算机程序指令模块，从而被数据处理设备执行，或者控制数据处理设备的操作。可替选地或另外，程序指令能够被编码在人工生成的传播信号上，例如机器生成的电、光或电磁信号，其被生成以对发送至适合由数据处理设备执行的接收器设备的信息编码。计算机存储介质能够为机器可读存储装置、机器可读存储基板、随机或串行存取存储器装置，或一个或多个这些装置的组合。

[0102] 术语“数据处理设备”包括用于处理数据的各种设备、装置和机器，包括例如可编程处理器、计算机或多个处理器或计算机。该设备能够包括专用逻辑电路，例如FPGA(现场可编程门阵列)或ASIC(专用集成电路)。除了硬件之外，该设备也能够包括为所述计算机程序创建执行环境的代码，例如组成处理器固件、协议栈、数据库管理系统、操作系统或其中一个或更多个的组合的代码。

[0103] 计算机程序(也可被称为或描述为程序、软件、软件应用、模块、软件模块、脚本或代码)能够以任何形式的编程语言编写，包括编译或解释语言，或声明性或程序性语言，并且能够以任何形式部署，包括作为独立程序或模块、部件、子程序或适用于计算环境的其它单元。计算机程序可以但不需要对应于文件系统中的文件。能够将程序存储在保持其它程

序或数据的一部分文件中，例如存储在标记语言文档中的一个或多个脚本，存储在专用于所述程序的单个文件中，或者存储在多个协调文件中，例如存储一个或多个模块、子程序或代码部分的文件。能够将计算机程序部署为将在一个计算机上，或在位于一个站点上，或者分布在多个站点上并由通信网络互连的多个计算机上执行。

[0104] 本说明书中所述的过程和逻辑流程能够由执行一个或多个计算机程序的一个或多个可编程计算机执行，以通过对输入数据进行操作并生成输出执行功能。过程和逻辑流程也可以由专用逻辑电路执行，并且设备也可被实施为专用逻辑电路，该专用逻辑电路为例如FPGA(现场可编程门阵列)或ASIC(专用集成电路)。

[0105] 作为示例，适用于执行计算机程序的计算机能够基于通用或专用微处理器或两者，或任何其它类型的中央处理单元。通常，中央处理单元将从只读存储器或随机存取存储器或两者接收指令和数据。计算机的基本元件是用于完成或执行指令的中央处理单元以及用于存储指令和数据的一个或多个存储器装置。通常，计算机也将包括或被可操作地耦合，以从用于存储数据的一个或多个大容量存储装置，例如磁，磁光盘或光盘接收数据或向其传输数据或两者。然而，计算机不需要具有这些装置。此外，计算机能够被植入另一装置中，仅举几例，例如移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制台、全球定位系统(GPS)接收器或便携式存储设备，例如通用串行总线(USB)闪存驱动器。

[0106] 适用于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、介质和存储器装置，作为示例，包括半导体存储器装置，例如EPROM、EEPROM和闪速存储器装置；磁盘，例如内部硬盘或可移动盘；磁光盘；以及CD ROM和DVD-ROM盘。处理器和存储器可由专用逻辑电路补充或被并入其中。

[0107] 为了发送与用户的互动，本说明书中所述的主题的实施例能够在具有显示装置(例如CRT(阴极射线管)或LCD(液晶显示器)监视器)，以向用户显示信息，以及键盘和指向装置，例如用户能够通过其将输入发送至计算机的鼠标或者轨迹球的计算机上实施。也能够使用其它类型的装置发送与用户的交互；例如，提供给用户的反馈能够为任何形式的感觉反馈，例如视觉反馈、听觉反馈或触觉反馈；并且能够以任何形式接收来自用户的输入，包括声音，语音或触觉输入。此外，计算机能够通过向由用户使用的装置发送文档以及从该装置接收文档与用户互动；例如，通过响应于从网络浏览器接收的请求将网页发送至用户的客户端装置上的网络浏览器。

[0108] 本说明书中描述的主题的实施例能够在下列计算系统中实施，该计算系统包括后端部件，例如作为数据服务器，或包括中间件部件，例如应用服务器，或者包括前端部件，例如具有图形用户界面或网络浏览器的客户端计算机，或者一个或多个这种后端，中间件或前端部件的任何组合，用户能够通过这些部件与本说明书中所述的主题的实施进行互动。系统的部件可以通过数字数据通信的如何形式或介质，例如通信网络被相互连接。通信网络的示例包括局域网(“LAN”)和广域网(“WAN”)，例如因特网。

[0109] 计算系统能够包括客户端和服务器。客户端和服务器通常彼此远离，并且通常通过通信网络进行交互。客户端和服务器之间的关系是由于相应计算机上运行，并且彼此具有客户端-服务器关系的计算机程序而发生的。

[0110] 虽然本说明书包括许多特定的实施细节，但是这些细节不应被解释为对任何发明或所要求的范围的限制，而是应被解释为对特殊发明的特殊实施例特定的特征的说明。在

本说明书中在单独实施例的背景下描述的某些特征也能够在单个实施例中组合地实施。相反，在单个实施例的背景下描述的各种特征也可以单独地或以任何合适的子组合在多个实施例中加以实施。此外，虽然这些特征可能在上文被描述为以某些组合起作用，并且甚至同样最初要求如此，但是在一些情况下，能够从组合中去除来自所要求组合的一个或多个特征，并且所要求的组合可以针对子组合或子组合的变形。

[0111] 类似地，虽然在附图中示出特定顺序的操作，但是这不应被理解为要求以所示的特定顺序或按顺序执行这些操作，或者执行所有所示的操作以实现期望的结果。在某些情况下，多任务和并行处理可能是有利的。此外，上述实施例中的各种系统模块和部件的分离不应被理解为在所有实施例中都需要这种分离，并且应理解，所描述的程序部件和系统通常能够被集成在单个软件产品中，或者被打包成多个软件产品。

[0112] 在下列示例中总结了进一步的实施：

[0113] 示例1：一种方法，包括：接收在硬件电路上处理神经网络的请求，所述神经网络包括具有大于1的步长的第一卷积神经网络层；和作为响应产生指令，当被所述硬件电路执行时，所述指令引起所述硬件电路在所述神经网络处理输入张量期间通过执行下列操作产生等效于所述第一卷积神经网络层的输出的层输出张量，所述操作包括：使用具有等于1否则等效于所述第一卷积神经网络层的步长的第二卷积神经网络层处理对所述第一卷积神经网络层的所述输入张量以产生第一张量；将如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长则本就不会被生成的所述第一张量的元素归零以生成第二张量；以及在所述第二张量上执行最大池化以生成所述层输出张量。

[0114] 示例2：根据示例1所述的方法，其中对所述第一张量的元素归零包括：将所述第一张量的元素子集乘以0；和将所述子集中未包括的所述第一张量的元素乘以1。

[0115] 示例3：根据示例1所述的方法，其中将所述第一张量的元素归零包括：执行掩蔽张量和所述第一张量的元素乘法以生成所述第二张量，其中所述掩蔽张量包括(i)在对应于如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长则本就不会被生成的所述第一张量的元素的所述掩蔽张量的每个元素位置处的零，和(ii)在所述掩蔽张量的每个其它元素位置处的1。

[0116] 示例4：根据示例3所述的方法，其中所述掩蔽张量被存储在可由硬件电路访问的存储器处，并且其中由在所述硬件电路中所包括的硬件中实施的向量计算单元执行所述掩蔽张量和所述第一张量的元素乘法。

[0117] 示例5：根据示例1所述的方法，其中对所述第一张量的元素归零包括：执行所述第一掩蔽张量和所述第一张量的元素乘法以生成修改的第一张量，其中所述第一掩蔽张量包括(i)在对应于如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长则本就不会被生成的所述第一张量的元素的所述掩蔽张量的每个元素位置处的零，和(ii)在对应于如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长则本就会被生成的所述第一张量的元素的所述掩蔽张量的每个元素位置处的相应的非零值；以及执行第二掩蔽张量和所述修改的第一张量的元素乘法，其中所述第二掩蔽张量在对应于如果所述第二卷积神经网络层具有第一卷积神经网络层的步长则本就会被生成的所述第一张量的元素的每个元素位置处都包括所述第一掩蔽张量的相应的非零值的逆。

[0118] 示例6：根据示例1至5任一项所述的方法，其中执行最大池化包括对于由所述第一

卷积神经网络层的步长限定的所述第二张量的一个或多个窗口中的每个窗口都获得处于所述窗口内的元素的最大值元素。

[0119] 示例7:根据示例6所述的方法,其中所述第二张量的一个或多个窗口中的每个窗口都为具有对应于所述卷积神经网络层的步长的尺寸的矩形窗口,并且包括所述第二张量的不同元素。

[0120] 示例8:根据示例1至7任一项所述的方法,其中执行最大池化包括对于所述第二张量的元素的一个或多个子集中的每个子集都获得所述子集的最大值元素。

[0121] 示例9:根据示例1至8任一项所述的方法,其中由硬件电路的池化电路执行在所述第二张量上执行的最大池化。

[0122] 示例10:根据示例1至9任一项所述的方法,其中所述卷积神经网络层是所述神经网络中的第一神经网络层,并且其中所述输入张量是包括对应于数字图像的像素的元素的数字图像的表示。

[0123] 示例11:根据示例1至10任一项所述的方法,其中所述输入张量被存储在硬件电路的统一的缓冲器处,并且所述第二卷积神经网络层的权重被存储在硬件电路的动态存储器处,并且其中使用所述第二卷积神经网络层处理对所述第一卷积神经网络层的所述输入张量包括:将所述输入张量从所述统一的缓冲器发送到在硬件中实施的硬件电路的矩阵计算单元;将所述第二卷积神经网络层的权重从所述动态存储器发送到所述硬件电路的矩阵计算单元;以及由所述硬件电路的矩阵计算单元,使用所述第二卷积神经网络层的权重处理所述输入张量以生成所述第一张量。

[0124] 示例12:一种系统,包括:硬件电路;和存储可操作指令的一个或多个存储装置,当被所述硬件电路执行时,所述指令引起所述硬件电路执行下列操作,包括:使用具有等于1否则等效于所述卷积神经网络层的步长的第二卷积神经网络层处理对具有大于1的步长的所述卷积神经网络层的所述输入张量以产生第一张量;将如果所述第二卷积神经网络层具有所述卷积神经网络层的步长则本就不会被生成的所述第一张量的元素归零以生成第二张量;以及在所述第二张量上执行最大池化以生成层输出张量。

[0125] 示例13:根据示例12所述的系统,其中将所述第一张量的元素归零包括:执行掩蔽张量和所述第一张量的元素乘法以生成所述第二张量,其中所述掩蔽张量包括(i)在对应于如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长则本就不会被生成的所述第一张量的元素的所述掩蔽张量的每个元素位置处的零,和(ii)在所述掩蔽张量的每个其它元素位置处的1。

[0126] 示例14:根据示例13所述的系统,其中所述掩蔽张量被存储在可由硬件电路访问的存储器处,并且其中由在所述硬件电路中所包括的硬件中实施的向量计算单元执行所述掩蔽张量和所述第一张量的元素乘法。

[0127] 示例15:根据示例12至14任一项所述的系统,其中执行最大池化包括对于由所述第一卷积神经网络层的步长限定的所述第二张量的一个或多个窗口中的每个窗口都获得处于所述窗口内的元素的最大值元素。

[0128] 示例16:根据示例15所述的系统,其中所述第二张量的一个或多个窗口中的每个窗口都为具有对应于所述卷积神经网络层的步长的尺寸的矩形窗口,并且包括所述第二张量的不同元素。

[0129] 示例17:根据示例12至16任一项所述的系统,其中由硬件电路的池化电路执行在所述第二张量上执行的最大池化。

[0130] 示例18:根据示例12至17任一项所述的系统,其中所述卷积神经网络层是所述神经网络中的第一神经网络层,并且其中所述输入张量是包括对应于数字图像的像素的元素的数字图像的表示。

[0131] 示例19:根据示例12至18任一项所述的系统,其中所述输入张量被存储在硬件电路的统一的缓冲器处,并且所述第二卷积神经网络层的权重被存储在硬件电路的动态存储器处,并且其中使用所述第二卷积神经网络层处理对所述第一卷积神经网络层的所述输入张量包括:将所述输入张量从所述统一的缓冲器发送到在硬件中实施的硬件电路的矩阵计算单元;将所述第二卷积神经网络层的权重从所述动态存储器发送到所述硬件电路的矩阵计算单元;以及由所述硬件电路的矩阵计算单元,使用所述第二卷积神经网络层的权重处理所述输入张量以生成所述第一张量。

[0132] 示例20:一种利用计算机程序编码的计算机可读存储装置,所述程序包括指令,如果被一个或多个计算机执行,所述指令引起所述一个或多个计算机执行下列操作:接收在硬件电路上处理神经网络的请求,所述神经网络包括具有大于1的步长的第一卷积神经网络层;和作为响应产生指令,当被所述硬件电路执行时,所述指令引起所述硬件电路在所述神经网络处理输入张量期间通过执行下列操作产生等效于所述第一卷积神经网络层的输出的层输出张量,所述操作包括:使用具有等于1否则等效于所述第一卷积神经网络层的步长的第二卷积神经网络层处理对所述第一卷积神经网络层的所述输入张量以产生第一张量;将如果所述第二卷积神经网络层具有所述第一卷积神经网络层的步长则本就不会被生成的所述第一张量的元素归零以生成第二张量;以及在所述第二张量上执行最大池化以生成所述层输出张量。

[0133] 已经描述了主题的特殊实施例。其它实施例在所附权利要求的范围内。例如,权利要求中所述的动作可以被以不同的顺序执行,并且仍然实现期望的结果。作为一个示例,附图中所示的过程不一定需要所示的特定顺序或相继顺序来实现期望的结果。在某些实施中,多任务和并行处理可能是有利的。

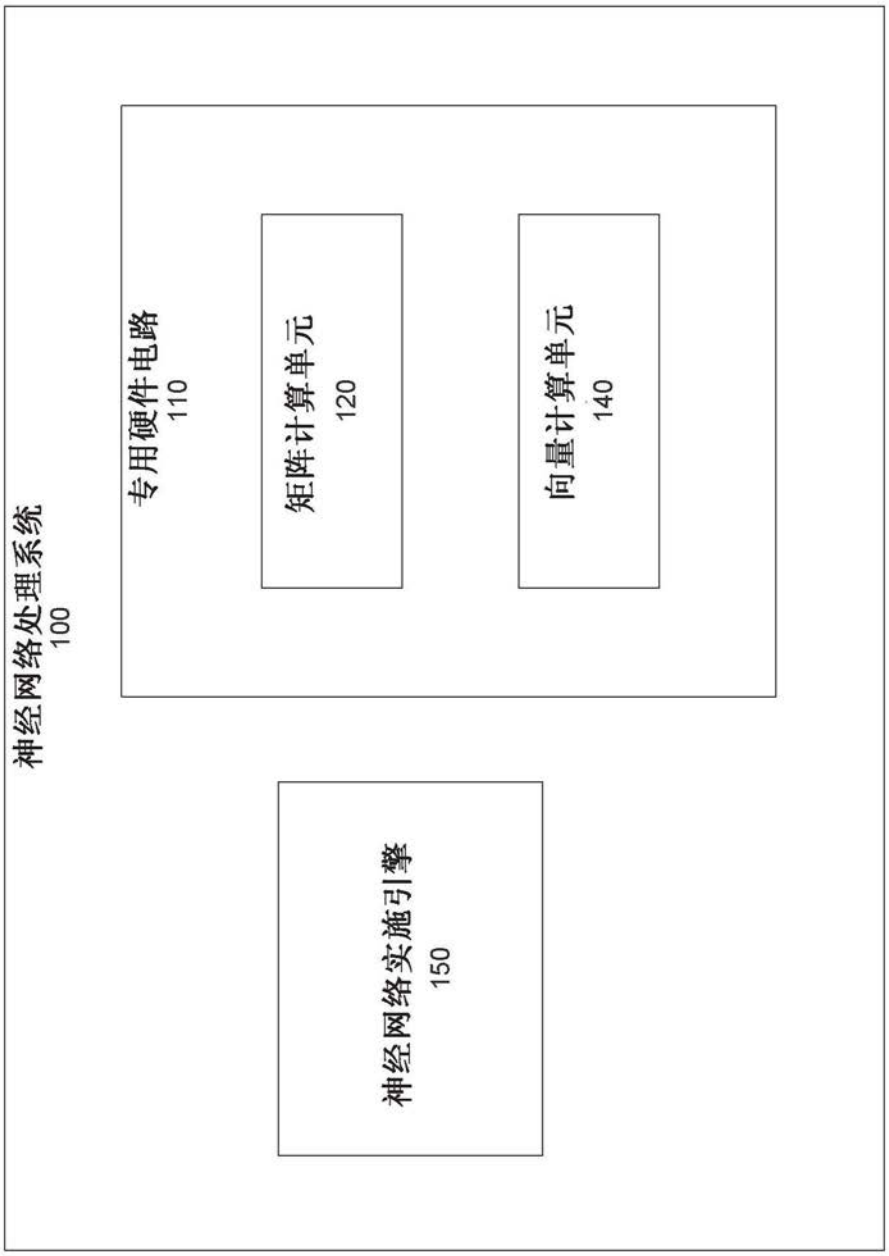


图1

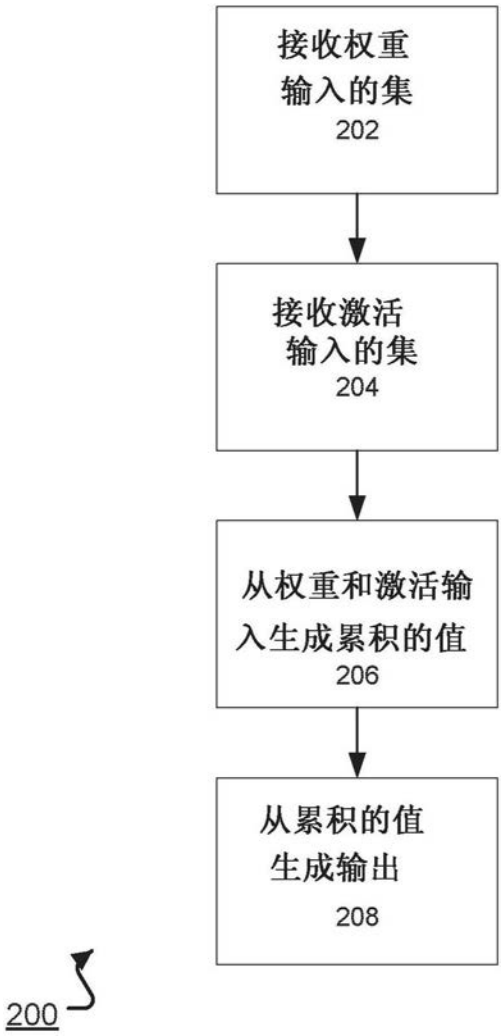

接收权重输入的集
202
接收激活输入的集
204
从权重和激活输入生成累积的值
206
从累积的值生成输出
208
200

图2

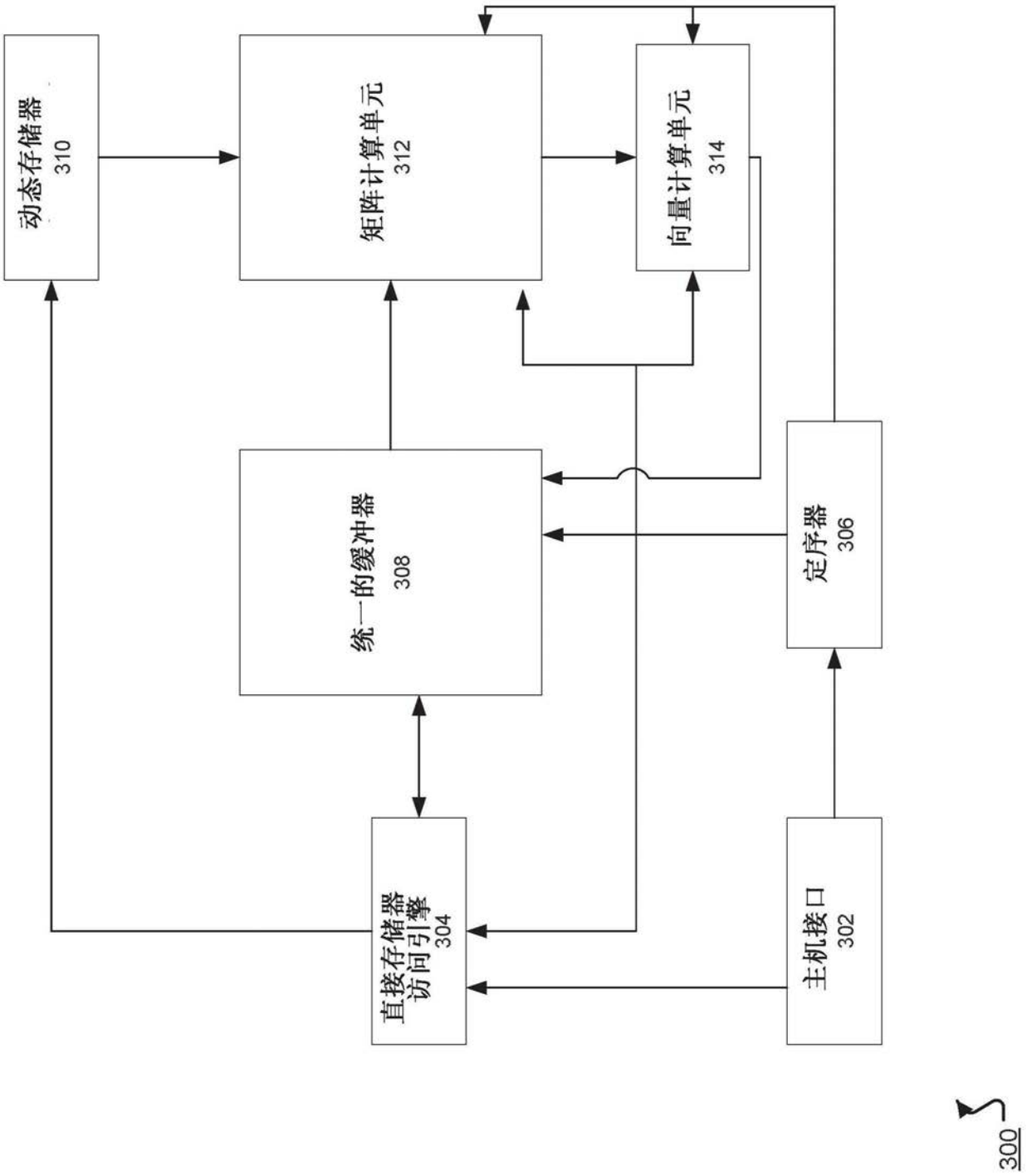

动态存储器
310
矩阵计算单元
312
向量计算单元
314
统一的缓冲器
308
定序器
306
直接存储器
访问引擎
304
主机接口
302
300

图3

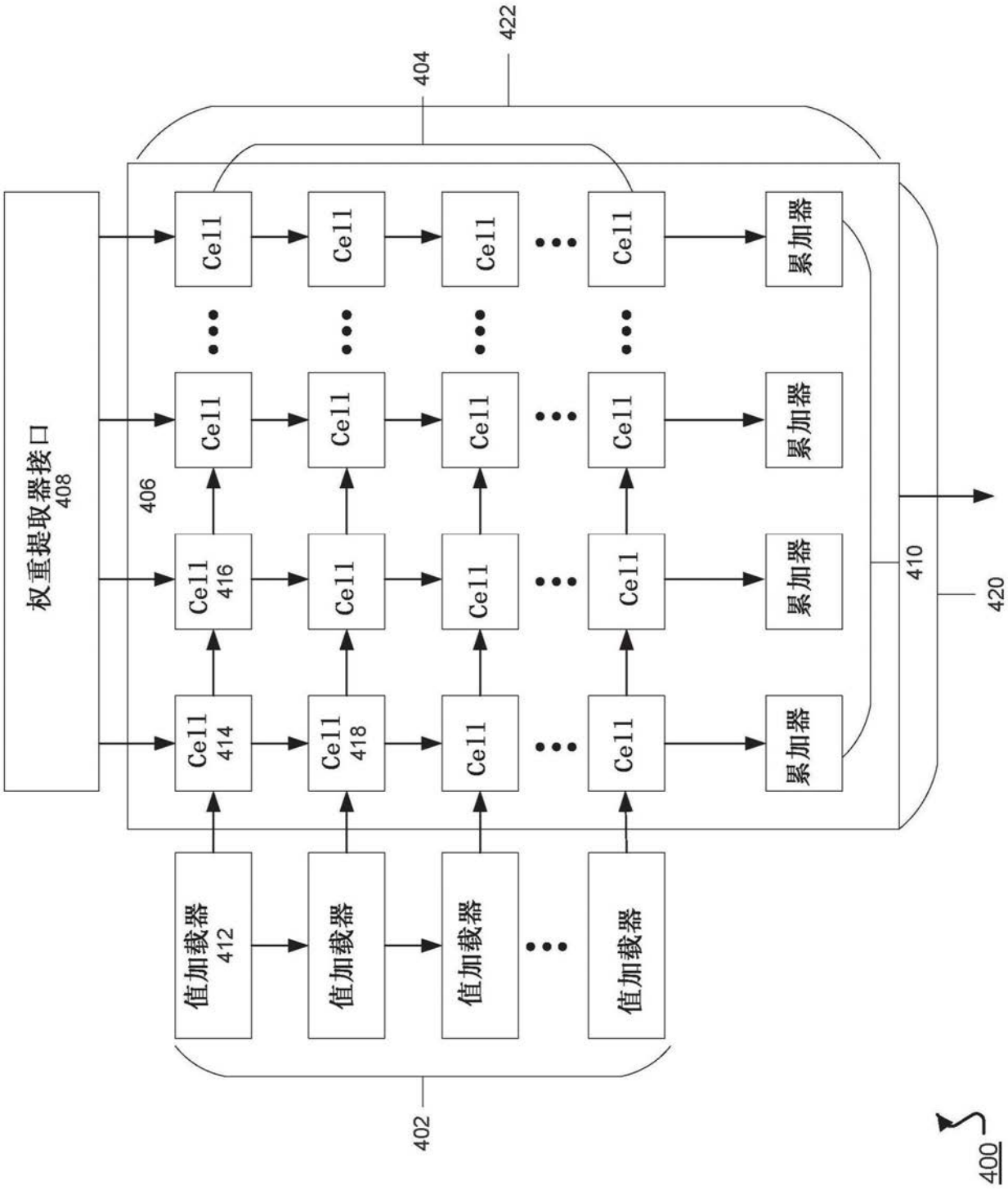

422
404
权重提取器接口
408
406
Cell
累加器
Cell 416
Cell 414
Cell 418
值加载器
412
410
420
402
400

图4

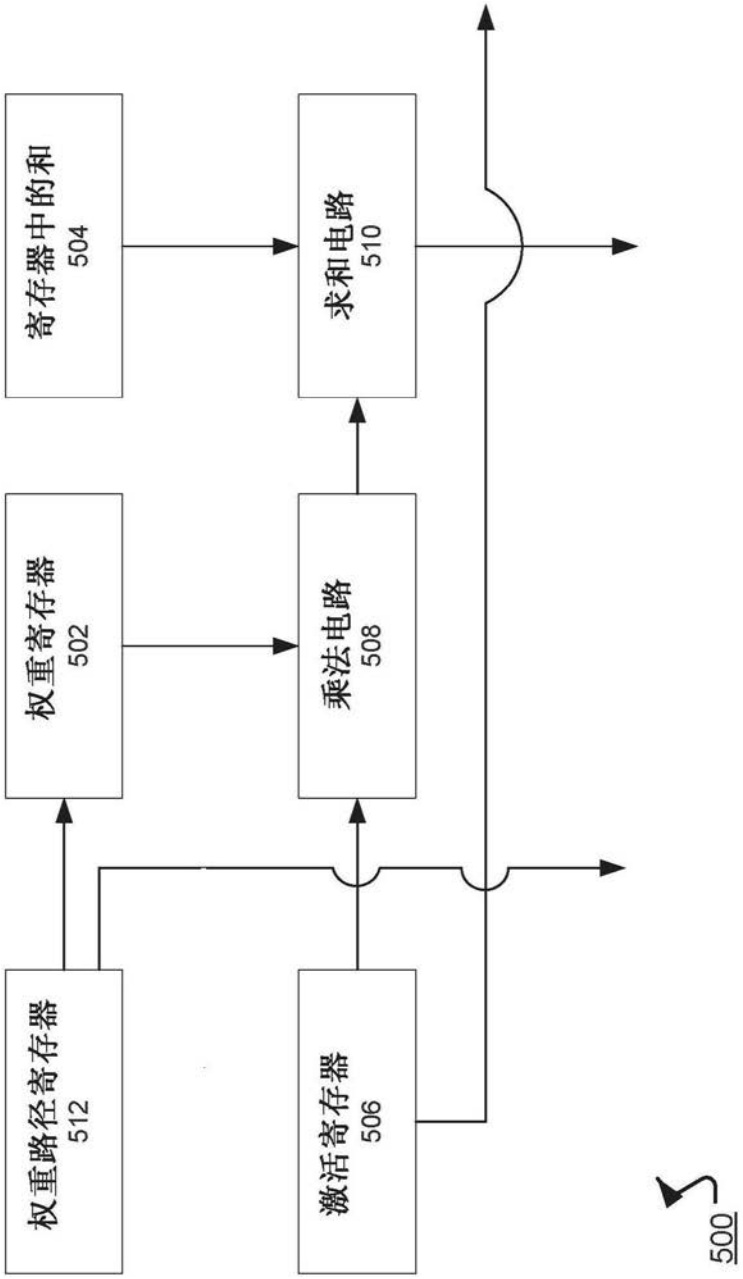

寄存器中的和
504
求和电路
510
权重寄存器
502
乘法电路
508
权重路径寄存器
512
激活寄存器
506
500

图5

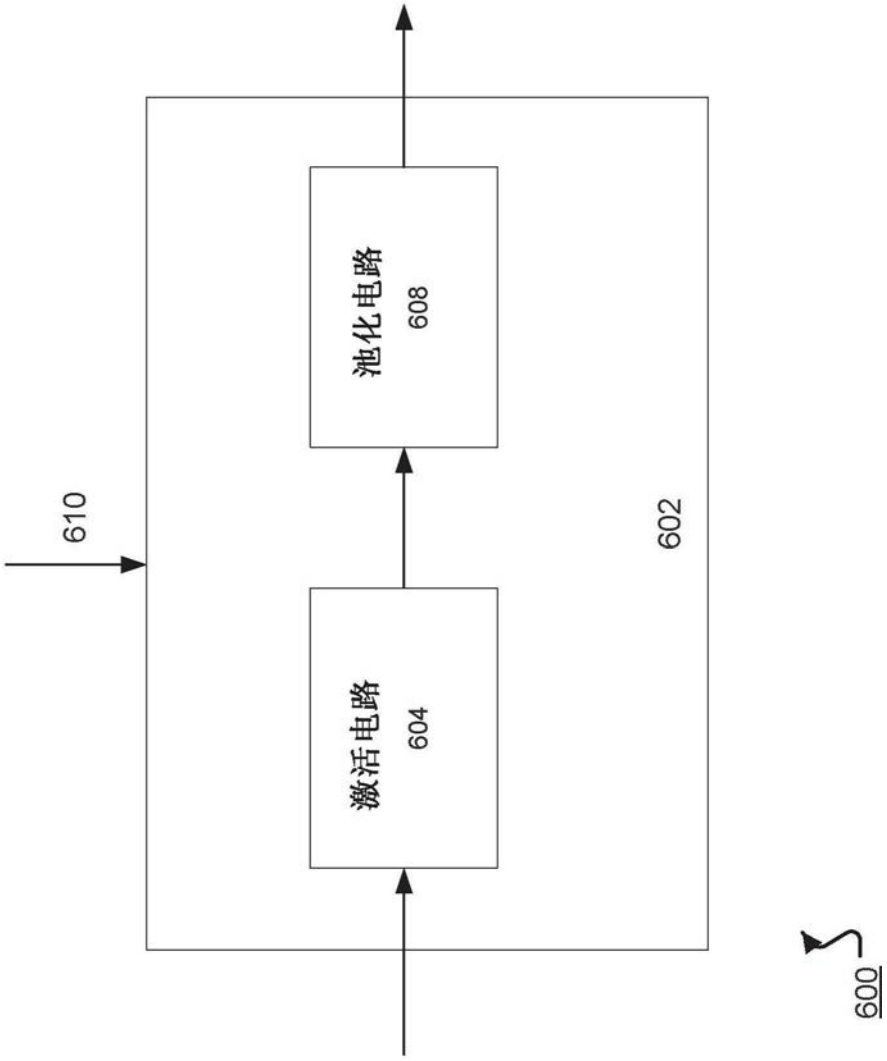


图6

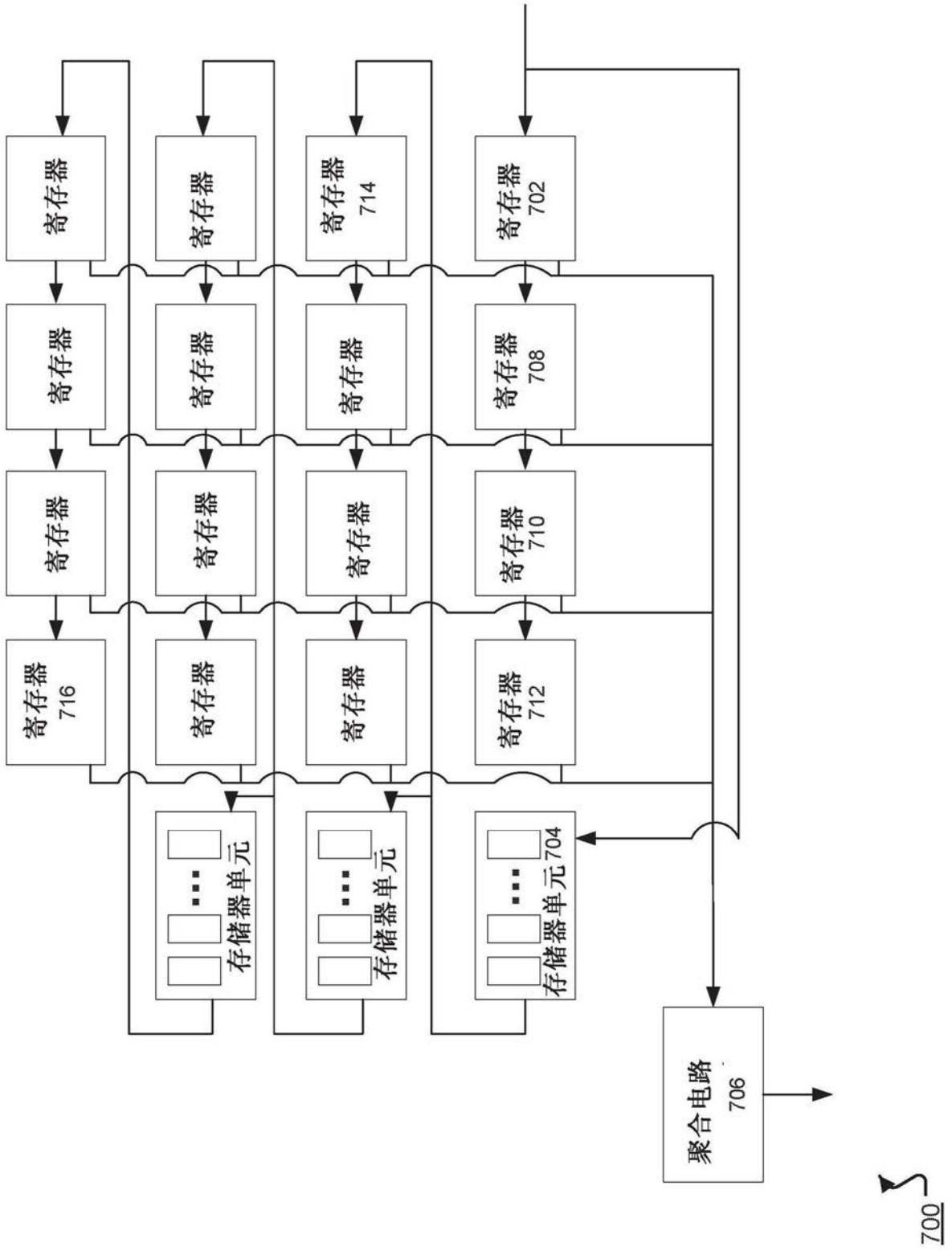

寄存器
寄存器
寄存器 714
寄存器 702
寄存器
寄存器
寄存器
寄存器 708
寄存器
寄存器
寄存器
寄存器 710
寄存器 716
寄存器
寄存器
寄存器 712
存储器单元
存储器单元
存储器单元 704
聚合电路 706
700

图7

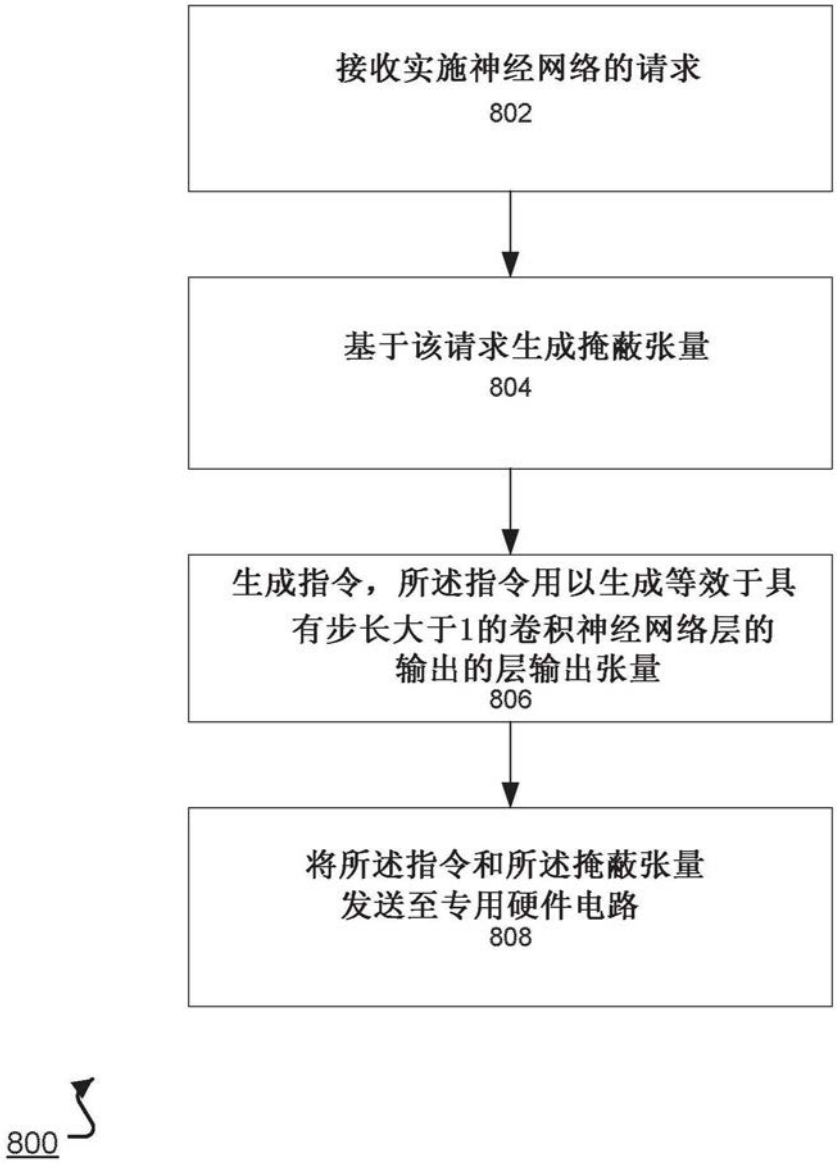

接收实施神经网络的请求
802
基于该请求生成掩蔽张量
804
生成指令，所述指令用以生成等效于具有步长大于1的卷积神经网络层的输出的层输出张量
806
将所述指令和所述掩蔽张量发送至专用硬件电路
808
800

图8

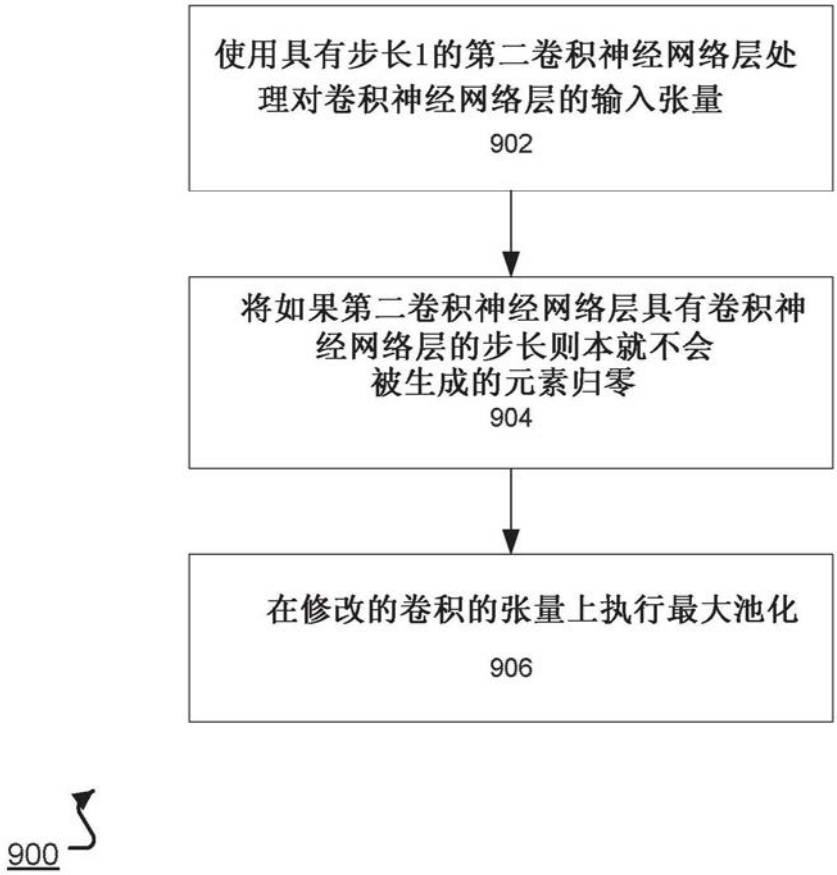
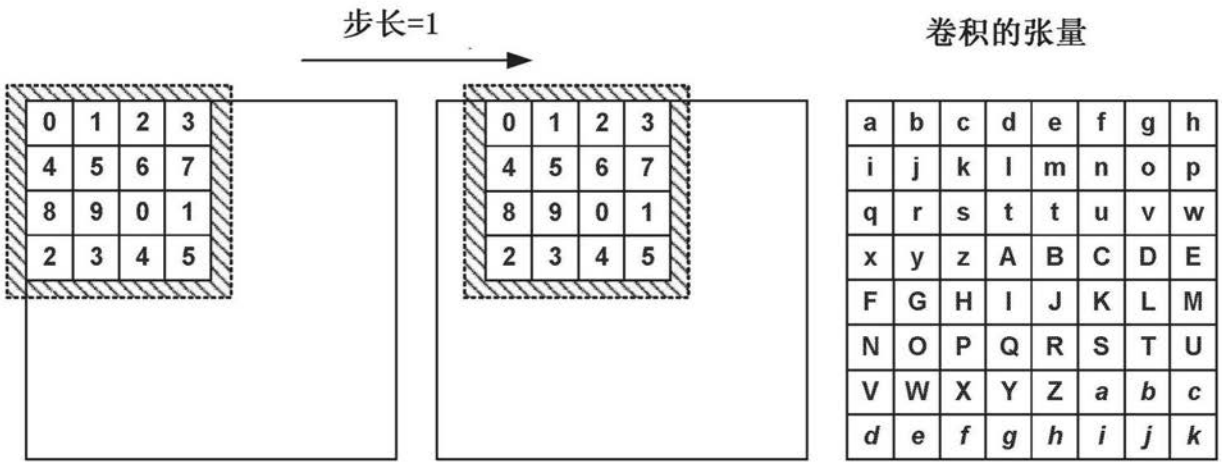
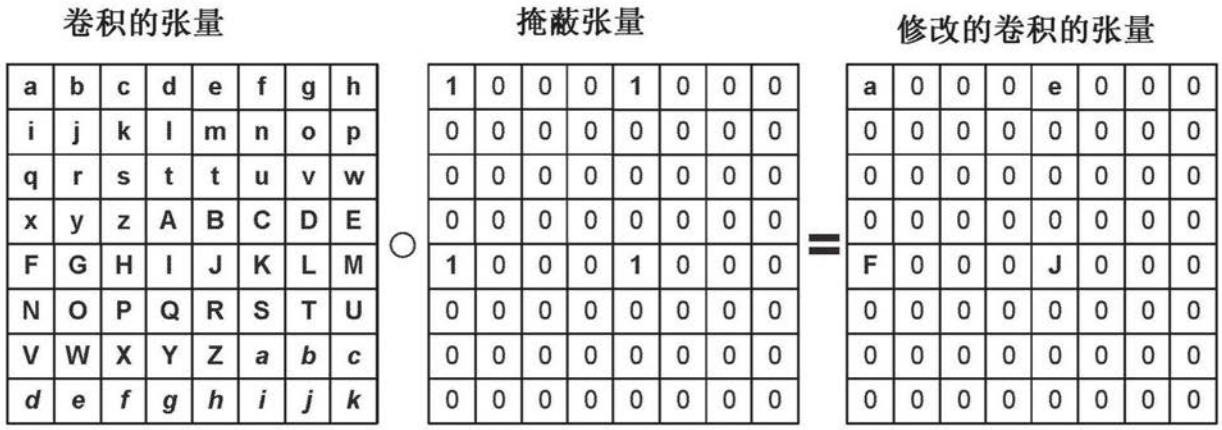

使用具有步长1的第二卷积神经网络层处理对卷积神经网络层的输入张量
902
将如果第二卷积神经网络层具有卷积神经网络层的步长则本就不会被生成的元素归零
904
在修改的卷积的张量上执行最大池化
906
900

图9

(a)


步长=1
卷积的张量
0 1 2 3
4 5 6 7
8 9 0 1
2 3 4 5
a b c d e f g h
i j k l m n o p
q r s t t u v w
x y z A B C D E
F G H I J K L M
N O P Q R S T U
V W X Y Z a b c
d e f g h i j k

(b)


卷积的张量
掩蔽张量
修改的卷积的张量
a b c d e f g h
i j k l m n o p
q r s t t u v w
x y z A B C D E
F G H I J K L M
N O P Q R S T U
V W X Y Z a b c
d e f g h i j k
1 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
1 0 0 0 1 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
a 0 0 0 e 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
F 0 0 0 J 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0

(c)


修改的卷积的张量
a 0 0 0 e 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
F 0 0 0 J 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
最大池化
a e
F J

图10