(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107480770 A
(43)申请公布日 2017.12.15

(21)申请号 201710624244.0

(22)申请日 2017.07.27

(71)申请人 中国科学院自动化研究所
地址 100080 北京市海淀区中关村东路95号

(72)发明人 程健 贺翔宇 胡庆浩

(74)专利代理机构 北京瀚仁知识产权代理事务所(普通合伙) 11482
代理人 郭文浩 王世超

(51)Int.Cl.
G06N 3/04(2006.01)
G06N 3/08(2006.01)

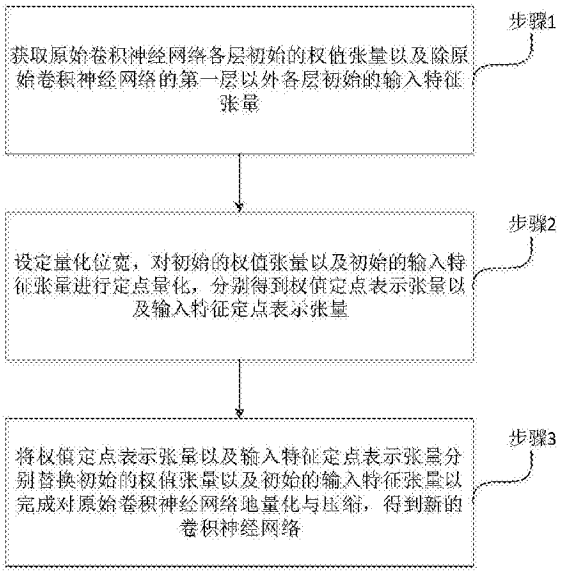
权利要求书2页 说明书7页 附图3页

(54)发明名称

可调节量化位宽的神经网络量化与压缩的方法及装置

(57)摘要

本发明涉及神经网络技术领域，具体提出一种卷积神经网络量化与压缩的方法及装置。旨在解决现有对神经网络量化与压缩的方法对网络性能造成较大损失的问题。本发明的方法包括获取原始卷积神经网络的权值张量和输入特征张量，并基于预先设定的量化位宽，对权值张量和输入特征张量进行定点量化，并将得到的权值定点表示张量以及输入特征定点表示张量替换原来的权值张量和输入特征张量，得到对原始卷积神经网络量化与压缩后的新的卷积神经网络。本发明能够根据不同的任务需要灵活地调整位宽，无需调整算法结构和网络结构即可实现对卷积神经网络的量化与压缩，减少对内存以及存储资源的占用。本发明还提出一种存储装置和处理装置，具有上述有益效果。



CN 107480770 A

1.一种卷积神经网络量化与压缩的方法,其特征在于,包括:

获取原始卷积神经网络卷积层初始的权值张量、以及除所述原始卷积神经网络的第一层以外各层初始的输入特征张量;

基于预先设定的量化位宽,对所述初始的权值张量以及所述初始的输入特征张量进行定点量化,分别得到权值定点表示张量以及输入特征定点表示张量;

利用所述权值定点表示张量以及所述输入特征定点表示张量,分别替换所述初始的权值张量以及所述初始的输入特征张量,得到对所述原始卷积神经网络量化与压缩后的新的卷积神经网络。

2.根据权利要求1所述的卷积神经网络量化与压缩的方法,其特征在于,所述对所述初始的权值张量以及所述初始的输入特征张量进行定点量化,其方法为:

对所述原始卷积神经网络中各卷积层,将所述初始的输入特征张量进行由浮点数到定点数的量化,获得与所述初始的输入特征张量近似的第一输入特征张量,所述第一输入特征张量为输入特征定点表示张量,直至所述原始卷积神经网络中所有卷积层被遍历;

对所述原始卷积神经网络中各卷积层,将所述初始的权值张量进行由浮点数到定点数的量化,获得权值缩放系数以及第一权值张量,直至所述原始卷积神经网络中所有卷积层被遍历。

3.根据权利要求2所述的卷积神经网络量化与压缩的方法,其特征在于,将所述初始的输入特征张量与权值张量进行由浮点数到定点数的量化,其方法为:

使用最近邻法对所述初始的输入特征张量与权值张量进行由浮点数到定点数的量化。

4.根据权利要求3所述的卷积神经网络量化与压缩的方法,其特征在于,所述得到权值定点表示张量,其方法包括:

基于预先设定的量化位宽直接量化所述初始的权值张量得到权值定点表示张量,或者基于预先设定的量化位宽量化所述初始的权值张量,将得到的结果进行迭代训练后,得到权值定点表示张量。

5.根据权利要求4所述的卷积神经网络量化与压缩的方法,其特征在于,所述将得到的结果进行迭代训练,其方法为:

对所述权值缩放系数进行初始化;

将所述第一权值张量依照所述权值缩放系数进行缩放,将缩放结果按照所述量化位宽进行量化,得到第二权值张量;

求解所述权值缩放系数,直至所述权值缩放系数收敛于设定的稳定值,得到稳定的权值缩放系数;

对所述第二权值张量与所述稳定的权值缩放系数进行乘积操作,得到所述权值定点表示张量。

6.根据权利要求5所述的卷积神经网络量化与压缩的方法,其特征在于,所述求解所述权值缩放系数,其方法为:

将所述第二权值张量与所述权值缩放系数代入公式计算对所述第一权值张量的最优表示,根据计算结果更新所述权值缩放系数,具体公式为:

$$\alpha\_gQ\_g \cong W\_g,$$

其中,α_g表示权值缩放系数,Q_g表示第二权值张量,W_g表示第一权值张量。

7.根据权利要求6所述的卷积神经网络量化与压缩的方法，其特征在于，所述第二权值张量与所述权值缩放系数一一对应。

8.根据权利要求7所述的卷积神经网络量化与压缩的方法，其特征在于，在分别替换所述初始的权值张量以及所述初始的输入特征张量之后，在得到新的卷积神经网络之前，该方法还包括：

对所述原始卷积神经网络中各卷积层，对所述权值定点表示张量与所述输入特征定点表示张量进行卷积运算，得到第一输出特征张量，直至所述原始卷积神经网络中所有卷积层被遍历；

对所述原始卷积神经网络中各卷积层，对所述第一输出特征张量与所述稳定的权值缩放系数进行乘积运算，得到最终的输出特征张量，直至所述原始卷积神经网络中所有卷积层被遍历，所述最终的输出特征张量构成所述新的卷积神经网络。

9.一种存储装置，其中存储有多条程序，其特征在于，所述程序适于由处理器加载并执行以实现权利要求1-8任一项所述的卷积神经网络量化与压缩的方法。

10.一种处理装置，包括处理器、存储设备；处理器，适于执行各条程序；存储设备，适于存储多条程序；其特征在于，所述程序适于由处理器加载并执行以实现权利要求1-8任一项所述的卷积神经网络量化与压缩的方法。

可调节量化位宽的神经网络量化与压缩的方法及装置

技术领域

[0001] 本发明属于神经网络技术领域，具体提供一种卷积神经网络量化与压缩的方法及装置。

背景技术

[0002] 近年来，随着卷积神经网络在目标检测识别领域的发展，其检测正确率已经达到商用水平，与此同时，便携设备(例如移动终端、智能设备)的高速发展，让研究者看到了将卷积神经网络与便携设备相结合的契机。然而，基于卷积神经网络的目标识别往往依赖于高性能的GPU(Graphics Processing Unit，图形处理器)设备，需要巨大的运算量，消耗较大的内存，如果在智能手机或者嵌入式设备上运行卷积神经网络模型，将迅速消耗智能手机或者嵌入式设备有限的内存资源、硬盘存储资源以及电量，这对用户来说，显然是难以接受的。

[0003] 为解决上述问题，现有技术的工作主要聚焦于如何缩减卷积神经网络的规模以实现高效地网络训练以及如何在运行时加速针对卷积神经网络的量化与压缩。通过用定点数表示卷积神经网络模型，将相应地减少内存及存储资源的占用，对于FPGA(Field Programmable Gate Array，现场可编程门阵列)等专用器件来说将会节省资源，节约传输时间；采用较少的显存位宽也可以压缩逻辑电路的尺寸，在单位时间内，可进行运算的数目将会增多。但是现有技术的方法有些仅考虑压缩存储问题，有些量化或者压缩问题的解决方案对于硬件实现并不友好，有些追求较低的显存位宽导致网络性能地大幅下降，对大型卷积神经网络的所有层同时进行可变位宽的量化与压缩且对性能不造成较大损失的工作还有待研究。

[0004] 因此，如何提供一种解决上述技术问题的方案是本领域技术人员目前需要解决的问题。

发明内容

[0005] 为了解决现有技术中的上述问题，即为了解决现有技术对卷积神经网络进行量化与压缩时对卷积神经网络的性能造成较大损失的问题，本发明的一方面提供了一种卷积神经网络量化与压缩的方法，包括：

[0006] 获取原始卷积神经网络卷积层初始的权值张量、以及除所述原始卷积神经网络的第一层以外各层初始的输入特征张量；

[0007] 基于预先设定的量化位宽，对所述初始的权值张量以及所述初始的输入特征张量进行定点量化，分别得到权值定点表示张量以及输入特征定点表示张量；

[0008] 利用所述权值定点表示张量以及所述输入特征定点表示张量，分别替换所述初始的权值张量以及所述初始的输入特征张量，得到对所述原始卷积神经网络量化与压缩后的新的卷积神经网络。

[0009] 在上述方法的优选技术方案中，所述对所述初始的权值张量以及所述初始的输入

特征张量进行定点量化，其方法为：

[0010] 对所述原始卷积神经网络中各卷积层，将所述初始的输入特征张量进行由浮点数到定点数的量化，获得与所述初始的输入特征张量近似的第一输入特征张量，所述第一输入特征张量为输入特征定点表示张量，直至所述原始卷积神经网络中所有卷积层被遍历；

[0011] 对所述原始卷积神经网络中各卷积层，将所述初始的权值张量进行由浮点数到定点数的量化，获得权值缩放系数以及第一权值张量，直至所述原始卷积神经网络中所有卷积层被遍历。

[0012] 在上述方法的优选技术方案中，将所述初始的输入特征张量与权值张量进行由浮点数到定点数的量化，其方法为：

[0013] 使用最近邻法对所述初始的输入特征张量与权值张量进行由浮点数到定点数的量化。

[0014] 在上述方法的优选技术方案中，所述得到权值定点表示张量，其方法包括：

[0015] 基于预先设定的量化位宽直接量化所述初始的权值张量得到权值定点表示张量，或者基于预先设定的量化位宽量化所述初始的权值张量，将得到的结果进行迭代训练后，得到权值定点表示张量。

[0016] 在上述方法的优选技术方案中，所述将得到的结果进行迭代训练，其方法为：

[0017] 对所述权值缩放系数进行初始化；

[0018] 将所述第一权值张量依照所述权值缩放系数进行缩放，将缩放结果按照所述量化位宽进行量化，得到第二权值张量；

[0019] 求解所述权值缩放系数，直至所述权值缩放系数收敛于设定的稳定值，得到稳定的权值缩放系数；

[0020] 对所述第二权值张量与所述稳定的权值缩放系数进行乘积操作，得到所述权值定点表示张量。

[0021] 在上述方法的优选技术方案中，所述求解所述权值缩放系数，其方法为：

[0022] 将所述第二权值张量与所述权值缩放系数代入公式计算对所述第一权值张量的最优表示，根据计算结果更新所述权值缩放系数，具体公式为：

[0023] $$\alpha\_gQ\_g \cong W\_g,$$

[0024] 其中，α_g表示权值缩放系数，Q_g表示第二权值张量，W_g表示第一权值张量。

[0025] 在上述方法的优选技术方案中，所述第二权值张量与所述权值缩放系数一一对应。

[0026] 在上述方法的优选技术方案中，在分别替换所述初始的权值张量以及所述初始的输入特征张量之后，在得到新的卷积神经网络之前，该方法还包括：

[0027] 对所述原始卷积神经网络中各卷积层，对所述权值定点表示张量与所述输入特征定点表示张量进行卷积运算，得到第一输出特征张量，直至所述原始卷积神经网络中所有卷积层被遍历；

[0028] 对所述原始卷积神经网络中各卷积层，对所述第一输出特征张量与所述稳定的权值缩放系数进行乘积运算，得到最终的输出特征张量，直至所述原始卷积神经网络中所有卷积层被遍历，所述最终的输出特征张量构成所述新的卷积神经网络。

[0029] 本发明的另一个方面，提供了一种存储装置，其中存储有多条程序，所述程序适于

由处理器加载并执行以实现上述所述的卷积神经网络量化与压缩的方法。

[0030] 本发明的第三方面，提供了一种处理装置，包括处理器、存储设备；处理器，适于执行各条程序；存储设备，适于存储多条程序；所述程序适于由处理器加载并执行以实现上述所述的卷积神经网络量化与压缩的方法。

[0031] 本发明提供了一种卷积神经网络量化与压缩的方法，包括获取原始卷积神经网络卷积层初始的权值张量、以及除原始卷积神经网络的第一层以外各层初始的输入特征张量；基于预先设定的量化位宽，对初始的权值张量以及初始的输入特征张量进行定点量化，分别得到权值定点表示张量以及输入特征定点表示张量；利用权值定点表示张量以及输入特征定点表示张量，分别替换初始的权值张量以及初始的输入特征张量，得到对原始卷积神经网络量化与压缩后的新的卷积神经网络。

[0032] 本领域技术人员能够理解的是，本发明通过对卷积神经网络的权值利用缩放系数及定点权值进行近似，对卷积神经网络的输入特征张量利用高位位宽定点数进行近似，仅需根据已有网络模型参数，将权值张量与输入特征张量进行定点量化，并将定点量化后的参数替换原来的参数即可得到新的网络模型。能够根据不同的任务需要灵活地调整位宽，无需调整算法结构和网络结构，在实现卷积网络所有层进行可变位宽地量化与压缩的同时，不对卷积神经网络的性能造成较大的损失，减少了对内存以及存储资源的占用，节约传输时间，取得性能及网络大小之间的平衡，更加有利于卷积神经网络应用于便携设备。

附图说明

[0033] 图1为本发明提供的卷积神经网络量化与压缩的方法的流程示意图；

[0034] 图2为本发明提供的卷积神经网络用于图像分类过程的示意图；

[0035] 图3为本发明提供的卷积神经网络图像分类过程中卷积层的卷积操作示意图；

[0036] 图4为本发明提供的对权值张量进行定点量化操作的示意图；

[0037] 图5为本发明提供的获得新的卷积神经网络的过程示意图。

具体实施方式

[0038] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0039] 如图1所述，为本发明提供的卷积神经网络量化与压缩的方法的流程示意图，包括：

[0040] 步骤1：获取原始卷积神经网络卷积层初始的权值张量、以及除原始卷积神经网络的第一层以外各层初始的输入特征张量；

[0041] 在实际应用中，如图2所示，为本发明提供的卷积神经网络用于图像分类过程示意图，卷积神经网络包含多个卷积层和多个全连接层，为了表述形式的统一，将所有的全连接层视为特殊形式的卷积层，每一卷积层都有一组卷积核，该组卷积核共同组成该层的权值张量。其中，卷积神经网络的第一层全连接层视为该层的卷积核与该层的输入特征张量大小相同的卷积层，除第一层以外，其他全连接层视为G×1×1×T的卷积层，其中G表示全连

接层输出节点的个数，T表示全连接层输入节点的个数，该全连接层对应的卷积层的卷积核的宽度与高度均为1。由于第一层全连接层的输入为图片的原始数据，数值分布区间为0-255，若将其量化为低位位宽，网络性能将受到较大的影响，所以不获取卷积神经网络第一层的输入特征张量。

[0042] 步骤2：基于预先设定的量化位宽，对初始的权值张量以及初始的输入特征张量进行定点量化，分别得到权值定点表示张量以及输入特征定点表示张量；

[0043] 具体地，如图3所示，为本发明提供的卷积神经网络图像分类过程中卷积层的卷积操作示意图，设定量化张量的数目为G，设定量化位宽为L，其中，G和L均为不小于1的正整数，设定初始的权值张量为W，设定初始的输入特征张量为X，W为G×w×h×T的四维张量，X为N×m×n×T的四维张量，其中，G表示输出特征张量的通道数，N表示输入样本数量，T表示输入特征张量的通道数，w表示卷积核的宽度，h表示卷积核的高度，m表示输入特征张量的宽度，n表示输入特征张量的高度。权值张量W和输入特征张量除了可以是四维张量以外，还可以是其他维度的张量，这里不做限定。在实际应用中，为了方便表示，可以将卷积核的宽度和高度合并，记住维度D，可以将四维张量转换成三维张量，即G×D×T。

[0044] 具体地，定点数是指小数点位置固定的数，浮点数是指小数点位置可以变动的数，其中，相对于浮点数，定点数要求的处理硬件更为简单，更能减少对便携设备的负担。根据设定的量化位宽，分别对初始的权值张量以及初始的输入特征张量进行由浮点数到定点数地量化，其中，对初始的权值张量地量化将得到一组权值缩放系数以及一组第一权值张量，权值缩放系数与第一权值张量一一对应，对初始的输入特征张量的量化将得到一组与输入特征张量近似的第一输入特征张量。

[0045] 步骤3：利用所述权值定点表示张量以及所述输入特征定点表示张量，分别替换所述初始的权值张量以及所述初始的输入特征张量，得到对所述原始卷积神经网络量化与压缩后的新的卷积神经网络。

[0046] 根据步骤2中得到的权值缩放系数和第一权值张量，将两者进行乘积计算，得到第二权值张量，并将权值缩放系数进行迭代直至收敛于稳定的权值缩放系数，再将稳定的权值缩放系数与第二权值张量进行乘积，并将得到的结果替换初始的权值张量；根据步骤2中得到的第一输入特征张量，将得到的第一输入特征张量利用较高位宽的定点数进行近似，并将得到的结果替换初始的输入特征张量，从而达到对原始卷积神经网络进行量化与压缩的目的，进而得到新的卷积神经网络，新的卷积神经网络能够在对其所有层同时进行可变位宽地量化与压缩同时对性能不造成较大损失。

[0047] 本发明通过对卷积神经网络的权值利用缩放系数及定点权值进行近似，对卷积神经网络的输入特征张量利用高位位宽定点数进行近似，仅需根据已有网络模型参数，将权值张量与输入特征张量进行定点量化，并将定点量化后的参数替换原来的参数即可得到新的网络模型。能够根据不同的任务需要灵活地调整位宽，无需调整算法结构和网络结构，在实现卷积网络所有层进行可变位宽地量化与压缩的同时，不对卷积神经网络的性能造成较大的损失，减少了对内存以及存储资源的占用，节约传输时间，取得性能及网络大小之间的平衡，更加有利于卷积神经网络应用于便携设备。

[0048] 作为一种优选的实施例，对初始的权值张量以及初始的输入特征张量进行定点量化，其方法为：

对原始卷积神经网络中各卷积层，将初始的输入特征张量进行由浮点数到定点数的量化，获得与初始的输入特征张量近似的第一输入特征张量，第一输入特征张量为输入特征定点表示张量，直至原始卷积神经网络中所有卷积层被遍历；

对原始卷积神经网络中各卷积层，将初始的权值张量进行由浮点数到定点数的量化，获得权值缩放系数以及第一权值张量，直至原始卷积神经网络中所有卷积层被遍历。

[0049] 在实际应用中，根据设定的定点量化的位宽，对初始的输入特征张量为X进行定点量化，获得一组(N个)定点量化后的对X近似的第一输入特征张量$\tilde{X}_1$，$\tilde{X}\_2 \ldots \tilde{X}\_N$，其中，N为不小于1的正整数，对初始的权值张量W进行定点量化，获得一组(g个)权值缩放系数α_1，α_2…α_g以及一组(g个)定点量化的第一权值张量W_1，W_2…W_g，其中N和g均为不小于1的正整数。

[0050] 作为一种优选的实施例，将初始的输入特征张量与权值张量进行由浮点数到定点数的量化，其方法为：

[0051] 使用最近邻法对初始的输入特征张量与权值张量进行由浮点数到定点数的量化。

[0052] 具体地，如图4所示，图4为本发明提供的对权值张量进行定点量化操作的示意图，设定量化位宽为2bit，则对应于00、01、10、11四种量化方式，在实际应用中，2的幂次定点量化均可用于此发明，通过最近邻法将其转化为-1、-0.5、0、0.5四种值，以原始数值为0.52、2.11、0.24、-0.25、-0.7、0.31的向量为例，通过最近邻法将浮点数量化为定点数，具体被量化为0.5、0.5、0、0、-0.5、0。若量化位宽为1bit，则对应于0、1两种量化方式，通过最近邻法将其转化为1、-1两种值，以原始数值为0.51、0.26、0.24、-0.24、-0.26、-0.66为例，则上述原始数值将被量化为1、1、1、-1、-1、-1。

[0053] 作为一种优选的实施例，得到权值定点表示张量，其方法包括：

[0054] 基于预先设定的量化位宽直接量化初始的权值张量得到权值定点表示张量，或者基于预先设定的量化位宽量化初始的权值张量，将得到的结果进行迭代训练后，得到权值定点表示张量。

[0055] 在实际应用中，对于权值张量与输入特征张量的量化位宽可以是不同的，也可以是相同的，这里不做限定。权值张量的量化位宽的范围为1-32bit，输入特征张量的量化位宽的范围为8-32bit，在实际应用中，将不小于8bit的视为高位位宽，将小于8bit的视为低位位宽，当权值张量的量化位宽为高位位宽时，无需再训练缩放系数及权值张量的定点数表示，直接将权值张量量化为定点数；当权值张量的量化位宽为低位位宽时，对缩放系数及权值张量的定点数表示进行再训练后，也能够将网络性能恢复到接近于32bit全精度网络的性能。本发明通过多次实验获得如下结论：

[0056] 当权值张量的量化位宽为8bit，输入特征张量的量化位宽为8bit时，无需再训练权值缩放系数，直接将权值张量进行量化，网络性能没有损失；

[0057] 当权值张量的量化位宽为8bit，输入特征张量的量化位宽为32bit时，无需在训练权值缩放系数，直接将权值张量进行量化，网络性能可以获得提升；

[0058] 当权值张量的量化位宽为3bit，输入特征张量的量化位宽为32bit时，对权值缩放系数进行再训练，网络性能可以获得提升。

[0059] 作为一种优选的实施例，所述将得到的结果进行迭代训练，其方法为：

[0060] 对权值缩放系数进行初始化；

[0061] 将第一权值张量依照权值缩放系数进行缩放，将缩放结果按照量化位宽进行量化，得到第二权值张量；

[0062] 求解权值缩放系数，直至权值缩放系数收敛于设定的稳定值，得到稳定的权值缩放系数；

[0063] 对第二权值张量与稳定的权值缩放系数进行乘积操作，得到权值定点表示张量。

[0064] 作为一种优选的实施例，求解权值缩放系数，其方法为：

[0065] 将第二权值张量与权值缩放系数代入公式计算对第一权值张量的最优表示，根据计算结果更新权值缩放系数，具体公式为：

[0066] $$\alpha\_gQ\_g \cong W\_g,$$

[0067] 其中，α_g表示权值缩放系数，Q_g表示第二权值张量，W_g表示第一权值张量。

[0068] 作为一种优选的实施例，第二权值张量与权值缩放系数一一对应。

[0069] 具体地，以第g个权值缩放系数与第g个第一权值张量W_g为例，将第g个权值缩放系数α_g进行初始化，将第g个第一权值张量W_g按照量化位宽L，参照权值缩放系数α_g进行缩放，采用浮点数到定点数量化的方法，定点量化为第二权值张量Q_g，将第g个第二权值张量与第g个权值缩放系数α_g代入$\alpha_gQ_g \cong W_g$，重复上述步骤直至g个权值缩放系数α_g收敛于稳定值，得到稳定的权值缩放系数，将g个第二权值张量与g个稳定的权值缩放系数α_g代入$\alpha_gQ_g \cong W_g$，求解对第一权值张量W_g的最优表示。在实际应用中，求解对第一权值张量W_g的最优表示可以视为对||W-alpha*Q||^{2}问题的解析求解过程，其中，Q为定点量化张量，alpha为缩放系数，通过迭代求解Q＝Quantize(W/alpha)，alpha＝(W*Q)/(Q*Q)，直至alpha的数值收敛，其中，Quantize表示浮点数的定点量化算法，可以但不限于所述最近邻量化算法。

[0070] 作为一种优选的实施例，在分别替换初始的权值张量以及初始的输入特征张量之后，在得到新的卷积神经网络之前，该方法还包括：

[0071] 对原始卷积神经网络中各卷积层，对权值定点表示张量与输入特征定点表示张量进行卷积运算，得到第一输出特征张量，直至原始卷积神经网络中所有卷积层被遍历；

[0072] 对原始卷积神经网络中各卷积层，对第一输出特征张量与稳定的权值缩放系数进行乘积运算，得到最终的输出特征张量，直至原始卷积神经网络中所有卷积层被遍历，最终的输出特征张量构成新的卷积神经网络。

[0073] 如图5所示，为本发明提供的获得新的卷积神经网络的过程示意图。其中，过程1表示对原始卷积神经网络进行卷积操作的过程，过程2表示获得的新的卷积神经网络进行卷积操作的过程；在图5中，每个指针表示由输入特征图通过卷积计算输出特征图的过程，指针尾部的方块对应于输入特征图的卷积区域，指针头部的方块对应于卷积计算出的值，指针头部方块所处的阴影立方体区域表示该卷积核最终输出的特征张量。

[0074] 本发明提供的方法应用前后的空间复杂度进行了分析，具体分析如下：以VGG16卷积神经网络为例，假设原始VGG16卷积神经网络占用的存储空间为50兆字节，量化位宽为2bit，则得到的新的VGG16卷积神经网络占用的存储空间为25兆字节，在此基础上，设定量化位宽为1bit，则得到的新的VGG16深度卷积神经网络占用的存储空间为18兆字节，通过多次实验表明，本发明的方法最高可以达到29倍的压缩率。原始卷积神经网络中的卷积层的

每一层权值及特征张量均以32bit浮点数表示，假定量化位宽为K bit，显然压缩率接近于32/K。可见，量化后的卷积层权值占用的空间远小于原始卷积神经网络卷积层，因此可以显著地减低卷积神经网络权值的存储开销。

[0075] 结合本文中所公开的实施例描述的方法或算法的步骤可以用硬件、处理器执行的软件模块，或者二者的结合来实施。软件模块可以置于随机存储器(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。

[0076] 本发明一种实施例的存储装置，其中存储有多条程序，所述程序适于由处理器加载并执行以实现上述所述的卷积神经网络量化与压缩的方法。

[0077] 本发明一种实施例的处理装置，包括处理器、存储设备；处理器，适于执行各条程序；存储设备，适于存储多条程序；所述程序适于由处理器加载并执行以实现上述所述的卷积神经网络量化与压缩的方法。

[0078] 所属技术领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的装置的具体工作过程及有关说明，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0079] 本领域技术人员应该能够意识到，结合本文中所公开的实施例描述的各示例的方法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明电子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示例的组成及步骤。这些功能究竟以电子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。

[0080] 术语“第一”、“第二”等是用于区别类似的对象，而不是用于描述或表示特定的顺序或先后次序。术语“包括”或者任何其它类似用语旨在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法不仅包括那些要素，而且还包括没有明确列出的其它要素，或者还包括这些过程、方法所固有的要素。

[0081] 至此，已经结合附图所示的优选实施方式描述了本发明的技术方案，但是，本领域技术人员容易理解的是，本发明的保护范围显然不局限于这些具体实施方式。在不偏离本发明的原理的前提下，本领域技术人员可以对相关技术特征作出等同的更改或替换，这些更改或替换之后的技术方案都将落入本发明的保护范围之内。

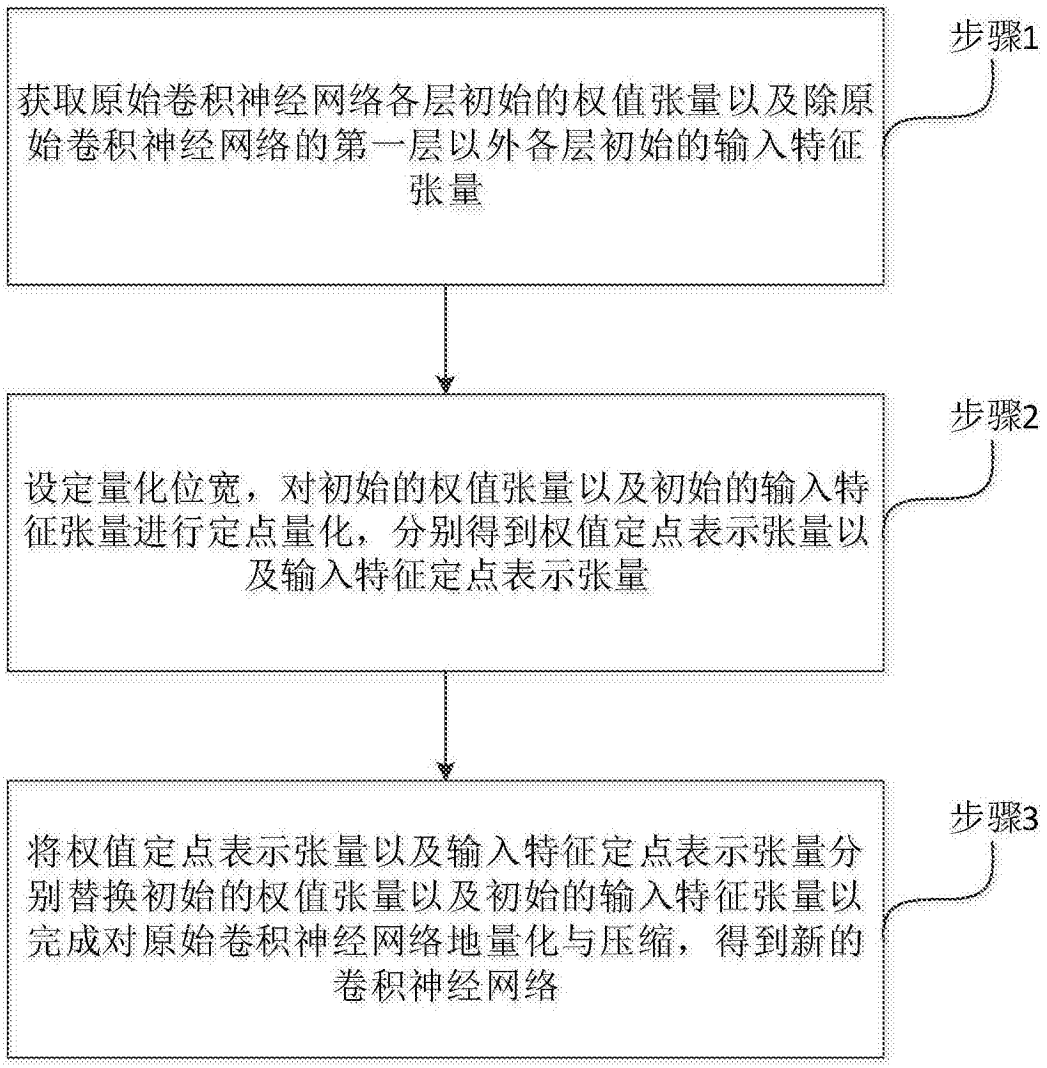

步骤1
获取原始卷积神经网络各层初始的权值张量以及除原始卷积神经网络的第一层以外各层初始的输入特征张量
步骤2
设定量化位宽，对初始的权值张量以及初始的输入特征张量进行定点量化，分别得到权值定点表示张量以及输入特征定点表示张量
步骤3
将权值定点表示张量以及输入特征定点表示张量分别替换初始的权值张量以及初始的输入特征张量以完成对原始卷积神经网络地量化与压缩，得到新的卷积神经网络

图1

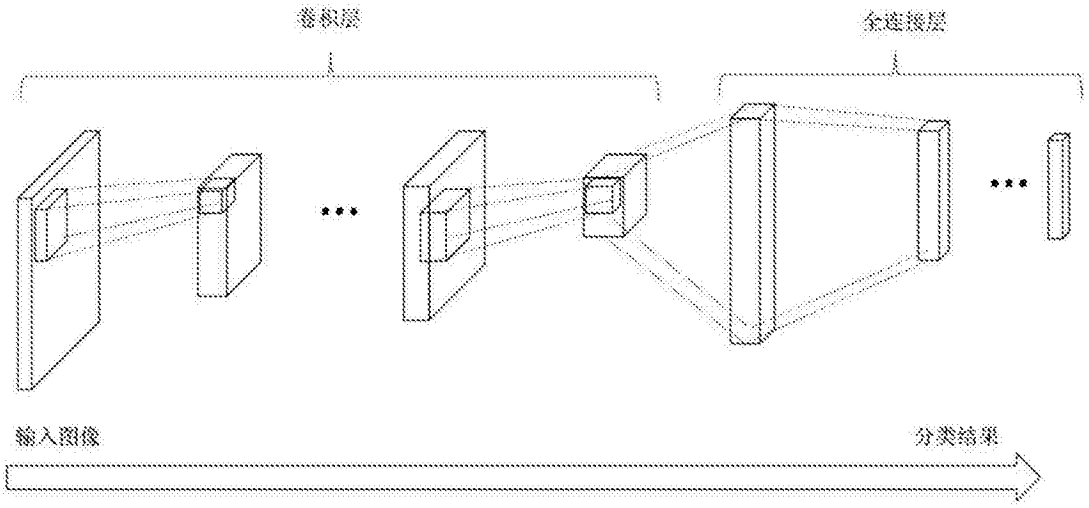

卷积层
全连接层
输入图像
分类结果

图2

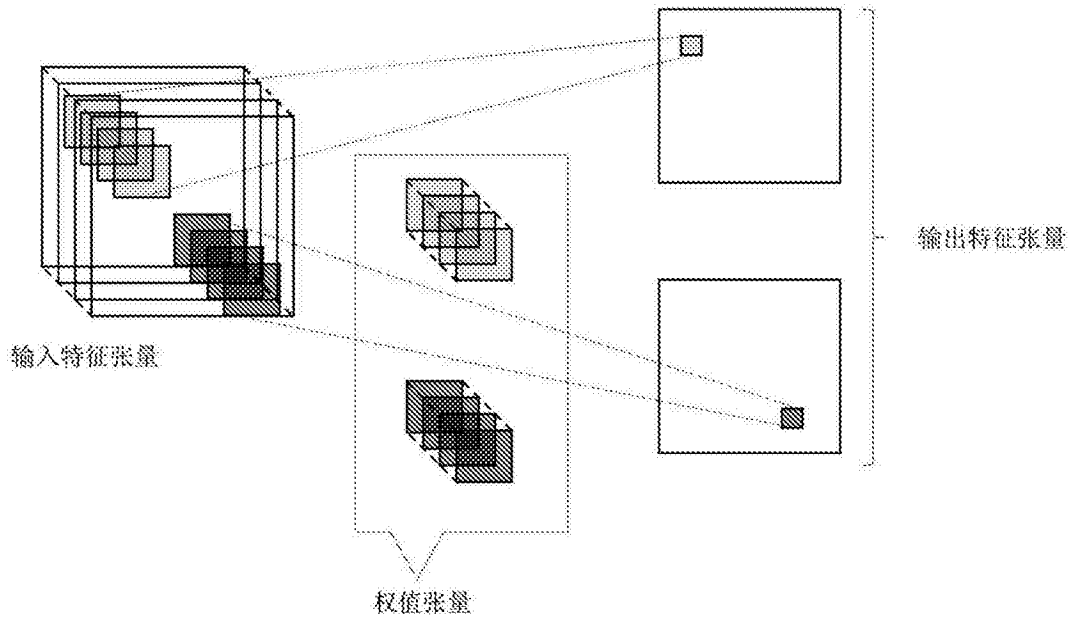

输入特征张量
输出特征张量
权值张量

图3

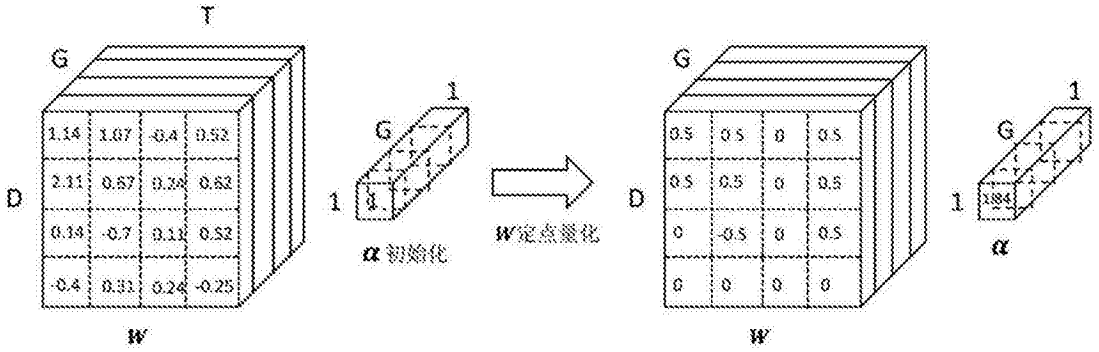

T
G
D
1
1.14
1.07
-0.4
0.52
2.11
0.57
0.36
0.62
0.14
-0.7
0.11
0.52
-0.4
0.31
0.24
-0.25
W
G
1
α
初始化
W定点量化
0.5
0.5
0
0.5
0.5
0.5
0
0.5
0
-0.5
0
0.5
0
0
0
0
0.884

图4

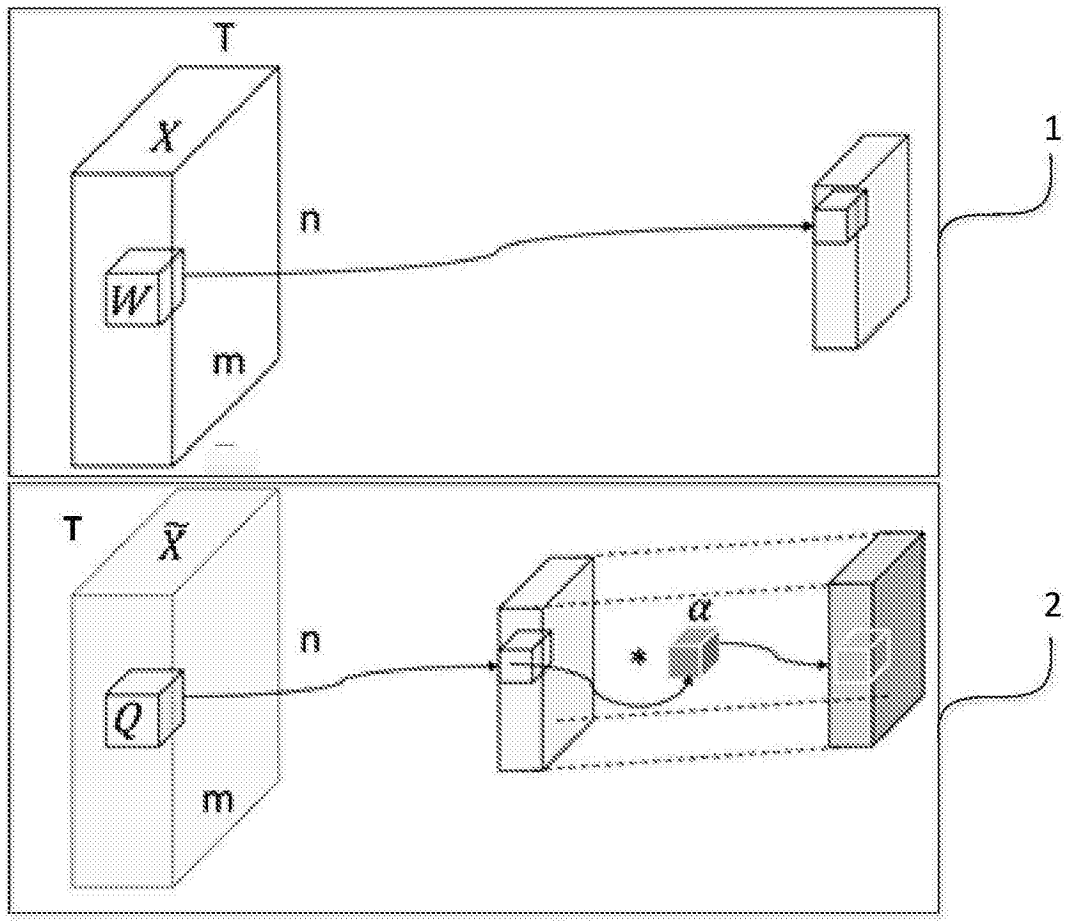


图5