(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 107633295 B
(45)授权公告日 2020.04.28

(21)申请号 201710874198.X

(22)申请日 2017.09.25

(65)同一申请的已公布的文献号
申请公布号 CN 107633295 A

(43)申请公布日 2018.01.26

(73)专利权人 南京地平线机器人技术有限公司
地址 210046 江苏省南京市经济技术开发区兴智路兴智科技园A栋20层

(72)发明人 凌坤 陈亮 李建军 李德林 黄畅

(74)专利代理机构 北京市正见永申律师事务所 11497
代理人 黄小临 王怀章

(51)Int.Cl.
G06N 3/04(2006.01)

(56)对比文件
US 2005157939 A1,2005.07.21,全文.
US 4918472 A,1990.04.17,全文.

审查员 谭碧云

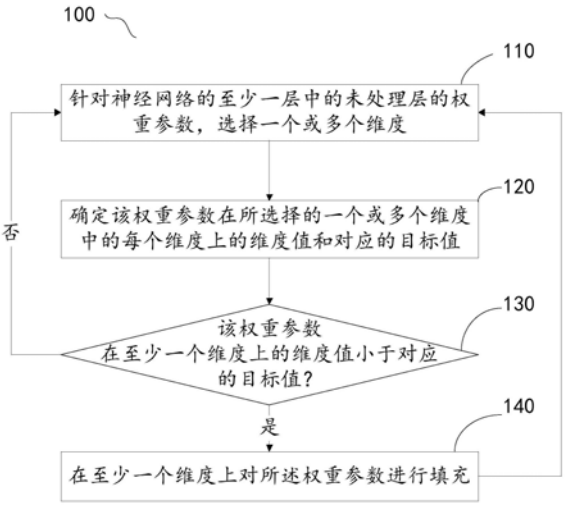
权利要求书3页 说明书11页 附图6页

(54)发明名称

用于适配神经网络的参数的方法和装置

(57)摘要

公开了一种用于适配神经网络的参数的方法和装置。该方法包括:针对神经网络的至少一层中的每一层的权重参数,选择一个或多个维度;确定权重参数在每个维度上的维度值和对应的目标值;以及如果权重参数在至少一个维度上的维度值小于对应的目标值,则对权重参数进行填充,使得填充后的权重参数在每个维度上的维度值等于对应的目标值。通过该方法能够获得具有规范形式的神经网络参数,有利于简化神经网络及相关硬件的设计和实现。



CN 107633295 B

1.一种用于适配神经网络的参数的方法,包括:

针对神经网络的至少一层中的每一层的权重参数,选择一个或多个维度;

确定所述权重参数在所述一个或多个维度中的每个维度上的维度值;

基于支持神经网络的计算的硬件参数,确定所述每个维度上对应的目标值;以及

如果所述权重参数在所述一个或多个维度中的至少一个维度上的维度值小于对应的目标值,则在所述至少一个维度上对所述权重参数进行填充,使得在填充之后所获得的权重参数在所述一个或多个维度中的每个维度上的维度值等于对应的目标值,

其中,维度目标值的确定步骤包括:

基于支持神经网络的计算的硬件的使用率,确定所述权重参数在所述每个维度上的目标值的候选值集合;以及

基于支持神经网络的计算的硬件可处理的权重参数的形式和/或数量,确定所述每个维度上对应的目标值,其中所述每个维度上对应的目标值是所述权重参数在所述每个维度上的目标值的候选值集合中大于或等于所述每个维度上对应维度值的最小值。

2.根据权利要求1所述的方法,其中,所述一个或多个维度包括宽度、高度、深度和数量中的至少一个。

3.根据权利要求1所述的方法,其中,所述一个或多个维度包括第一维度,并且所述权重参数在所述第一维度上的对应的目标值是基于与所述第一维度相关联的候选值的集合的子集所确定的,所述子集中的每个候选值大于或等于所述权重参数在所述第一维度上的维度值。

4.根据权利要求3所述的方法,其中,所述权重参数在所述第一维度上的对应的目标值是所述子集中的最小的候选值。

5.根据权利要求3所述的方法,其中,所述第一维度包括深度或数量,并且所述集合包括第一参数和第二参数的公倍数中的一个或多个公倍数,所述第一参数和所述第二参数是与支持神经网络的乘加运算的每组乘法器和加法器的布置以及组数相关的参数。

6.根据权利要求1所述的方法,其中,所述一个或多个维度包括第一维度和至少一个第二维度,并且所述权重参数在所述第一维度和所述至少一个第二维度上的对应的目标值是基于候选值数组的集合的子集所确定的,所述集合中的每个候选值数组包括与所述第一维度相对应的第一候选值以及分别与所述至少一个第二维度相对应的至少一个第二候选值,并且所述子集中的每个候选值数组中的每个候选值分别大于或等于所述权重参数在对应的维度上的维度值。

7.根据权利要求6所述的方法,其中,所述权重参数在所述第一维度上的对应的目标值是基于所述子集所确定的最小的第一候选值。

8.根据权利要求6所述的方法,其中,所述权重参数在所述第一维度以及所述至少一个第二维度上的对应的目标值分别是所述子集中的第一候选值数组中的对应的候选值,所述第一候选值数组中的所有候选值的乘积小于所述子集中的任何其他候选值数组中的所有候选值的乘积。

9.根据权利要求6所述的方法,其中,所述第一维度包括宽度或高度或深度,并且对于基于支持神经网络的计算的硬件参数所预先确定的、所述权重参数在所述第一维度与所述至少一个第二维度上的对应的目标值的每种组合,所述集合包括一个或多个对应的候选值

数组。

10.根据权利要求1所述的方法，还包括：

针对所述至少一个维度中的每个维度，确定与该维度相对应的填充模式，所述填充模式指示在该维度上对所述权重参数进行填充的一个或多个填充位置、填充量以及设置填充值的规则中的至少一个。

11.根据权利要求10所述的方法，其中，所述一个或多个维度包括宽度、高度、深度和数量中的至少一个，

与宽度相对应的填充模式所指示的信息包括在宽度方向上的左侧和/或右侧和/或中间的一个或多个位置处填充一列或多列，

与高度相对应的填充模式所指示的信息包括在高度方向上的上方和/或下方和/或中间的一个或多个位置处填充一行或多行，

与深度相对应的填充模式所指示的信息包括在深度方向上的前方和/或后方和/或中间的一个或多个位置处填充一排或多排，并且

与数量相对应的填充模式所指示的信息包括在所述权重参数的序列之前和/或之后和/或之中的一个或多个位置处填充一个或多个填充值。

12.根据权利要求10所述的方法，其中，设置填充值的规则包括以下中的至少一个：

使用零值进行填充；

使用预先定义的一个或多个非零值进行填充；以及

使用在所述权重参数中与要填充的位置相邻的位置处的值进行填充。

13.根据权利要求1所述的方法，还包括：

在所述至少一个维度包括宽度和/或高度和/或深度的情况下，对所述权重参数所在层的特征数据在对应的宽度和/或高度和/或深度上进行填充。

14.根据权利要求13所述的方法，其中，在所述特征数据的要填充的每个维度上的填充位置和填充量分别与在所述权重参数的对应的维度上的填充位置和填充量相同。

15.根据权利要求13所述的方法，其中，使用零值、预先定义的一个或多个非零值、在所述特征数据中与要填充的位置相邻的位置处的值、以及随机值中的至少一个对所述特征数据进行填充。

16.根据权利要求1至15中的任一项所述的方法，其中，所述神经网络是卷积神经网络。

17.一种用于适配神经网络的参数的装置，包括：

一个或多个处理器，被配置为执行如权利要求1至16中的任一项所述的方法。

18.一种用于适配神经网络的参数的装置，包括：

检测器，被配置为针对神经网络的至少一层中的每一层的权重参数所选择的一个或多个维度，确定所述权重参数在所述一个或多个维度中的每个维度上的维度值，并且基于支持神经网络的计算的硬件参数，确定所述每个维度上对应的目标值；以及

填充器，被配置为在所述权重参数在所述一个或多个维度中的至少一个维度上的维度值小于对应的目标值的情况下，在所述至少一个维度上对所述权重参数进行填充，使得在填充之后所获得的权重参数在所述一个或多个维度中的每个维度上的维度值等于对应的目标值，

其中，维度目标值的确定包括：

基于支持神经网络的计算的硬件的使用率，确定所述权重参数在所述每个维度上的目标值的候选值集合；以及基于支持神经网络的计算的硬件可处理的权重参数的形式和/或数量，确定所述每个维度上对应的目标值，其中所述每个维度上对应的目标值是所述权重参数在所述每个维度上的目标值的候选值集合中大于或等于所述每个维度上对应维度值的最小值。

19.一种非临时性的计算机可读写存储介质，在其上存储有程序指令，所述程序指令由处理器执行以实现如权利要求1至16中的任一项所述的方法。

用于适配神经网络的参数的方法和装置

技术领域

[0001] 本申请总体上涉及人工神经网络的技术领域，并且具体地涉及用于适配人工神经网络的参数的方法和装置。

背景技术

[0002] 人工神经网络(在本文中简称为神经网络)是一种模仿生物神经网络行为特征，进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间相互连接的关系，达到处理信息的目的。

[0003] 在通常的神经网络中，一个或多个神经元被联结在一起，并形成一种多层结构。每层中的每个神经元针对该层的每个特征数据(也可以被称为输入数据或特征图)，使用对应的一个或多个权重参数，基于特定的激活函数执行运算，以获得激活值，作为神经网络的输出结果或者下一层的特征数据。

[0004] 针对不同的应用场景，往往需要设计不同的神经网络架构，并且需要在某一类型的计算架构上使用一系列的运算来实现。因此，期望能够高效地表示神经网络，并且能够通过较低的硬件成本高效地实现神经网络中的运算，或者能够通过模拟的方式将神经网络中的运算高效地映射到计算架构所支持的运算中。

发明内容

[0005] 一方面，本申请提供一种用于适配神经网络的参数的方法，该方法包括：针对神经网络的至少一层中的每一层的权重参数，选择一个或多个维度；确定该权重参数在每个维度上的维度值和对应的目标值；以及如果权重参数在至少一个维度上的维度值小于对应的目标值，则在至少一个维度上对权重参数进行填充，使得在填充之后所获得的权重参数在每个维度上的维度值等于对应的目标值。

[0006] 另一方面，本申请还提供一种用于适配神经网络的参数的装置，其包括被配置为执行用于上述方法的一个或多个处理器。

[0007] 另一方面，本申请还提供一种用于适配神经网络的参数的装置，其包括：检测器，被配置为针对神经网络的至少一层中的每一层的权重参数所选择的一个或多个维度，确定权重参数在每个维度上的维度值和对应的目标值；以及填充器，被配置为在权重参数在至少一个维度上的维度值小于对应的目标值的情况下，在至少一个维度上对权重参数进行填充，使得在填充之后所获得的权重参数在每个维度上的维度值等于对应的目标值。

[0008] 另外，本申请还提供一种非临时性存储介质，在其上存储有用于执行上述方法的程序指令。

[0009] 通过根据本申请的示例的方法和装置，能够获得具有规范形式的神经网络参数，有利于简化神经网络及相关硬件的设计和实现。

附图说明

[0010] 在下文中将结合附图来描述根据本申请的示例,附图中:

[0011] 图1示出根据本申请的实施例的用于对神经网络的参数进行适配的示例方法;

[0012] 图2示出根据本申请的实施例的权利参数和特征数据的维度的示例;

[0013] 图3示出根据本申请的实施例的在数量的维度上对权重参数进行填充的示例;

[0014] 图4示出根据本申请的实施例的在宽度和/或高度的维度上对权重参数进行填充的示例;

[0015] 图5示出根据本申请的实施例的在宽度和/或高度的维度上对权重参数进行填充的示例;

[0016] 图6示出根据本申请的实施例的在宽度和/或高度的维度上对权重参数进行填充的示例;

[0017] 图7示出根据本申请的实施例的在深度的维度上对权重参数进行填充的示例;

[0018] 图8示出根据本申请的实施例的用于对神经网络的参数进行适配的示例装置的框图;以及

[0019] 图9示出根据本申请的实施例的用于对神经网络的参数进行适配的示例装置的框图。

具体实施方式

[0020] 基于不同的应用背景和处理目标,神经网络的不同层的特征数据和/或权重参数可能具有不同的形式和/或数量。在本文中,取决于神经网络的类型以及神经网络的层中的关键运算,权重参数包括与该层的关键运算相关联的关键参数。例如,对于卷积神经网络的卷积层而言,关键运算可以是卷积运算,并且权重参数可以包括一个或多个卷积核。

[0021] 根据所关注的运算以及权重参数的类型和/或观察角度或所关注的特征,可以确定或选择权重参数的一个或多个维度。例如,如果相应的权重参数是一个或多个数值,则权重参数至少可以具有数量这样的维度;如果相应的权重参数是一个或多个一维数据(例如,包括一个或多个数值的一维数组),则权重参数至少可以具有宽度(例如,一维数组中的数据的数量)和数量(例如,一维数组的数量)这样的维度;如果相应的权重参数是一个或多个二维数据(例如矩阵),则权重参数至少可以具有宽度(例如矩阵的列数)、高度(例如矩阵的行数)和数量这样的维度;如果相应的权重参数是一个或多个三维数据(例如立方体),则权重参数至少可以具有宽度、高度、深度和数量这样的维度。类似地,每层的特征数据也可以具有一个或多个维度(例如,宽度、高度、深度和数量)。应当理解,根据需要,对于神经网络的某个层的特征数据和/或权重参数,可以考虑其他类型和数量的维度,而不局限于上述示例。

[0022] 例如出于设计复杂度和成本等方面的考虑,用于支持神经网络中的运算的硬件(例如专用加速器、累加器、与累加器之间的连线布置等)可能只能够直接处理具有某些形式和/或数量的权重参数和/或特征数据,或者可能只有在某些形式和/或数量的权重参数和/或特征数据的情况下才能够具有相对良好的处理性能。因此,对特征数据和/或权重参数进行填充以获得具有某些规范的形式和/或数量的特征数据和/或权重参数,至少对于简化硬件设计、节约成本、提高处理效率等方面而言是有利的。

[0023] 图1示出根据本申请的实施例的用于适配神经网络的参数的示例方法100。

[0024] 如图1所示，针对神经网络的至少一层中的未处理的一层的权重参数，在步骤110中，选择一个或多个维度。

[0025] 在步骤120中，确定该权重参数在所选择的一个或多个维度中的每个维度上的维度值和对应的目标值。

[0026] 在步骤130中，针对每个维度，比较该权重参数在该维度上的维度值和对应的目标值。如果该权重参数在至少一个维度上的维度值小于对应目标值，则方法100继续到步骤140。否则，方法100返回到步骤110，并继续处理下一个未处理层的权重参数。

[0027] 在步骤140中，在基于步骤130中的比较结果所确定的每个维度上，对该权重参数进行填充，使得在填充之后所获得的权重参数在所选择的一个或多个维度中的每个维度上的维度值均等于对应的目标值。然后，方法100返回到步骤110，以继续处理下一个未处理层的权重参数。

[0028] 在下文中以卷积神经网络为例，更详细地描述根据本申请的实施例的方法。应当理解，本申请的原理并不局限于卷积神经网络，而是可以适用于其他类型的神经网络。

[0029] 图2示出卷积神经网络中的特征数据和权重参数的维度的示例。如图2所示，特征数据可以至少具有宽度、高度、深度(或通道数)和数量(未示出)这样的维度。如图2所示，权重参数可以是具有相同形状和大小的一个或多个卷积核，并且可以至少具有宽度(每个卷积核的宽度)、高度(每个卷积核的高度)、深度(每个卷积核的深度或通道数)和数量(卷积核的个数)这样的维度。

[0030] 对于图2所示的权重参数的示例，例如，示例方法100可以在步骤110中选择宽度、高度、深度和数量中的一个或多个，作为要进一步处理的维度。

[0031] 根据一个实施例，例如在示例方法100的步骤110中所选择的一个或多个维度中的部分或全部维度的对应的目标值可以独立地确定。

[0032] 例如，在支持神经网络的乘加运算的硬件电路中，可以包括一组或多组乘法器和加法器的布置，每组乘法器和加法器的布置中可以包括一个或多个乘法器以及一个或多个加法器。基于每组乘法器和加法器的布置，确定在权重参数的深度值为A时，乘法器和加法器的使用率最高(或相对较高)，并且例如可以省去选择/使能电器的设计/布置，则可以将权重参数在深度上的目标值确定为A的某个倍数，其中该倍数大于或等于权重参数在深度上的维度值。例如，假设A＝3，则对于当前的神经网络而言，权重参数在深度上的目标值的候选值的集合为{x|x＝n*A，n为正整数}。假设当前层的权重参数在深度上的维度值为5，则可以将该权重参数在深度上的目标值确定为前述候选值的集合中的大于或等于5的某个值，例如6、9或12等。例如，可以从候选值的集合的子集{x|x＝m*A，n是大于或等于5的正整数}中选择最小值(即，6)作为目标值。

[0033] 在另外的示例中，基于神经网络的架构设计和/或支持神经网络的硬件的设计，确定权重参数在深度上的目标值应当为集合{3，5，7，9，12，256，1024，……}中的某个值，例如在权重参数的深度为该集合中的某个值的情况下，可以减少额外处理或者可以硬件的处理性能比较好，则可以将该权重参数在深度上的目标值确定为该集合中的大于或等于5的某个值，例如5、7、9、12、256、1024等。例如，可以从这些候选值中选择最小的值5作为目标值，在这样的情况下，维度值等于目标值，因此不需要在深度上对该权重参数进行填充。在另外

的示例中，可以选择例如7作为目标值，在这样的情况下，维度值小于目标值，因此需要在深度上对该权重参数进行填充。

[0034] 在另外的示例中，基于每组乘法器和加法器的布置或者神经网络的架构设计等，确定在权重参数的数量为B时，乘法器和加法器的使用率最高(或相对较高)，并且例如可以省去选择/使能电器的设计/布置，则可以将权重参数在数量上的目标值确定为B的某个倍数，其中该倍数大于或等于权重参数在数量上的维度值。例如，假设B＝5，则对于当前的神经网络而言，权重参数在数量上的目标值的候选值的集合为{y|y＝n*B，n为正整数}。假设当前层的权重参数在数量上的维度值为8，则可以将该权重参数在数量上的目标值确定为前述候选值的集合中的大于或等于8的某个值，例如10、15或20等。例如，可以从候选值的集合的子集{y|y＝n*B，n是大于或等于5的正整数}中选择最小值10作为目标值。

[0035] 在另外的示例中，在支持神经网络的乘加运算的硬件电路中，可以包括一组或多组乘法器和加法器的布置，每组乘法器和加法器的布置中可以包括一个或多个乘法器以及一个或多个加法器。基于每组乘法器和加法器的布置(例如，用于执行最后一步的加法运算的加法器的数量)以及组数，可以确定第一参数C和第二参数D。在这样的情况下，可以将权重参数在深度或数量上的目标值的候选值的集合确定为C和D的公倍数的集合。

[0036] 在另外的示例中，基于硬件能够直接处理的权重参数的形状或者神经网络的架构设计等，确定权重参数在宽度上的目标值应当为集合{3,5,6,7,9,11}中的一个值。假设当前层的权重参数的宽度为4，则可以将该权重参数在宽度上的目标值确定为前述集合的子集{5,6,7,9,11}中的一个值。例如，可以将目标值确定为子集中的最小值5。

[0037] 根据另外的实施例，例如，在示例方法100的步骤110中所选择的一个或多个维度中，部分或全部维度的对应的目标值可以相互关联地确定。

[0038] 例如，可以确定硬件能够直接处理的权重参数的宽度和高度的组合为集合{(3,3)，(3,6)，(4,7)，(5,5)，(7,7)}中的某个组合，则可以将该集合作为权重参数在宽度和高度上的目标值的候选值数组的集合。假设当前层的权重参数的宽度和高度分别4和4，则可以从前述示例集合的子集{(4,7)，(5,5)，(7,7)}中选择一个候选值数组，用于确定权重参数在宽度和高度上的对应的目标值，其中，可以基于该子集中的每个候选值数组所确定的在宽度和高度上的对应的目标值均分别大于或等于该权重参数在宽度和高度上的维度值。例如，可以选择宽度作为主要参考维度，并且从上述子集中选择包括最小宽度值的候选值数组(4,7)，并于基于该候选值数组将该权重参数在宽度和高度上的对应的目标值分别确定为4和7。在另外的示例中，也可以从上述子集中选择候选值数组中的所有分量的乘积最小的候选值数组(5,5)(其中，4x7＝28，5x5＝25，7x7＝49，候选值数组(5,5)中的所有分量的乘积最小)，基于该候选值数组将该权重参数在宽度和高度上的对应的目标值分别确定为5和5，这相当使得填充的权重参数在宽度和高度两个维度上的“面积”最小且能够完全“覆盖”原来的权重参数。

[0039] 在另外的示例中，基于神经网络的架构设计和/或硬件参数，确定权重参数在宽度、高度和深度上的目标值的组合为集合{(3,3,3)，(3,6,6)，(4,7,6)，(5,5,5)，(7,7,7)}中的某个组合。假设当前层的权重参数的宽度、高度和深度分别4、4和4，则可以从前述示例集合的子集{(4,7,6)，(5,5,5)，(7,7,7)}中选择一个候选值数组，用于确定权重参数在宽度、高度和深度上的对应的目标值，其中，可以基于该子集中的每个候选值数组所确定的在

宽度、高度和深度上的对应的目标值均分别大于或等于该权重参数在宽度、高度和深度上的维度值。例如，可以选择宽度作为主要参考维度，并且从上述子集中选择包括最小宽度值的候选值数组(4,7,6)，并于基于该候选值数组将该权重参数在宽度、高度和深度上的对应的目标值分别确定为4、7和6。在另外的示例中，也可以从上述子集中选择候选值数组中的所有分量的乘积最小的候选值数组(5,5,5)(其中，4x7x6＝168，5x5x5＝125，7x7x7＝343，候选值数组(5,5,5)中的所有分量的乘积最小)，基于该候选值数组将该权重参数在宽度、高度和深度上的对应的目标值分别确定为5、5和5，这相当使得填充的权重参数在宽度、高度和深度三个维度上的“体积”最小且能够完全“包围”原来的权重参数。

[0040] 根据另外的实施例，例如，在示例方法100的步骤110中所选择的一个或多个维度中，部分维度上的对应的目标值可以相互关联地确定，而另一部分维度中的每个维度上的对应的目标值可以独立地确定。

[0041] 根据一个实施例，方法100还可以包括：针对在步骤110中所选择的一个或多个维度的每个维度，或者针对在步骤130中所确定的至少一个维度中的每个维度，确定与该维度相对应的填充模式，其中，所确定的填充模式指示在该维度上对权重参数进行填充的一个或多个填充位置、填充量以及设置填充值的规则中的至少一个。

[0042] 例如，与宽度相对应的填充模式所指示的信息包括在宽度方向上的左侧和/或右侧和/或中间的一个或多个位置处填充一列或多列。相应地，填充量可以为分别在所指示的一个或多个位置处要填充的列数。

[0043] 例如，与高度相对应的填充模式所指示的信息包括在高度方向上的上方和/或下方和/或中间的一个或多个位置处填充一行或多行。相应地，填充量可以为分别在所指示的一个或多个位置处要填充的行数。

[0044] 例如，与深度相对应的填充模式所指示的信息包括在深度方向上的前方和/或后方和/或中间的一个或多个位置处填充一排或多排。相应地，填充量可以为分别在所指示的一个或多个位置处要填充的排数。

[0045] 例如，与数量相对应的填充模式所指示的信息包括在权重参数的序列之前和/或之后和/或之中的一个或多个位置处填充一个或多个填充值。相应地，填充量可以为分别在所指示的一个或多个位置处要填充的填充值的数量。

[0046] 根据一个实施例，可以使用零值对权重参数进行填充。根据不同的实施例，设置填充值的规则也可以包括但不限于使用使用预先定义的非零值进行填充、使用与要填充的位置相邻的位置处的值进行填充等。

[0047] 在本文中，零值可以指在数值上等于0的值，或者所有的分量的值均为0的数组、多元组、矩阵、立方体等，或者可以被视为相当于0的其他形式的数据值。相应地，非零值可以指在数值上不等于0的值，或者部分或全部分量的值不为0的数组、多元组、矩阵、立方体等，或者可以不能被视为相当于0的其他形式的数据值。

[0048] 根据一个实施例，在神经网络的某层的权重参数的大小和形状在填充之后改变或者能够确定权重参数的大小和形状将发生改变的情况下，方法100还可以包括：例如，在诸如宽度和/或高度和/或深度这样的使权重参数的大小和形状发生变化的维度上对权重参数进行填充之后，或者在确定要对权重参数在诸如宽度和/或高度和/或深度这样的可能使权重参数的大小和形状发生变化上进行填充的情况下，针对该层的特征数据，在使权重参

数的大小和形状发生变化的维度(例如,宽度和/或高度和/或深度)上进行填充。例如,在权重参数在宽度上被填充的情况下,对应的特征数据也可以在宽度上进行填充。

[0049] 在一个示例中,可以使在特征数据的要填充的每个维度上的填充位置和填充量分别与在权重参数的对应的维度上的填充位置和填充量相同,或者可以使用针对权重参数所确定的填充模式对特征数据进行填充。

[0050] 根据不同的实施例,针对特征数据的设置填充值的规则可以包括但不限于使用零值进行填充、使用预先定义的非零值进行填充、使用与要填充的位置相邻的位置处的值进行填充、使用随机值或任意值进行填充等。

[0051] 应当理解,图1中所示的方法100也仅仅是示例性的。在另外的示例中,可以在步骤110中确定每个维度的对应的目标值,或者可以将目标值的确定作为单独的步骤。例如,确定填充模式的可选步骤可以安排在对权重参数和/或特征数据进行填充之前的任何一个阶段。例如,对特征数据的填充和对权重参数的填充可以并行执行。

[0052] 图3示出使用根据本申请的实施例的方法对卷积神经网络中的权重参数在数量的维度上进行填充的示例。

[0053] 在图3所示的示例中,基于神经网络的设计架构(例如,由神经网络的算法设计人员基于特定应用和目标所预先定义),针对某个层的特征数据300,预先设置三个卷积核,即卷积核301、卷积核302和卷积核303。相应地,在图3所示的示例中,权重参数在数量的维度上的维度值为3。

[0054] 例如,基于支持神经网络的计算的硬件相关的参数,确定在设计例如5个卷积核的情况下,软件和/或硬件方面的成本和性能相对较好,例如,可以省去加法器的选择/使能电路的设计,或者可以取得相对较好的处理上的并行度,或者确定在接下来的一层要进行池化处理并且例如硬件能够在相应的池化层的特征数据的通道数为5的情况下获得最好或相对较好的处理性通(例如,并行度)。相应地,可以将权重参数在数量的维度上的目标值确定为5。

[0055] 由于权重参数在数量的维度上的维度值(即,3)小于对应的目标值(即,5),所以可以对权重参数在数量上进行填充。

[0056] 在图3的示例中,可以在卷积核301、卷积核302和卷积核303的基础上,添加与卷积核301至303中的每一个具有相同的大小和形状(即,分别在宽度、高度和深度上完全相同)的两个卷积核,即卷积核304和卷积核305(在图3中用虚线表示),使得填充之后的卷积核的总数量为5,从而与所确定的目标值相同。

[0057] 在一个示例中,填充用的卷积核304和卷积核305可以是相同的。换句话说,可以使用一个或多个相同的卷积核进行数量维度上的填充。在另外的示例中,根据需要,也可以选择包含不同值的一个或多个卷积核进行数量维度上的填充,并且可以使所选择的所有填充用的卷积核(例如,图3中的卷积核304和305)与原始设置的卷积核(例如图3中的卷积核301、302和303)具有相同的形状和大小,即,所有的卷积核(包括图3中的卷积核301至305)至少具有相同的宽度、相同的高度和相同的深度。

[0058] 在一个示例中,用于填充的卷积核304和卷积核305可以均为相当于零值的卷积核,即,卷积核304和305中所包含的所有分量值均为零。在另一个示例中,卷积核304和卷积核305也可以从预先定义(例如,在神经网络的设计阶段时由设计人员预先定义)的一组模

板数据(可以包括零值和/或非零值)中选择,或者也可以基于预先定义的填充用卷积核的生成规则来生成,例如,使用一个或多个预定定义的数值(可以是非零值)来生成。

[0059] 在图3所示的示例中,卷积核304和305被填充在卷积核303的下方,或者说,卷积核304和305被顺序地填充在包括卷积核301、卷积核302和卷积核303的有序序列的尾部。在另外的示例中,可以根据需要选择一个或多个另外的填充位置。例如,可以将卷积核304和305填充到包括卷积核301、卷积核302和卷积核303的有序序列的头部,也可以将卷积核304填充在序列的头部(即,图3的示例中的卷积核301的上方)并将卷积核305填充在卷积核302和303之间,或者也可以将卷积核304填充在卷积核301和302之间并将卷积核305填充在卷积核302和303之间,或者也可以将卷积核304和305均填充在卷积核301和302之间。

[0060] 如前文所述,可以基于默认的或者所确定的填充模式来确定在数量维度上的填充位置、填充量以及设置填充值的规则等。

[0061] 图4示出使用根据本申请的实施例的方法对卷积神经网络中的某一层的权重参数在宽度和高度两个维度上进行填充的示例。在图4中,各个卷积核中的数字1、2、……、12等仅用于例示卷积核中的数值(或权重值)的相对位置,而不代表实际的值。为了清楚,在图4中没有示出卷积和特征数据的深度(通道)。

[0062] 在图4所示的示例中,例如,基于神经网络的设计架构,针对特征数据400预先设置的每个卷积核(例如,图4中的卷积核401、卷积核402和卷积核403)的宽度和高度分别为4和3。

[0063] 例如,基于支持神经网络的计算的硬件相关的参数,确定硬件(例如,神经网络的专用加速器)能够直接处理的卷积核的宽度和高度可以分别为5和5。相应地,可以将权重参数在数量的宽度和高度上的目标值分别确定为5和5。

[0064] 由于权重参数在宽度和高度这两个维度上的维度值分别小于对应的目标值,因此可以对权重参数在宽度和高度这两个维度上进行填充。

[0065] 在图4所示的示例中,针对卷积核401、402和403中的每一个,在左侧填充一列并且在下方填充两行(在图4中,使用P来表示所填充的行和列),使得填充后所得到的卷积核401’、402’和403’中的每一个的宽度和高度分别为5和5,从而与所确定的相应的目标值相同。

[0066] 在一个示例中,可以先在宽度上进行填充,即在左侧填充一列,得到的宽度和高度分别为5和3的中间结果;然后,再在高度上对宽度和高度分别为5和3的中间结果进行填充,即在下方填充两行,从而得到最终的宽度和高度分别为5和5的填充后的卷积核。在另外的示例中,也可以先在高度上进行填充,然后在宽度上进行填充。在另外的示例中,也可以同时在宽度和高度上进行填充。

[0067] 在对权重参数的宽度和/或高度进行填充的情况下,权重参数(每个卷积核)的形状改变。相应地,对于对应的特征数据400,可以在宽度和/或高度进行相应的填充,以使得填充后的特征数据400’能够与填充后的卷积核401’、402’和403’相匹配,从而确保神经网络在使用填充后的特征数据400’和填充后的卷积核401’、402’和403’的情况下所能够获得的最终结果与在使用原始的特征数据400和原始的卷积核401、402和403的情况下所能够获得的最终结果相同。在图4所示的示例中,对于卷积核401、402和403中的每一个,在左侧填充一列并且在下方填充两行,并且相应地,对于特征数据400,可以在其左侧填充一列并在

其下方填充两行，由此得到填充后的卷积核401’、402’和403’以及填充后的特征数据400’。

[0068] 在一个示例中，可以使用零值对卷积核和特征数据在宽度和高度上进行填充。例如，在图4所示的示例中，各个卷积核和特征数据的各个填充位置P处的填充值可以为零值。在另外的示例中，对于特征数据也可以使用一个或多个非零值或者通过复制特征数据中与要填充的位置相邻的位置(例如，图4中的特征数据400的最左边的一列和最下方的一行)处的值进行填充。图5示出例如针对图4中的特征数据400使用复制的方式所得到的填充后的特征数据500’的示例，其中，填充位A、B、C、D和E表示分别使用特征数据400中的位置A、B、C、D和E处的值作为填充值。在另外的示例中，根据需要，也可以使用其他值(例如随机值或任意值)对卷积核进行填充。

[0069] 在图4和图5所示的示例中，在每个卷积核和/或特征数据的左侧和下方进行填充。在另外的示例中，也可以选择在卷积核和/或特征数据的右侧和下方进行填充，或者在卷积核和/或特征数据的左侧或右侧以及上方进行填充，或者也可以在卷积核和/或特征数据的左侧或右侧填充一列并且在卷积核和/或特征数据的上方和下方分别填充一行。在一个示例中，可以使针对卷积核的填充方式与针对特征数据的填充方式相同。例如，在针对卷积核在右侧填充一列并在上方和下方分别填充一行的情况下，针对特征数据也在右侧填充一列并在上方和下方分别填充一行。

[0070] 如前文所述，可以基于默认的或者所确定的填充模式，确定在宽度和/或高度维度上的填充位置、填充量以及设置填充值的规则等，并且针对特征数据在宽度和/或高度维度上的填充位置和填充量可以分别与针对权重参数在对应的维度上的填充位置和填充量相同。

[0071] 为了清楚，在图4和图5中没有示出卷积和/或特征数据的深度(通道)。实际上，图4和图5的示例也可以被视为在针对权重参数和/或特征数据在每个通道(或深度)上的填充处理的示例。在考虑深度的情况下，针对权重参数和/或特征数据的所有通道(或深度)，可以在相同的位置处，按照相同的填充值的设置规则，以相同的填充量进行填充。图6示出对宽度、高度和深度均为2的示例卷积核在右侧填充一列以使其成为宽度、高度和深度分别为3、2和2的卷积核的示例。

[0072] 虽然图4和图5示出同时在宽度和高度两个维度上对权重参数和特征数据进行填充，但是如图6所示，根据卷积核的形状和大小、以及与在宽度和/或高度上的对应的目标值之间的比较结果，也可以在宽度和高度之中的一个维度上对权重参数和特征数据进行填充。

[0073] 图7示出使用根据本申请的实施例的方法对卷积神经网络中的某一层的权重参数在深度(通道)上进行填充的示例。

[0074] 在图7所示的示例中，基于神经网络的设计架构，针对神经网络的某一层的通道数(深度)为3的特征数据700，预先设置深度(或通道数)为3的若干卷积核701、702和703。

[0075] 例如，基于支持神经网络的计算的硬件相关的参数，确定在设计例如深度为5的卷积核的情况下，硬件方面的成本和性能相对较好，例如，可以省去加法器的选择/使能电路的设计，或者可以取得相对较好的处理上的并行度，或者可以由硬件直接处理。相应地，可以将权重参数在深度的维度上的目标值确定为5。

[0076] 由于权重参数在深度上的维度值小于对应的目标值，因此可以对卷积核在深度上

进行填充。

[0077] 如图7所示，针对卷积核701、702和703中的每一个，填充两排(在图7中用虚线表示)，使得填充后所得到的卷积核701’、702’和703’中的每一个的深度均为5，从而与所确定的相应的目标值相同。

[0078] 在深度上对权重参数进行填充的情况下，卷积核701、702和703的形状改变。相应地，可以在深度上对该层的特征数据700进行填充，以使得填充后的特征数据700’能够与填充后的卷积核701’、702’和703’相匹配，从而确保神经网络在使用填充后的特征数据700’和填充后的卷积核701’、702’和703’的情况下所能够获得的最终结果与在使用原始的特征数据700和原始的卷积核701、702和703的情况下所能够获得的最终结果相同。在图7所示的示例中，对于卷积核701、702和703中的每一个，在后方填充两排(每个填充后的卷积核在每个深度上具有相同的宽度和高度)，并且相应地，可以在特征数据700的后方填充两排(填充后的特征数据700’在每个深度上具有相同的宽度和高度)，由此得到填充后的卷积核701’、702’和703’以及填充后的特征数据700’。

[0079] 在一个示例中，可以使用零值对卷积核在深度上进行填充。例如，在图7中的卷积核701’、702’和703’中的每一个的后方所填充的两排的数据可以全部是0。在另外的示例中，根据需要，也可以使用其他值对卷积核在深度上进行填充。类似地，可以使用零值或其他值对特征数据在深度上进行填充。

[0080] 在图7的示例中，在每个卷积核和特征数据的后方进行填充。在另外的示例中，也可以选择在前方填充两排，或者可以某两个连续的通道之间填充两排，或者可以在通道1和通道2之间填充一排并且在通道2和通道3之间填充一排，或者可以某两个连续的通道之间填充一排并且在前方或后方填充一排。在一个示例中，可以使针对卷积核的填充方式与针对特征数据的填充方式相同。例如，在针对卷积核在前方填充一排并且在通道2和通道3之间填充一排的情况下，针对特征数据也在前方填充一排并且在通道2和通道3之间填充一排。

[0081] 如前文所述，可以基于默认的或者所确定的填充模式，确定在深度维度上的填充位置、填充量以及设置填充值的规则等，并且针对特征数据在深度维度上的填充位置和填充量可以分别与针对权重参数在对应的维度上的填充位置和填充量相同。另外，用于填充特征数据的填充值可以根据用于填充权重参数的填充值的来确定，也可以独立地确定。

[0082] 图3至图7示出使用根据本申请的实施例的方法对权重参数和/或特征数据在数量、宽度、高度和深度等维度上进行填充的示例。虽然图2至7中所采用的特征数据和权重参数的表示方式可能不同，但这并不意味着这些特征数据和/或权重参数的实际形式一定是不同的。应当理解，采用的不同的形式表示特征数据和/或权重参数仅仅是出于方便和/或强调的目的。

[0083] 还应当理解，根据本申请的实施例的方法不局限于图3至图7所示的示例，对于卷积神经网络或者其他类型的神经网络中的某个层的权重参数和特征数据，可以根据需要考虑权重参数和/或特征数据实际具有的一个或多个维度，或者可以等效地从一个或多个维度观察权重参数和/或特征数据，并根据需要使用根据本申请的实施例的方法对权重参数和/或特征数据进行填充，以获得具有规范形式的权重参数和/或特征数据。例如，对于卷积神经网中可能存在的池化层，例如可以对特征数据设置一个池化核(也被称为感受野)，并

使用这个或这些池化核对特征数据的每个通道分别进行池化处理。相应地，可以将该池化核视为具有深度这样的维度，并且其深度值等于特征数据的通道数(或深度值)，并且根据需要可以对该池化层的池化核和/或特征数据在例如深度上进行填充，例如可以采用在深度上复制池化核和/或特征数据的方式进行填充。

[0084] 通过使用根据本申请的实施例的方法对权重参数和/或特征数据进行填充，能够获得具有规范形式的权重参数和/或特征数据，从而能够简化神经网络的架构设计以及支持神经网络的运算的硬件(例如，专用加速器、乘加运算单元等)的设计，能够避免为了处理不同的形式的权重参数所需的额外处理以及可能导致的错误，并且能够提高软件和/或硬件的处理性能(例如，提高处理的并行度或硬件的利用率等)。

[0085] 根据本申请的实施例的方法(例如图1所示的示例方法100)可以用于针对神经网络中的一层或多层中的特征数据和权重参数的预处理。例如，可以在由神经网络架构中的编译器层向硬件层提交数据和指令之前，针对神经网络中的一层或多层中的特征数据和权重参数，使用根据本申请的实施例的方法进行检查和填充。也可以神经网络中的一层或多层中的每一层的运算之前，使用根据本申请的实施例的方法对该层的特征数据和权重参数进行检查和填充。

[0086] 虽然上文以卷积神经网络为例描述了本申请的用于适配神经网络的参数的方法的原理，但是应当理解，本申请的原理可以适用于其他类型的神经网络。例如，如果权重参数是一个或多个单值数据，则可以使用例如图1所示的示例方法100并且参照例如图3所示的示例，在数量上对权重参数进行填充。例如，如果权重参数是一个或多个二维数据，则可以使用例如图1所示的示例方法100并且参照例如图4和图5所示的示例，在二维数据的两个维度上对权重参数进行填充。相应地，可以根据需要和神经网络的类型，确定权重参数在相应的维度上的目标值的确定方式和填充方式。

[0087] 图8示出根据本申请的实施例的可以用于对神经网络的参数进行适配的装置的框图。

[0088] 如图8所示，示例装置800可以包括一个或多个处理器810。处理器810可以是具有数据处理能力和/或指令执行能力的任何形式的处理单元，例如通用CPU、GPU或者专用的神经网络处理器或加速器等。例如，处理器810可以执行根据本申请的实施例的用于对神经网络的参数进行适配的方法。另外，处理器810还可以控制装置810中的其他部件，以执行所期望的功能。

[0089] 处理器810可以通过总线系统和/或其他形式的连接机构(未示出)与存储器830以及I/O接口830相连。

[0090] 存储器830可以包括各种形式的计算机可读写存储介质，例如易失性存储器和/或非易失性存储器。所述易失性存储器例如可以包括随机存取存储器(RAM)和/或高速缓冲存储器(cache)等。所述非易失性存储器例如可以包括只读存储器(ROM)、硬盘、闪速存储器等。可读写存储介质例如可以包括但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。例如，在配合神经网络专用处理器使用的情况下，存储器830也可以是承载专用处理器的芯片上的RAM。存储器830可以包括用于指示装置800执行根据本申请实施例的对神经网络的参数进行适配的方法的程序指令。

[0091] I/O接口830可以用于向处理器810提供参数或数据并且输出经过处理器810处理

的结果数据。

[0092] 图9示出根据本申请的另外的实施例的可以用于对神经网络的参数进行适配的装置的框图。

[0093] 如图9所示，示例装置900可以包括检测器910和填充器920。检测器910可以被配置为针对神经网络的至少一层中的每一层的权重参数所选择的一个或多个维度，确定该权重参数在每个维度上的维度值和对应的目标值。填充器920可以被配置为在权重参数在至少一个维度上的维度值小于对应的目标值的情况下，对所述权重参数进行填充，使得在填充之后所获得的权重参数在每个维度上的维度值等于对应的目标值。

[0094] 示例装置900还可以包括处理器930(例如，通用CPU、GPU，或者神经网络的专用处理器或加速器)，接收来自填充器920的输出数据并进行与神经网络相关的运算。

[0095] 应当理解，图8和图9所示的装置800和900仅是示例性的，而非限制性的。根据需要，根据本申请的实施例的可以用于对神经网络的参数进行适配的装置可以具有其他部件和/或结构。

[0096] 除非上下文清楚地另有要求，否则贯穿说明书和权利要求书，措词“包括”、“包含”等应当以与排他性或穷尽性的意义相反的包括性的意义来解释，也就是说，应当以“包括但不限于”的意义来解释。另外，措词“在本文中”、“上文”、“下文”以及相似含义的措词在本申请中使用时应当指作为整体的本申请，而不是本申请的任何具体部分。在上下文允许时，在使用单数或复数的以上描述中的措词也可以分别包括复数或单数。关于在提及两个或多个项目的列表时的措词“或”，该措词涵盖该措词的以下解释中的全部：列表中的任何项目，列表中的所有项目，以及列表中的项目的任何组合。

[0097] 本发明实施例的以上详细描述不打算是穷尽性的或者将本发明局限于上文所公开的确切形式。尽管以上出于说明的目的而描述了本发明的具体实施例和示例，但是如本领域技术人员将认识到的那样，在本发明范围内可能有各种等效的修改。例如，尽管处理或块以给定的次序呈现，但是替代的实施例可以以不同的次序执行具有这些步骤的处理或者以不同的次序采用具有这些块的系统，并且一些处理或块可以被删除、移动、添加、细分、组合和/或修改。这些处理或块中的每个可以以各种不同的方式来实现。另外，虽然处理或块有时被示为串行执行，但是替代地，这些处理或块也可以并行执行，或者可以在不同时间执行。

[0098] 可以将在本文中所提供的本发明的教导应用于其他系统，而不必是上述的系统。可以组合上述的各个实施例的元件和动作，以提供另外的实施例。

[0099] 虽然已经描述了本发明的一些实施例，但是这些实施例仅作为示例而呈现，而不打算限制本申请的范围。实际上，在本文中所描述的新颖方法和系统可以以多种其他形式来实施。另外，可以在不脱离本申请的范围的情况下，在本文中所描述的方法和系统的形式上做出各种省略、替换和改变。

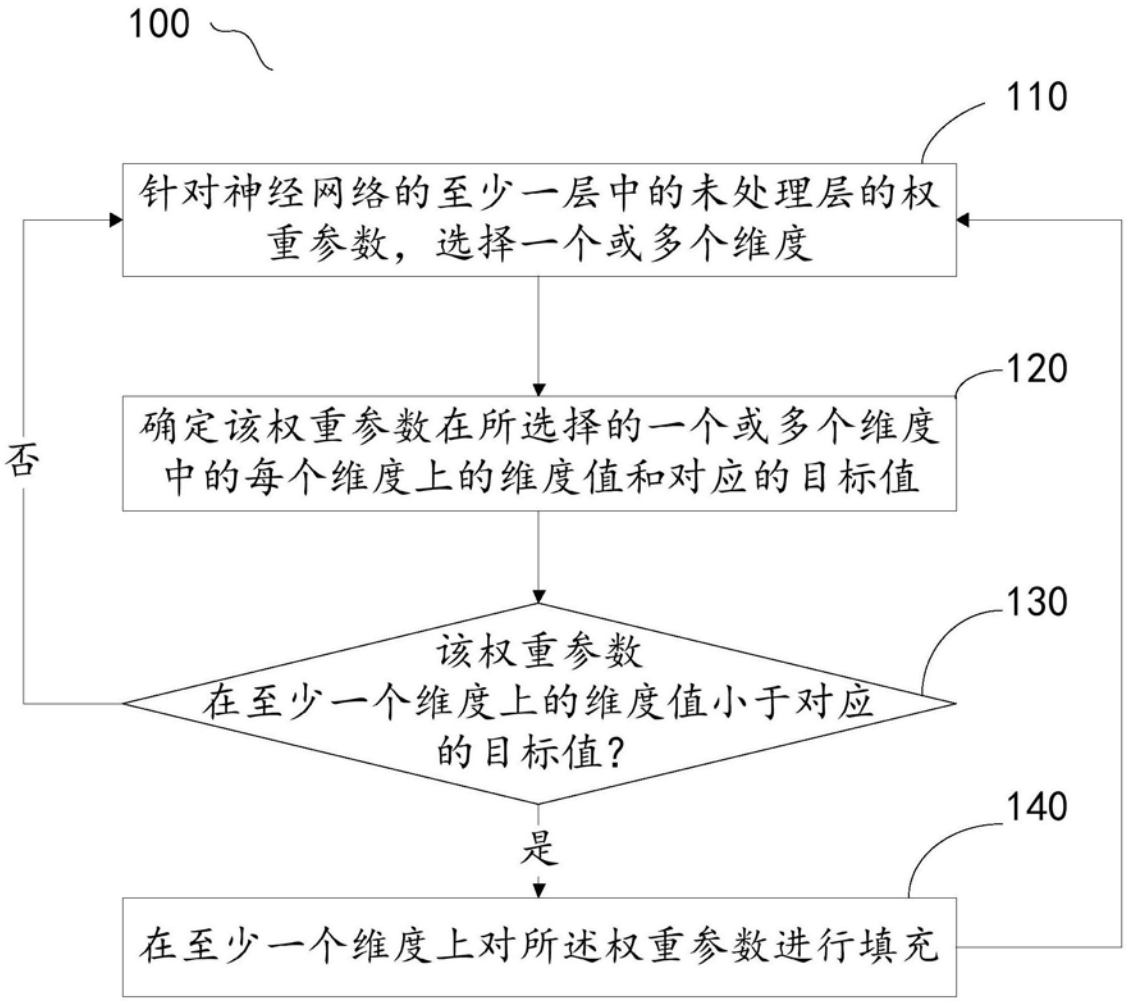

100
110
针对神经网络的至少一层中的未处理层的权重参数，选择一个或多个维度
120
确定该权重参数在所选择的一个或多个维度中的每个维度上的维度值和对应的目标值
130
该权重参数在至少一个维度上的维度值小于对应的目标值？
否
是
140
在至少一个维度上对所述权重参数进行填充

图1

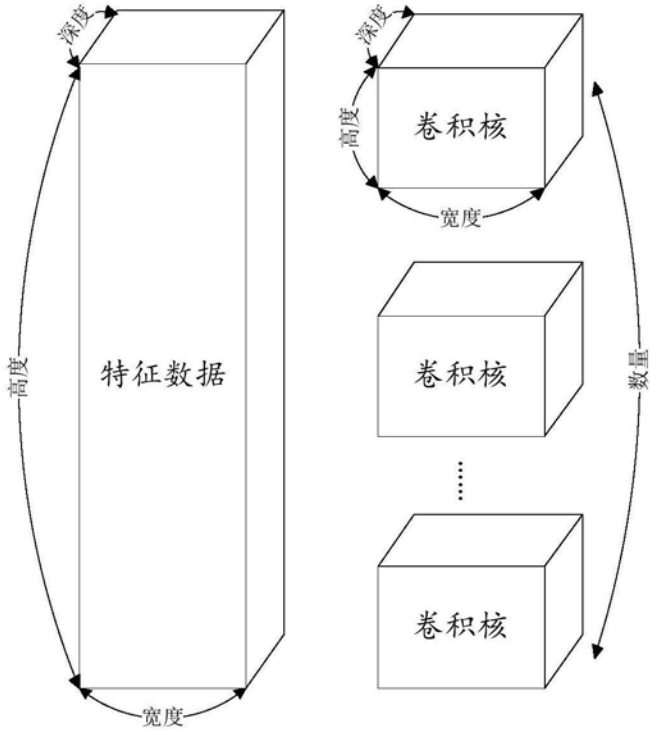


图2

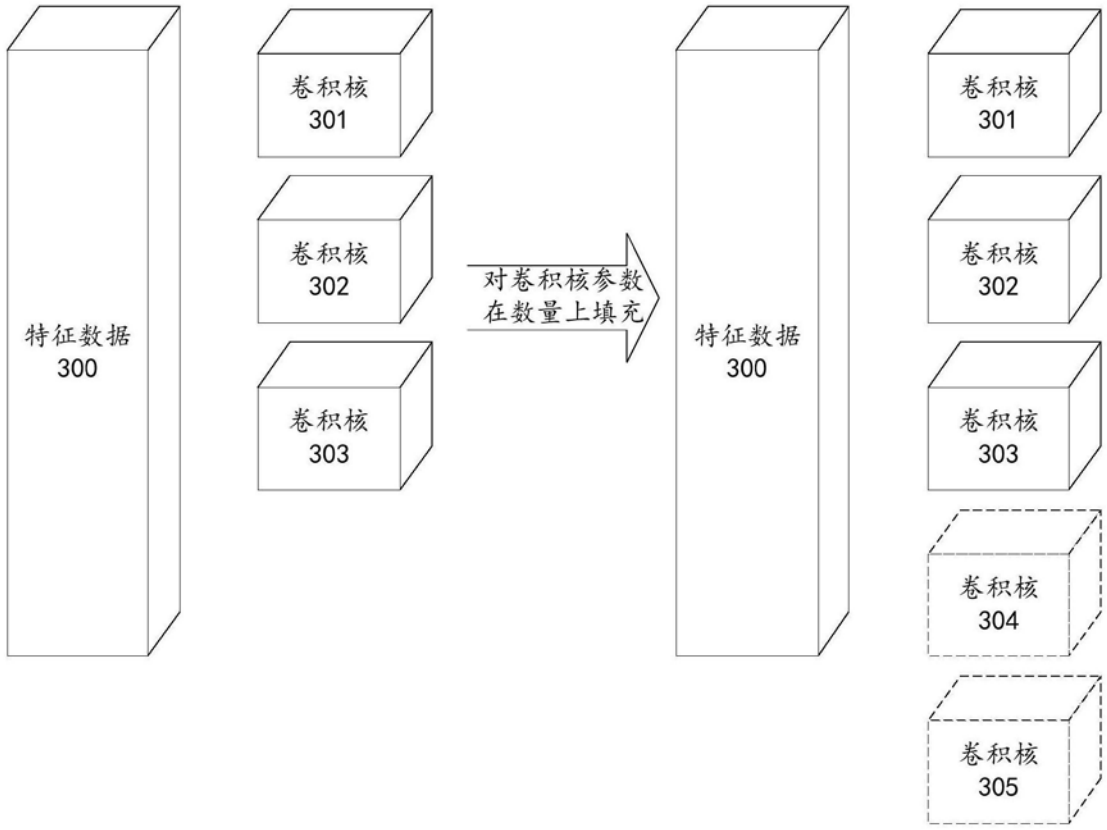

特征数据
300
卷积核
301
卷积核
302
卷积核
303
对卷积核参数
在数量上填充
特征数据
300
卷积核
301
卷积核
302
卷积核
303
卷积核
304
卷积核
305

图3

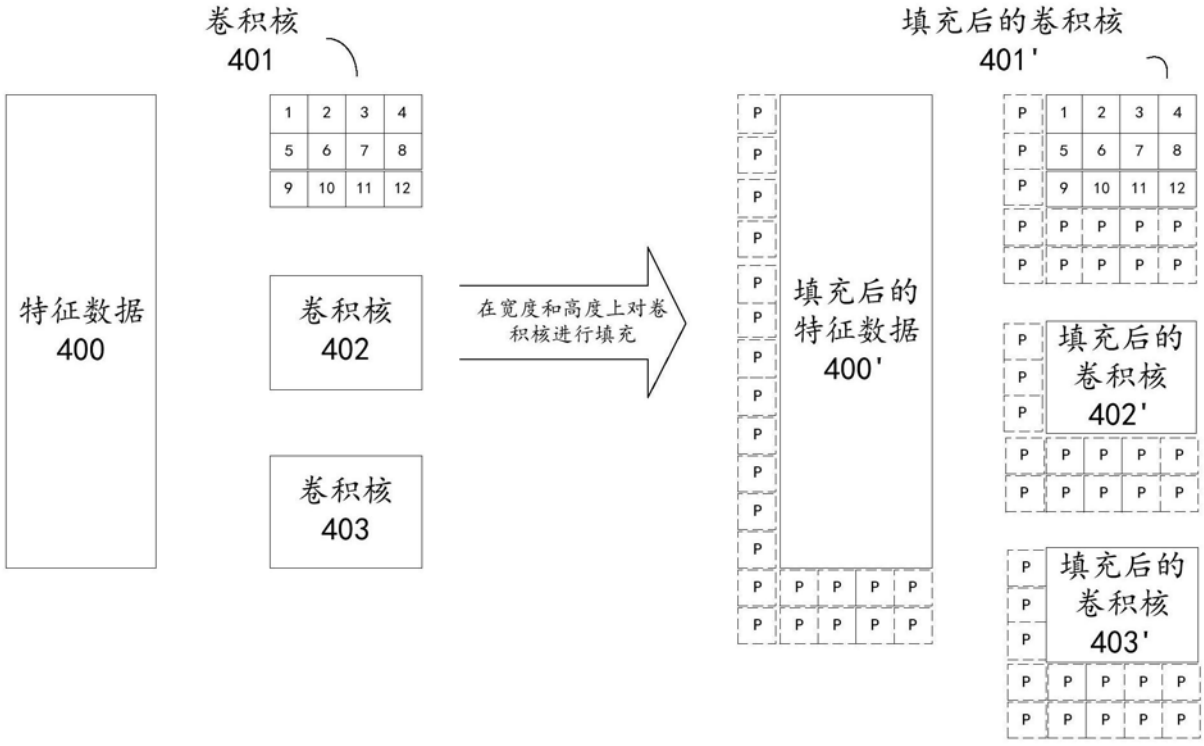

卷积核
401
特征数据
400
1 2 3 4
5 6 7 8
9 10 11 12
卷积核
402
卷积核
403
在宽度和高度上对卷
积核进行填充
填充后的
特征数据
400'
填充后的卷积核
401'
填充后的
卷积核
402'
填充后的
卷积核
403'

图4

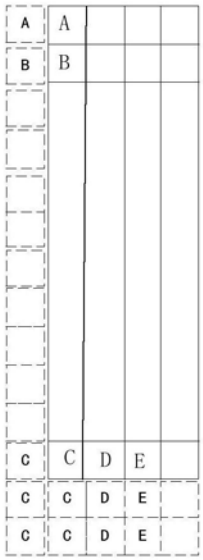

A
B
C
C
C
C
D
E
C
D
E
C
D
E
填充后的特征数据
500'

图5

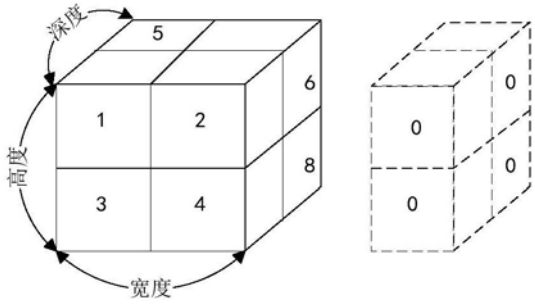

深度
高度
宽度
5
6
1
2
8
3
4
0
0
0
0
使用零值在右侧填充一列

图6

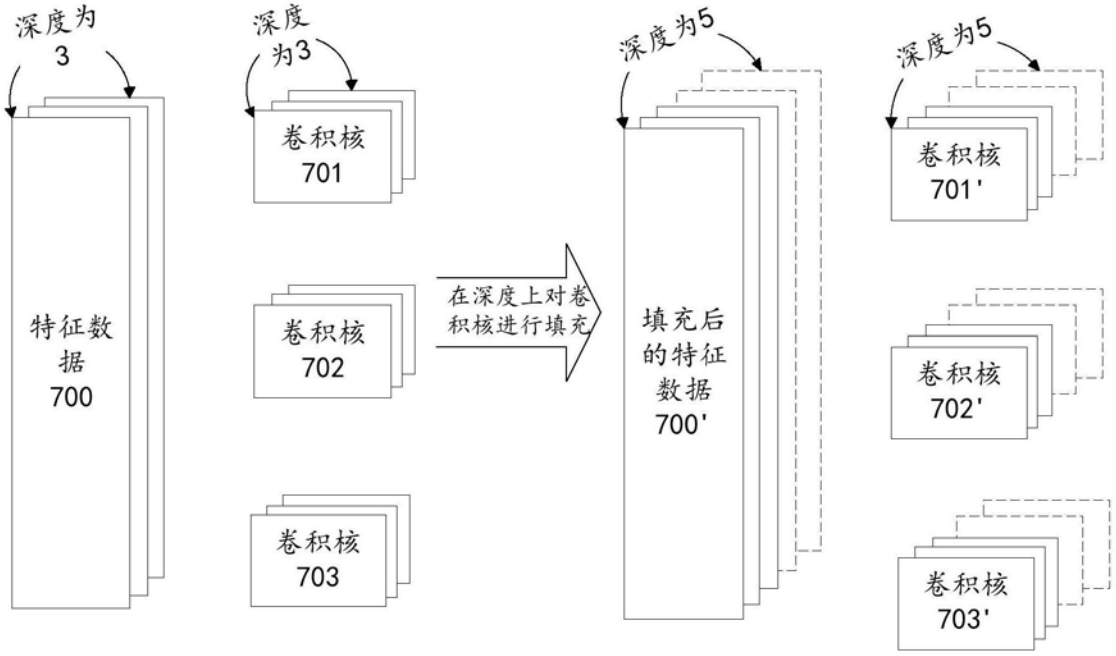

深度为
3
特征数
据
700
深度
为3
卷积核
701
卷积核
702
卷积核
703
在深度上对卷
积核进行填充
深度为5
填充后
的特征
数据
700'
深度为5
卷积核
701'
卷积核
702'
卷积核
703'

图7

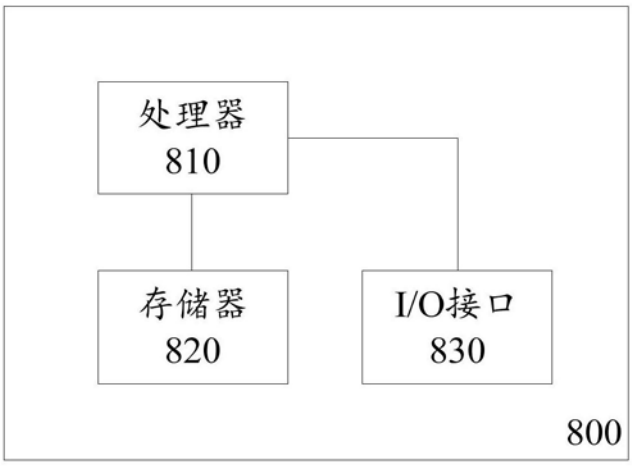

处理器
810
存储器
820
I/O接口
830
800

图8

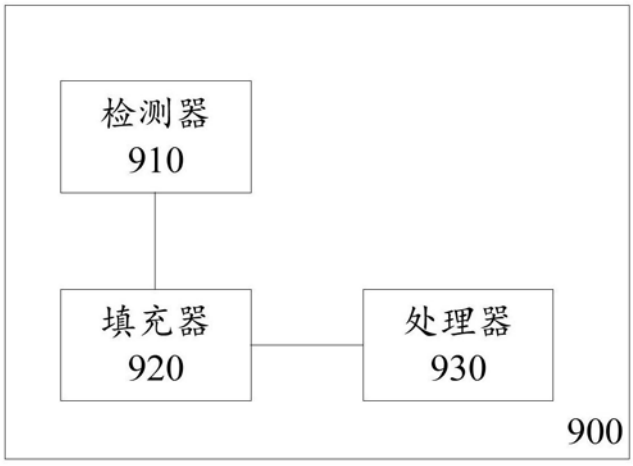


图9