(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117332830 A
(43)申请公布日 2024.01.02

(21)申请号 202310791759.5

(22)申请日 2023.06.29

(30)优先权数据
2209612.7 2022.06.30 GB
2209616.8 2022.06.30 GB
2216948.6 2022.11.14 GB
2216947.8 2022.11.14 GB

(71)申请人 想象技术有限公司
地址 英国赫特福德郡

(72)发明人 S•塞法尔瓦伊

(74)专利代理机构 北京三友知识产权代理有限公司 11127
专利代理师 王青芝 徐敏刚

(51)Int.Cl.
G06N 3/063(2023.01)
G06N 3/0464(2023.01)
G06N 3/084(2023.01)

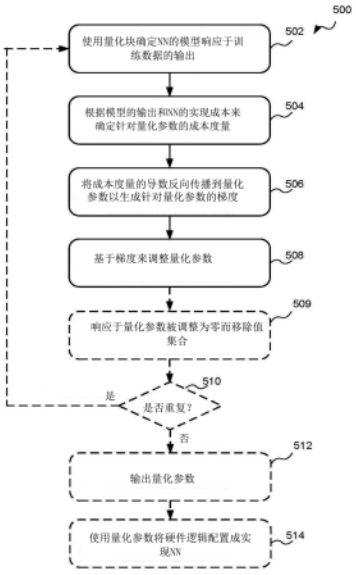
权利要求书4页 说明书45页 附图18页

(54)发明名称

标识用于量化待由神经网络处理的值的一个或多个量化参数

(57)摘要

本申请涉及标识用于量化待由神经网络处理的值的一个或多个量化参数。一种标识量化参数的计算机实现的方法，包括：(a)确定NN的模型的输出，模型包括量化块，每个量化块在模型根据NN的层处理值集合之前，将值集合变换为由量化参数定义的相应定点数格式；(b)确定NN的成本度量，成本度量是误差度量与实现度量的组合，实现度量表示基于量化参数的NN的实现成本，值集合已根据量化参数进行变换，对于NN的每个层，实现度量取决于：表示层的输出的实现成本的第一贡献；以及表示在层之前的层的输出的实现成本的第二贡献；(c)将成本度量的导数反向传播到至少一个量化参数，以生成针对至少一个量化参数的成本度量的梯度；以及(d)基于梯度来调整至少一个量化参数。



CN 117332830 A

1.一种标识一个或多个量化参数的计算机实现的方法,所述一个或多个量化参数用于将待由神经网络“NN”处理的值进行变换以在硬件中实现所述NN,所述方法包括在至少一个处理器中:

(a)确定所述NN的模型响应于训练数据的输出,所述NN的所述模型包括一个或多个量化块,所述一个或多个量化块中的每个量化块被配置成在所述模型根据所述NN的层处理一个或多个值集合之前,将输入到所述层的所述一个或多个值集合变换为由一个或多个量化参数定义的相应定点数格式;

(b)确定所述NN的成本度量,所述成本度量是误差度量与实现度量的组合,所述实现度量表示基于所述一个或多个量化参数的所述NN的实现成本,所述一个或多个值集合已根据所述一个或多个量化参数进行变换,对于所述NN的多个层中的每个层,所述实现度量取决于:

表示来自所述层的输出的实现成本的第一贡献;以及

表示来自在所述层之前的层的输出的实现成本的第二贡献;

(c)将所述成本度量的导数反向传播到所述一个或多个量化参数中的至少一个量化参数,以生成针对所述一个或多个量化参数中的所述至少一个量化参数的所述成本度量的梯度;以及

(d)基于针对所述一个或多个量化参数中的所述至少一个量化参数的所述梯度来调整所述一个或多个量化参数中的所述至少一个量化参数。

2.如权利要求1所述的计算机实现的方法,还包括:在调整步骤(d)之后,根据所述一个或多个量化参数中的经调整的至少一个量化参数而从所述NN的所述模型中移除值集合。

3.如权利要求1或2所述的计算机实现的方法,其中根据输入到所述层的权重数据的一个或多个输出通道的实现成本来形成所述第一贡献,并且根据输入到所述层的激活数据的一个或多个输入通道的实现成本来形成所述第二贡献。

4.如权利要求1或2所述的计算机实现的方法,其中根据输入到所述层的权重数据的一个或多个输出通道的实现成本来形成所述第一贡献,并且根据输入到所述层的权重数据的一个或多个输入通道的实现成本来形成所述第二贡献。

5.如权利要求1或2所述的计算机实现的方法,其中所述一个或多个量化参数中的每个量化参数包括相应位宽度,并且其中所述一个或多个值集合中的每个值集合是输入到所述层的值的通道,所述方法包括确定针对输入到所述层的权重数据的一个或多个输入通道中的每个输入通道的相应位宽度,以及确定针对输入到所述层的权重数据的一个或多个输出通道中的每个输出通道的相应位宽度。

6.如权利要求5所述的计算机实现的方法,其中对于输入到所述层的每个权重值,分别确定第一位宽度和第二位宽度,并且所述方法包括根据输入到所述层的每个权重值的相应第一位宽度和/或相应第二位宽度,可选地根据所述每个权重值的所述相应第一位宽度和所述相应第二位宽度中的较小者,对所述每个权重值进行变换。

7.如权利要求5所述的计算机实现的方法,所述方法包括:在所述调整步骤(d)之后,当针对输入到所述层的所述权重数据的对应输入通道的经调整的位宽度为零时,从所述NN的所述模型中移除输入到在前层的所述权重数据的输出通道。

8.如权利要求1或2所述的计算机实现的方法,其中:

根据输入到所述层的权重数据的一个或多个输出通道的实现成本以及输入到所述层的一个或多个偏置的实现成本来形成所述第一贡献;并且

根据输入到所述在前层的权重数据的一个或多个输出通道的实现成本以及输入到所述在前层的一个或多个偏置的实现成本来形成所述第二贡献。

9.如权利要求1或2所述的计算机实现的方法,其中根据输入到所述层的权重数据的一个或多个输出通道的实现成本来形成所述第一贡献,并且根据输入到所述在前层的权重数据的一个或多个输出通道的实现成本来形成所述第二贡献。

10.如权利要求1或2所述的计算机实现的方法,其中所述一个或多个量化参数中的每个量化参数包括相应位宽度,并且其中所述一个或多个值集合包括输入到所述层的权重数据的一个或多个输出通道以及输入到所述在前层的权重数据的一个或多个输出通道,所述方法包括根据相应位宽度对输入到所述层的权重数据的所述一个或多个输出通道中的每个输出通道进行变换,以及根据相应位宽度对输入到所述在前层的权重数据的所述一个或多个输出通道中的每个输出通道进行变换。

11.如权利要求10所述的计算机实现的方法,所述方法包括:在所述调整步骤(d)之后,当针对输入到所述在前层的所述权重数据的输出通道的所述经调整的位宽度为零时,从所述NN的所述模型中移除所述输出通道。

12.如权利要求9所述的计算机实现的方法,其中对于所述NN的多个层中的每个层,所述实现度量进一步取决于表示输入到所述在前层的一个或多个偏置的实现成本的另外的贡献。

13.如权利要求12所述的计算机实现的方法,所述方法包括:在所述调整步骤(d)之后,当针对输入到所述在前层的所述权重数据的输出通道的所述经调整的位宽度和所述输出通道的相关联偏置的绝对值为零时,从所述NN的所述模型中移除所述输出通道。

14.如权利要求1或2所述的计算机实现的方法,其中所述NN的层接收已从多于一个在前层的激活输出数据中导出的激活输入数据,并且其中针对所述层的所述实现度量取决于:

表示来自所述层的输出的实现成本的第一贡献;

表示来自在所述层之前的第一层的输出的实现成本的第二贡献;以及

表示来自在所述层之前的第二层的输出的实现成本的第三贡献。

15.如权利要求1或2所述的计算机实现的方法,其中所述NN的层输出被输入到第一后续层和第二后续层的激活数据,其中所述方法还包括在所述层与所述第一后续层之间向所述NN添加新的层,并且其中针对所述第一后续层的所述实现度量取决于:

表示来自所述第一后续层的输出的实现成本的第一贡献;以及

表示来自所述新的层的输出的实现成本的第二贡献。

16.如权利要求1或2所述的计算机实现的方法,其中所述第二贡献表示来自紧接在所述层之前的层的输出的实现成本。

17.如权利要求1或2所述的计算机实现的方法,还包括输出所述一个或多个量化参数中的所述经调整的所述至少一个量化参数以用于将硬件逻辑配置成实现所述NN。

18.如权利要求1或2所述的计算机实现的方法,还包括将硬件逻辑配置成使用所述经调整的量化参数来实现所述NN,可选地其中所述硬件逻辑包括神经网络加速器。

19.一种上面存储有计算机可读指令的非暂态计算机可读存储介质,当在计算机系统处被执行时,所述计算机可读指令使得所述计算机系统执行标识一个或多个量化参数的计算机实现的方法,所述一个或多个量化参数用于将待由神经网络"NN"处理的值进行变换以在硬件中实现所述NN,所述方法包括在至少一个处理器中:

(a)确定所述NN的模型响应于训练数据的输出,所述NN的所述模型包括一个或多个量化块,所述一个或多个量化块中的每个量化块被配置成在所述模型根据所述NN的层处理一个或多个值集合之前,将输入到所述层的所述一个或多个值集合变换为由一个或多个量化参数定义的相应定点数格式;

(b)确定所述NN的成本度量,所述成本度量是误差度量与实现度量的组合,所述实现度量表示基于所述一个或多个量化参数的所述NN的实现成本,所述一个或多个值集合已根据所述一个或多个量化参数进行变换,对于所述NN的多个层中的每个层,所述实现度量取决于:

表示来自所述层的输出的实现成本的第一贡献;以及

表示来自在所述层之前的层的输出的实现成本的第二贡献;

(c)将所述成本度量的导数反向传播到所述一个或多个量化参数中的至少一个量化参数,以生成针对所述一个或多个量化参数中的所述至少一个量化参数的所述成本度量的梯度;以及

(d)基于针对所述一个或多个量化参数中的所述至少一个量化参数的所述梯度来调整所述一个或多个量化参数中的所述至少一个量化参数。

20.一种基于计算的设备,所述基于计算的设备被配置成标识一个或多个量化参数,所述一个或多个量化参数用于将待由神经网络"NN"处理的值进行变换以在硬件中实现所述NN,所述基于计算的设备包括:

至少一个处理器;以及

存储器,所述存储器耦接到所述至少一个处理器,所述存储器包括:

计算机可读代码,当由所述至少一个处理器执行时,所述计算机可读代码使得所述至少一个处理器:

(a)确定所述NN的模型响应于训练数据的输出,所述NN的所述模型包括一个或多个量化块,所述一个或多个量化块中的每个量化块被配置成在所述模型根据所述NN的层处理一个或多个值集合之前,将输入到所述层的所述一个或多个值集合变换为由一个或多个量化参数定义的相应定点数格式;

(b)确定所述NN的成本度量,所述成本度量是误差度量与实现度量的组合,所述实现度量表示基于所述一个或多个量化参数的所述NN的实现成本,所述一个或多个值集合已根据所述一个或多个量化参数进行变换,对于所述NN的多个层中的每个层,所述实现度量取决于:

表示来自所述层的输出的实现成本的第一贡献;以及

表示来自在所述层之前的层的输出的实现成本的第二贡献;

(c)将所述成本度量的导数反向传播到所述一个或多个量化参数中的至少一个量化参数,以生成针对所述一个或多个量化参数中的所述至少一个量化参数的所述成本度量的梯度;以及

(d)基于针对所述一个或多个量化参数中的所述至少一个量化参数的所述梯度来调整所述一个或多个量化参数中的所述至少一个量化参数。

标识用于量化待由神经网络处理的值的一个或多个量化参数

[0001] 相关申请的相交引用

[0002] 本申请要求2022年6月30日提交的英国专利申请2209612.7的优先权，该英国专利申请以全文引用方式并入本文中。本申请还要求2022年11月14日提交的英国专利申请2216948.6的优先权，该英国专利申请以全文引用方式并入本文中。本申请还要求于2022年6月30日提交的英国专利申请2209616.8的优先权，该申请以全文引用方式并入本文。本申请还要求于2022年11月14日提交的英国专利申请2216947.8的优先权，该申请以全文引用方式并入本文。

技术领域

[0003] 本申请涉及标识用于量化待由神经网络处理的值的一个或多个量化参数。

背景技术

[0004] 神经网络(NN)是人工网络的一种形式，包括可用于机器学习应用程序的多个互连层(例如“层”)。具体地，NN可用于信号处理应用程序中，该信号处理应用程序包括但不限于图像处理应用程序和计算机视觉应用程序。图1示出了包括多个层102-1、102-2、102-3的示例NN 100。每个层102-1、102-2、102-3接收输入激活数据，根据层来处理输入激活数据以产生输出数据。输出数据作为输入激活数据被提供给另一层，或者作为NN的最终输出数据被输出。例如，在图1的NN 100中，第一层102-1接收针对NN 100的原始输入激活数据104，并且根据第一层102-1来处理输入激活数据以产生输出数据。第一层102-1的输出数据成为针对第二层102-2的输入激活数据，该第二层根据第二层102-2来处理输入激活数据以产生输出数据。第二层102-2的输出数据成为针对第三层102-3的输入激活数据，该第三层根据第三层102-3来处理输入激活数据以产生输出数据。第三层102-3的输出数据作为NN的输出数据106被输出。

[0005] 对输入到层的激活数据执行的处理取决于层的类型。例如，NN的每个层可以是多种不同类型中的一种类型。示例NN层类型包括但不限于：卷积层、激活层、归一化层、池化层和全连接层。对于本领域技术人员将显而易见的是，这些是示例NN层类型，并且这不是详尽的列表，并且可存在其他NN层类型。

[0006] 在卷积层中，输入到层的激活数据与输入到该层的权重数据进行卷积。将激活数据与权重数据进行卷积的输出可以可选地与输入到卷积层的一个或多个偏移偏置进行组合。

[0007] 图2A示出了用于NN的卷积层中的数据格式的示例概况。输入到卷积层的激活数据包括多个数据值。参考图2A，输入到卷积层的激活数据可以具有维度$B\times C_{in}\times H_a\times W_a$。换句话说，激活数据可以被布置为$C_{in}$个输入通道(例如有时称为“数据通道”)，其中每个输入通道具有空间维度$H_a\times W_a$——其中H_a和W_a分别是高度维度和宽度维度。在图2A中，激活数据被示出为包括四个输入通道(即$C_{in}=4$)。每个输入通道是输入数据值集合。输入到卷积层的激活数据也可以由批次大小B来定义。批次大小B未在图2A中示出，但定义了输入到卷积层的

数据批次的数量。例如，在图像分类应用中，批次大小可以指输入到卷积层的数据中的单独图像的数量。

[0008] 输入到卷积层的权重数据包括多个权重值，该多个权重值也可以称为滤波器权重、系数或权重。权重数据被布置在一个或多个输入通道和一个或多个输出通道中。输出通道可以替代地称为内核或滤波器。再次参考图2A，权重数据可以具有维度$C_{out}\times C_{in}\times H_w\times W_w$。通常，权重数据中的输入通道的数量对应于(例如等于)激活数据中的输入通道的数量，该权重数据将与该激活数据进行组合(例如在图2A中所示出的示例中，$C_{in}=4$)。输入到卷积层的权重数据的每个滤波器的每个输入通道具有空间维度$H_w\times W_w$——其中H_w和W_w分别是高度维度和宽度维度。每个输入通道是权重值集合。每个输出通道是权重值集合。每个权重值被包括在一个输入通道和一个输出通道中(例如被一个输入通道和一个输出通道包括，或者是一个输入通道和一个输出通道的一部分)。C_{out}维度(例如输出通道的数量)未在图2A中示出——但表示了通过将权重数据与激活数据进行组合而生成的输出数据中的通道的数量。如图2A中所示出，在卷积层中，权重数据可以根据在方向s和t上跨多个步骤的卷积运算而与激活输入数据进行组合。

[0009] 图2B示意性地示出了示例卷积层202，该示例卷积层被布置成将输入激活数据206与输入权重数据208进行组合。图2B还示出了在层202内使用可选的偏移偏置212。在图2B中，输入到层202的激活数据206被布置在三个输入通道1、2、3中。权重数据208中的输入通道的数量对应于(例如等于)激活数据206中的输入通道的数量，该权重数据208将与该激活数据进行组合。因此，权重数据208被布置在三个输入通道1、2、3中。权重数据208还被布置在四个输出通道(例如滤波器)A、B、C、D中。权重数据208中的输出通道的数量对应于(例如等于)输出数据210中的通道(例如数据通道)的数量。每个权重值被包括在一个输入通道和一个输出通道中(例如被一个输入通道和一个输出通道包括，或者是一个输入通道和一个输出通道的一部分)。例如，权重值216被包括在输入通道1和输出通道A中。输入激活数据206与输入权重数据208进行卷积，以便生成具有四个数据通道A、B、C、D的输出数据210。权重数据208中的每个滤波器的第一输入通道与激活数据206的第一输入通道进行卷积，权重数据208中的每个滤波器的第二输入通道与激活数据206的第二输入通道进行卷积，并且权重数据208中的每个滤波器的第三输入通道与激活数据206的第三输入通道进行卷积。可以对与激活数据的每个输入通道的每个滤波器的所述卷积的结果进行求和(例如累积)，以便形成输出数据210的每个数据通道的输出数据值。如果卷积层202没有被配置成使用偏移偏置，则输出数据210将是该卷积层的输出。在图2B中，输出数据210是待与偏移偏置212进行组合的中间输出数据。输入到层202的权重数据208的四个输出通道A、B、C、D中的每一者与相应偏置A、B、C、D相关联。在卷积层中，偏置A、B、C、D与中间数据210的相应数据通道A、B、C、D进行求和，以便生成具有四个数据通道A、B、C、D的输出数据214。

[0010] 通常但不一定在卷积层之后的激活层对输入到该层的激活数据执行一个或多个激活函数。激活函数取单个数字，并且对其执行特定非线性数学运算。在一些示例中，激活层可通过实现ReLU函数(即，$f(x)=\max(0,x)$)充当整流线性单元(ReLU)，或者可通过实现PReLU函数充当参数化整流线性单元(PReLU)。

[0011] 归一化层被配置成对输入到该层的激活数据执行归一化函数，诸如局部响应归一化(LRN)函数。通常但不一定插入在连续卷积层之间的池化层执行池化函数(诸如最大值或

均值函数),以将输入到该层的激活数据的子集汇总。因此,池化层的目的是减小表示的空间大小,以减少网络中参数和计算的数量,并且因此还控制过度拟合。

[0012] 通常但不一定在多个卷积层和池化层之后的全连接层取三维输入激活数据集合,并且输出N维向量。在NN用于分类的情况下,N可以是类别的数量,并且向量中的每个值可以表示某个类别的概率。N维向量通过与权重数据的矩阵乘法来生成,可选地后面是偏置偏移。因此,全连接层接收激活数据、权重数据和可选的偏移偏置。如本领域技术人员所知,以与本文中针对卷积层描述的方式等效的方式,输入到全连接层的激活数据可以被布置在一个或多个输入通道中,并且输入到全连接层的权重数据可以被布置在一个或多个输入通道和一个或多个输出通道中,其中那些输出通道中的每个输出通道可选地与相应偏移偏置相关联。

[0013] 因此,如图3中所示出,NN的每个层302接收输入激活数据,并且生成输出数据;并且一些层(诸如卷积层和全连接层)还接收权重数据和/或偏置。

[0014] 用于实现NN的硬件(例如NN加速器)包括硬件逻辑,该硬件逻辑可以被配置成根据NN的层来处理针对NN的输入数据。具体地,用于实现NN的硬件包括硬件逻辑,该硬件逻辑可以被配置成根据该层来处理输入到每个层的激活数据,并且生成针对该层的输出数据,该输出数据成为针对另一层的输入激活数据,或者成为NN的输出。例如,如果NN包括后面是激活层的卷积层,则可被配置成实现NN的硬件逻辑包括可被配置成使用输入到卷积层的权重数据和可选的偏置对输入到NN的激活数据执行卷积以产生针对卷积层的输出数据的硬件逻辑,以及可被配置成将激活函数应用于输入到激活层的激活数据(即卷积层的输出数据)以生成针对NN的输出数据的硬件逻辑。

[0015] 如本领域技术人员所知,对于处理值集合的硬件,每个值以数字格式表示。两种最合适的数字格式是定点数格式和浮点数格式。如本领域的技术人员所知,定点数格式在基数点(例如,小数点或二进制点)之后具有固定数量的数位。相反,浮点数格式没有固定的基数点(即,可以“浮动”)。换句话说,基数点可被放置在表示内的多个位置。虽然以浮点数格式表示输入到NN的层以及从NN的层输出的值可以允许产生更准确或更精确的输出数据,但在硬件中以浮点数格式处理值较复杂,与以定点数格式处理值的硬件相比,这往往会增加硅面积、功耗和硬件的复杂度。因此,用于实现NN的硬件可以被配置成以定点数格式表示输入到NN的层的值,以减小硬件逻辑的面积、功耗和存储器带宽。

[0016] 一般来说,可用于表示输入到NN的层以及从NN的层输出的值的位数越少,NN在硬件中实现得就越高效。然而,通常,用于表示输入到NN的层以及从NN的层输出的值的位越少,NN就变得越不准确。因此,期望标识用于表示NN的值的定点数格式,该定点数格式将用于表示NN的值的位数与NN的准确度平衡。

[0017] 下文描述的实施方案仅借助于示例来提供,并且不对解决用于标识用于表示NN的值的定点数格式的方法和系统的任何或所有缺点的实现进行限制。

发明内容

[0018] 提供本发明内容是为了介绍在以下详细描述中进一步描述的一些概念。本发明内容不旨在标识所要求保护的主题的关键特征或必要特征,也不旨在用于限制所要求保护的主题的范围。

[0019] 根据本发明的第一方面，提供了一种标识一个或多个量化参数的计算机实现的方法，所述一个或多个量化参数用于将待由神经网络“NN”处理的值进行变换以在硬件中实现NN，所述方法包括在至少一个处理器中：(a)确定NN的模型响应于训练数据的输出，所述NN的模型包括一个或多个量化块，所述一个或多个量化块中的每个量化块被配置成在模型根据NN的层处理一个或多个值集合之前，将输入到所述层的一个或多个值集合变换为由一个或多个量化参数定义的相应定点数格式；(b)确定NN的成本度量，所述成本度量是误差度量与实现度量的组合，所述实现度量表示基于一个或多个量化参数的NN的实现成本，一个或多个值集合已根据所述一个或多个量化参数进行变换，对于NN的多个层中的每个层，所述实现度量取决于：表示来自所述层的输出的实现成本的第一贡献；以及表示来自在所述层之前的层的输出的实现成本的第二贡献；(c)将成本度量的导数反向传播到一个或多个量化参数中的至少一个量化参数，以生成针对一个或多个量化参数中的至少一个量化参数的成本度量的梯度；以及(d)基于针对一个或多个量化参数中的至少一个量化参数的梯度来调整一个或多个量化参数中的至少一个量化参数。

[0020] 一个或多个量化参数中的每个量化参数可以包括相应位宽度，并且所述方法还可以包括：在调整步骤(d)之后，当针对值集合或对应值集合的经调整的位宽度为零时，从NN的模型中移除所述值集合。

[0021] 可以根据输入到层的权重数据的一个或多个输出通道的实现成本来形成第一贡献，并且可以根据输入到层的激活数据的一个或多个输入通道的实现成本来形成第二贡献。

[0022] 一个或多个量化参数中的每个量化参数可以包括相应位宽度，一个或多个值集合中的每个值集合可以是输入到层的值的通道，并且所述方法可以包括根据相应位宽度对输入到层的激活数据的一个或多个输入通道中的每个输入通道进行变换，以及根据相应位宽度对输入到层的权重数据的一个或多个输出通道中的每个输出通道进行变换。

[0023] 所述方法可以包括：在调整步骤(d)之后，当针对输入到层的激活数据的对应输入通道的经调整的位宽度为零时，从NN的模型中移除输入到在前层的权重数据的输出通道。

[0024] 可以根据输入到层的权重数据的一个或多个输出通道的实现成本来形成第一贡献，并且可以根据输入到层的权重数据的一个或多个输入通道的实现成本来形成第二贡献。

[0025] 一个或多个量化参数中的每个量化参数可以包括相应位宽度，一个或多个值集合中的每个值集合可以是输入到层的值的通道，并且所述方法可以包括确定针对输入到层的权重数据的一个或多个输入通道中的每个输入通道的相应位宽度，以及确定针对输入到层的权重数据的一个或多个输出通道中的每个输出通道的相应位宽度。

[0026] 对于输入到层的每个权重值，可以分别确定第一位宽度和第二位宽度，并且所述方法可以包括根据输入到层的每个权重值的相应第一位宽度和/或相应第二位宽度，可选地根据所述每个权重值的相应第一位宽度和相应第二位宽度中的较小者，对所述每个权重值进行变换。

[0027] 所述方法可以包括：在调整步骤(d)之后，当针对输入到层的权重数据的对应输入通道的经调整的位宽度为零时，从NN的模型中移除输入到在前层的权重数据的输出通道。

[0028] 可以根据输入到层的权重数据的一个或多个输出通道的实现成本以及输入到层

的一个或多个偏置的实现成本来形成第一贡献，并且可以根据输入到在前层的权重数据的一个或多个输出通道的实现成本以及输入到在前层的一个或多个偏置的实现成本来形成第二贡献。

[0029] 一个或多个量化参数中的每个量化参数可以包括相应位宽度，一个或多个值集合可以包括输入到层的权重数据的一个或多个输出通道和相关联偏置，以及输入到在前层的权重数据的一个或多个输出通道和相关联偏置，并且所述方法可以包括根据相应位宽度对输入到层的权重数据的一个或多个输出通道中的每个输出通道进行变换，根据相应位宽度对输入到层的一个或多个偏置中的每个偏置进行变换，根据相应位宽度对输入到在前层的权重数据的一个或多个输出通道中的每个输出通道进行变换，以及根据相应位宽度对输入到在前层的一个或多个偏置中的每个偏置进行变换。

[0030] 可以使用相同位宽度来对权重数据的输出通道和所述输出通道的相关联偏置进行变换。

[0031] 所述方法可以包括：在调整步骤(d)之后，当针对输入到层的权重数据的输出通道的经调整的位宽度和所述输出通道的相关联偏置为零时，从NN的模型中移除所述输出通道。

[0032] 可以根据输入到层的权重数据的一个或多个输出通道的实现成本来形成第一贡献，并且可以根据输入到在前层的权重数据的一个或多个输出通道的实现成本来形成第二贡献。

[0033] 一个或多个量化参数中的每个量化参数可以包括相应位宽度，一个或多个值集合可以包括输入到层的权重数据的一个或多个输出通道以及输入到在前层的权重数据的一个或多个输出通道，并且所述方法可以包括根据相应位宽度对输入到层的权重数据的一个或多个输出通道中的每个输出通道进行变换，以及根据相应位宽度对输入到在前层的权重数据的一个或多个输出通道中的每个输出通道进行变换。

[0034] 所述方法可以包括：在调整步骤(d)之后，当针对输入到在前层的权重数据的输出通道的经调整的位宽度为零时，从NN的模型中移除所述输出通道。

[0035] 对于NN的多个层中的每个层，实现度量可以进一步取决于表示输入到在前层的一个或多个偏置的实现成本的另外的贡献。

[0036] 所述方法可以包括：在调整步骤(d)之后，当针对输入到在前层的权重数据的输出通道的经调整的位宽度和所述输出通道的相关联偏置的绝对值为零时，从NN的模型中移除所述输出通道。

[0037] NN的层可以接收已从多于一个在前层的激活输出数据中导出的激活输入数据，并且针对所述层的实现度量可以取决于：表示来自所述层的输出的实现成本的第一贡献；表示来自在所述层之前的第一层的输出的实现成本的第二贡献；以及表示来自在所述层之前的第二层的输出的实现成本的第三贡献。

[0038] NN的层可以输出被输入到第一后续层和第二后续层的激活数据，所述方法还可以包括在所述层与第一后续层之间向NN添加新的层，并且第一后续层的实现度量可以取决于：表示来自第一后续层的输出的实现成本的第一贡献；以及表示来自新的层的输出的实现成本的第二贡献。

[0039] 新的层可以不对层的输出激活数据执行任何计算。

[0040] 第二贡献可以表示来自紧接在所述层之前的层的输出的实现成本。

[0041] 所述方法可以包括使用一个或多个量化参数中的经调整的至少一个量化参数来重复(a)、(b)、(c)和(d)。

[0042] 所述方法还可以包括输出一个或多个量化参数中的经调整的至少一个量化参数以用于将硬件逻辑配置成实现NN。

[0043] 所述方法还可以包括将硬件逻辑配置成使用经调整的量化参数来实现NN。

[0044] 硬件逻辑可以包括神经网络加速器。

[0045] 根据本发明的第二方面,提供了一种基于计算的设备,所述基于计算的设备被配置成标识一个或多个量化参数,所述一个或多个量化参数用于将待由神经网络"NN"处理的值进行变换以在硬件中实现NN,所述基于计算的设备包括:至少一个处理器;以及存储器,所述存储器耦接到所述至少一个处理器,所述存储器包括:计算机可读代码,当由所述至少一个处理器执行时,所述计算机可读代码使得所述至少一个处理器:(a)确定NN的模型响应于训练数据的输出,所述NN的模型包括一个或多个量化块,所述一个或多个量化块中的每个量化块被配置成在模型根据NN的层处理一个或多个值集合之前,将输入到所述层的一个或多个值集合变换为由一个或多个量化参数定义的相应定点数格式;(b)确定NN的成本度量,所述成本度量是误差度量与实现度量的组合,所述实现度量表示基于一个或多个量化参数的NN的实现成本,一个或多个值集合已根据所述一个或多个量化参数进行变换,对于NN的多个层中的每个层,所述实现度量取决于:表示来自所述层的输出的实现成本的第一贡献;以及表示来自在所述层之前的层的输出的实现成本的第二贡献;(c)将成本度量的导数反向传播到一个或多个量化参数中的至少一个量化参数,以生成针对一个或多个量化参数中的至少一个量化参数的成本度量的梯度;以及(d)基于针对一个或多个量化参数中的至少一个量化参数的梯度来调整一个或多个量化参数中的至少一个量化参数。

[0046] 根据本发明的第三方面,提供了一种使用在硬件中实现的神经网络"NN"来处理数据的计算机实现的方法,所述NN包括多个层,每个层被配置成对输入到层的激活数据进行运算,以便形成针对层的输出数据,所述数据被布置在数据通道中,所述方法包括:对于针对层的输出数据的标识通道,对输入到层的激活数据进行运算,使得针对层的输出数据不包括标识通道;以及在被配置成对针对层的输出数据进行运算的NN的运算之前,根据指示本应包括有标识通道的针对层的输出数据的结构的信息,将替换通道插入到针对层的输出数据中以代替标识通道。

[0047] 根据本发明的第四方面,提供了一种基于计算的设备,所述基于计算的设备被配置成使用在硬件中实现的神经网络"NN"来处理数据,所述NN包括多个层,每个层被配置成对输入到层的激活数据进行运算,以便形成针对层的输出数据,所述数据被布置在数据通道中,所述基于计算的设备包括至少一个处理器,所述至少一个处理器被配置成:对于针对层的输出数据的标识通道,对输入到层的激活数据进行运算,使得针对层的输出数据不包括标识通道;以及在被配置成对针对层的输出数据进行运算的NN的运算之前,根据指示本应包括有标识通道的针对层的输出数据的结构的信息,将替换通道插入到针对层的输出数据中以代替标识通道。

[0048] 可配置成实现NN(例如NN加速器)的硬件逻辑可以在集成电路上的硬件中体现。可以提供一种在集成电路制造系统处制造可配置成实现NN(例如NN加速器)的硬件逻辑的方

法。可以提供一种集成电路定义数据集，当在集成电路制造系统中被处理时，所述集成电路定义数据集将系统配置成制造可配置成实现NN(例如NN加速器)的硬件逻辑。可以提供一种非暂态计算机可读存储介质，所述非暂态计算机可读存储介质上存储有可配置成实现NN(例如NN加速器)的硬件逻辑的计算机可读描述，当在集成电路制造系统中被处理时，所述计算机可读描述使得集成电路制造系统制造体现可配置成实现NN(例如NN加速器)的硬件逻辑的集成电路。

[0049] 可以提供一种集成电路制造系统，所述集成电路制造系统包括：非暂态计算机可读存储介质，所述非暂态计算机可读存储介质上存储有可配置成实现NN(例如NN加速器)的硬件逻辑的计算机可读描述；布局处理系统，所述布局处理系统被配置成处理计算机可读描述，以便生成体现可配置成实现NN(例如NN加速器)的硬件逻辑的集成电路的电路布局描述；以及集成电路生成系统，所述集成电路生成系统被配置成制造可配置成根据电路布局描述来实现NN(例如NN加速器)的硬件逻辑。

[0050] 可以提供用于执行如本文中描述的方法的计算机程序代码。可以提供上面存储有计算机可读指令的非暂态计算机可读存储介质，当在计算机系统处执行时，所述计算机可读指令使计算机系统执行如本文所述的方法。

[0051] 如对本领域的技术人员将显而易见，上述特征可以适当地组合，并且可与本文中所描述的示例的各方面中的任一方面进行组合。

附图说明

[0052] 现在将参考附图详细描述示例，在附图中：

[0053] 图1是示例神经网络(NN)的示意图；

[0054] 图2A示出了用于NN的卷积层中的数据格式的示例概况；

[0055] 图2B示意性地示出了示例卷积层。

[0056] 图3是示出输入到NN的层以及从NN的层输出的数据的示意图；

[0057] 图4是示出具有和不具有量化块的NN的示例模型的示意图；

[0058] 图5是用于标识针对NN的量化参数的示例方法的流程图；

[0059] 图6是示出用于生成误差度量的第一示例方法的示意图；

[0060] 图7是示出用于生成误差度量的第二示例方法的示意图；

[0061] 图8是示出示例成本度量相对于位宽度的示例梯度的曲线图；

[0062] 图9是示出NN的两个相邻层之间的交互的示意图；

[0063] 图10A是示出包括剩余层的NN的示意图。

[0064] 图10B是用于插入替换通道的示例方法的流程图。

[0065] 图10C至图10E是示出包括剩余层的NN的示意图。

[0066] 图11是用于标识NN的量化参数和权重的示例方法的流程图；

[0067] 图12是示出量化为示例定点数格式的示意图；

[0068] 图13是示例NN加速器的方块图；

[0069] 图14是示例基于计算的设备的方块图；

[0070] 图15是其中可实现NN加速器的示例计算机系统的方块图；并且

[0071] 图16是用于生成体现如本文中所描述的NN加速器的集成电路的示例集成电路制

造系统的方块图。

[0072] 附图示出了各种示例。技术人员将了解,附图中所示出的元件边界(例如框、框的组,或其他形状)表示边界的一个示例。在一些示例中,情况可能是一个元件可以被设计为多个元件,或者多个元件可以被设计为一个元件。在适当的情况下,贯穿各附图使用共同的附图标记来指示类似的特征。

具体实施方式

[0073] 借助于示例呈现以下描述,以使得本领域的技术人员能够制造和使用本发明。本发明不限于本文所描述的实施方案,并且对所公开的实施方案的各种修改对于本领域技术人员而言将是显而易见的。仅借助于示例来描述实施方案。

[0074] 由于高效表示值集合的位数是基于该集合中的值范围,因而通过将输入到NN的值划分成集合并且在每集合基础上选择定点数格式,可以高效实现NN,而不会显著降低其准确度。由于输入到相同层的值往往是相关的,因而每个集合可以是针对层的特定类型输入的全部或一部分。例如,每个集合可以是层的输入激活数据值的全部或一部分;层的输入权重数据的全部或一部分;或者层的偏置的全部或一部分。这些集合是包括针对层的特定类型输入的全部还是仅一部分取决于用于实现NN的硬件。例如,用于实现NN的一些硬件可以仅支持每层每种输入类型的单个定点数格式,而用于实现NN的其他硬件可以支持每层每种输入类型的多个定点数格式。

[0075] 每个定点数格式由一个或多个量化参数定义。常见的定点数格式是Q格式,其指定预定数量的整数位a和小数位b。因此,数字可以表示为Qa.b,这总共需要a+b+1位(包括符号位)。下表1中示出了示例Q格式。

[0076] 表1

[0077]

Q格式	描述	示例
Q4.4	4个整数位和4个小数位	0110.1110_2
Q0.8	0整数位和8个小数位	$.01101110_2$

[0078] 在Q格式用于表示NN的值的情况下,对于每个定点数格式,量化参数可以包括整数位数a和小数位数b。

[0079] 在其他情况下,可以使用由固定整数指数exp和b位尾数m定义的定点数格式,而不是使用Q格式来表示输入到NN的层的值,使得值z等于$z=2^{exp}m$。在一些情况下,尾数m可以二进制补码格式表示。然而,在其他情况下,可以使用其他经签名或未签名的整数格式。在这些情况下,对于以该格式表示的值集合,指数exp和尾数位数b仅需要被存储一次。在这种定点数格式用于表示NN的值的情况下,对于每个定点数格式,量化参数可以包括尾数位长度b(其在本文中也可以称为位宽度或位长度)和指数exp。

[0080] 在又其他情况下,8位非对称定点(Q8A)格式可用于表示输入到NN的层的值。此格式包括最小可表示数字r_{min}、最大可表示数字r_{max}、零点z以及针对每个值的8位数字,该8位数字标识最小数与最大数之间的线性内插因子。在其他情况下,可以使用Q8A格式的变型,其中用于存储内插因子的位数是可变的(例如用于存储内插因子的位数可以是多个可能的整数中的一个整数)。浮点值d_{float}可以从等式(1)中所示的这种格式中进行构造,其中b是由量化表示使用的位数,并且z是量化零点,该量化零点将总是精确地映射回0.f。在这种定点

数格式用于表示NN的值的情况下，对于每个定点数格式，量化参数可以包括最大可表示数字或值r_{max}、最小可表示数字或值r_{min}、量化零点z以及可选地，尾数位长度b(即当位长度不固定为8时)。

[0081] $$d_{float} = \frac{(r_{max} - r_{min})(d_{Q8A} - z)}{2^b - 1} \quad (1)$$

[0082] 虽然用于高效表示值集合的定点数格式(并且更具体地，其量化参数)可以仅根据该集合中的值范围来确定，但由于NN的层是互连的，因而当选择用于表示NN的值的定点数格式(并且更具体地，其量化参数)时，通过考虑层之间的交互，可以在用于表示NN的值的位数与NN的性能(例如准确度)之间实现更好的权衡。

[0083] 因此，本文中描述了用于标识定点数格式，并且特别是标识其量化参数(例如指数和尾数位长度)的方法和系统，该定点数格式用于使用反向传播来表示NN的值。如本领域技术人员所知，反向传播是一种可用于训练NN的技术。训练NN包括标识适当的权重以将NN配置成执行特定功能。

[0084] 具体地，为了经由反向传播来训练NN，NN的模型被配置成使用特定权重集合，随后将训练数据应用于该模型，并且记录模型响应于训练数据的输出。随后从记录输出中计算可微分误差度量，该可微分误差度量定量地指示使用该特定权重集合的NN的性能。在一些情况下，误差度量可以是针对该训练数据的记录输出与预期输出之间的距离(例如均方距离)。然而，这只是示例，可以使用任何合适的误差度量。误差度量的导数随后被反向传播到NN的权重，以产生误差度量相对于每个权重的梯度/导数。随后基于梯度来调整权重以便减小误差度量。此过程可以重复，直到误差度量收敛为止。

[0085] 通常使用NN的模型来训练NN，其中NN的值(例如激活数据、权重数据和偏置)以浮点数格式来表示和处理。使用浮点数格式来表示和处理NN的值的NN在本文中被称为浮点NN。浮点NN的模型在本文中可以称为NN的浮点模型。然而，如上文所描述，用于实现NN的硬件(例如NN加速器)可以使用定点数格式来表示NN的值(例如激活数据、权重数据和偏置)，以减小硬件的大小以及提高硬件的效率。对其值中的至少一些值使用定点数格式的NN在本文中称为定点NN。为了训练定点NN，可以向NN的浮点模型添加量化块，在处理这些值之前，这些量化块将NN的值量化(或将这些值的量化模拟)为预定定点数格式。这允许在训练NN时考虑将值量化为定点数格式。包括一个或多个量化块以量化一个或多个输入值集合(或模拟该输入值集合的量化)的NN的模型在本文中称为NN的量化模型。

[0086] 例如，图4示出了包括第一层402和第二层404的示例NN 400，该第一层根据第一权重数据集合W_1和第一偏置集合B_1来处理第一输入激活数据值集合X_1；该第二层根据第二权重数据集合W_2和第二偏置集合B_2来处理第二输入激活数据值集合X2(第一层402的输出)。这种NN 400的浮点模型可以使用一个或多个量化块来扩充，该一个或多个量化块各自将一个或多个输入值集合量化(或将该一个或多个输入值集合的量化模拟)为NN的层，使得在训练NN时可以考虑NN的值的量化。例如，如图4中所示出，可以通过添加以下各项而从NN的浮点模型中生成NN的量化模型420：第一量化块422，该第一量化块将第一输入激活数据值集合X_1量化(或将该第一输入激活数据值集合的量化模拟)为由相应量化参数集合定义的一个或多个定点数格式；第二量化块424，该第二量化块将第一权重数据集合W_1和第一偏置集合B_1量化(或将它们的量化模拟)为由相应量化参数集合定义的一个或多个定点数格式；第三

量化块426，该第三量化块将第二输入激活数据值集合X_2量化(或将该第二输入激活数据值集合的量化模拟)为由相应量化参数集合定义的一个或多个定点数格式；以及第四量化块428，该第四量化块将第二权重数据集合W_2和第二偏置集合B_2量化(或将它们的量化模拟)为由相应量化参数定义的一个或多个定点数格式。

[0087] 向NN的浮点模型添加量化块允许量化参数(例如尾数位长度和指数)本身经由反向传播来确定，只要量化参数是可微分的。具体地，这可以通过使量化参数(例如位长度b和指数exp)可学习并且基于误差度量和NN的实现成本来生成成本度量而实现。成本度量的导数随后可以被反向传播到量化参数(例如位深度b和指数exp)，以产生成本度量相对于量化参数中的每个量化参数的梯度/导数。每个梯度指示对应量化参数(例如位深度或指数)应比现在更高还是更低，以减小成本度量。随后可以基于梯度来调整量化参数以使成本度量最小化。类似于训练NN(即标识NN的权重)，可以重复此过程，直到成本度量收敛为止。

[0088] 测试已表明，使用反向传播来标识NN的量化参数可以生成具有良好性能水平(例如具有高于预定阈值的准确度)以及最少位数的定点NN，这允许NN在硬件中被高效实现。

[0089] 现在参考图5，该图示出了用于经由反向传播来标识NN的量化参数的示例方法500。在示例中，图5的方法500可用于经由反向传播来标识深度神经网络(DNN)——其为NN的一种类型——的量化参数。方法500可以由基于计算的设备来实现，该基于计算的设备诸如下文针对图14描述的基于计算的设备1400。例如，可能存在上面存储有计算机可读指令的计算机可读存储介质，当在基于计算的设备处被执行时，该计算机可读指令使得基于计算的设备执行图5的方法500。

[0090] 该方法开始于块502处，其中确定NN的量化模型响应于训练数据的输出。NN的模型是可用来确定NN响应于输入数据的输出的NN的表示。该模型可以是例如NN的软件实现或NN的硬件实现。确定NN的模型响应于训练数据的输出包括使训练数据传递通过NN的层并且获得其输出。这可以称为NN的前向传递，因为计算流程是从输入通过NN到达输出。该模型可以被配置成使用经训练的权重集合(例如通过训练NN的浮点模型获得的权重集合)。

[0091] NN的量化模型是包括一个或多个量化块的NN的模型(例如，如图4中所示出)。每个量化块被配置成在模型根据NN的层来处理一个或多个值集合之前，对输入到所述层的一个或多个值集合进行变换(例如量化或模拟该一个或多个值集合的量化)。量化块允许量化NN的一个或多个值集合对待测量的NN的输出的影响。

[0092] 如本领域技术人员所知，量化是将呈较高精度数字格式的数字转换为较低精度数字格式的过程。将呈较高精度格式的数字量化为较低精度格式一般包括基于特定舍入模式(诸如但不限于就近舍入(RTN)、向零舍入(RTZ)、就近舍入平局成偶(RTE)、向正无穷舍入(RTP)以及向负无穷舍入(RTNI))来选择呈较低精度格式的可表示数字中的一个可表示数字以表示呈较高精度格式的数字。

[0093] 例如，等式(2)列出用于将呈第一数字格式的值z量化为呈第二较低精度数字格式的值z^q的示例公式，其中X_{max}是呈第二数字格式的最高可表示数字，X_{min}是呈第二数字格式的最低可表示数字，并且RND(z)是舍入函数：

[0094] $$z^q = \begin{cases} X_{max}, \text{如果} z \geq X_{max} \\ X_{min}, \text{如果} z \leq X_{min} \\ 0, \text{如果} z = 0 \\ RND(z), \text{否则} \end{cases} \quad (2)$$

[0095] 等式(2)中列出的公式将呈第一数字格式的值量化为呈基于舍入模式RND(例如RTN、RTZ、RTE、RTP或RTNI)选择的第二数字格式的可表示数字中的一个可表示数字。

[0096] 在本文中描述的示例中,较低精度格式是定点数格式,并且较高精度格式可以是浮点数格式或定点数格式。换句话说,每个量化块被配置成接收呈输入数字格式(其可以是浮点数格式或定点数格式)的一个或多个值集合,并且将这些值集合量化(或将这些值集合的量化模拟)为一个或多个较低精度输出定点数格式。

[0097] 如上文针对图3所描述,NN的每个层接收输入激活数据,并且产生输出数据。层也可以接收权重数据和/或偏置。因此,由量化块变换的值集合可以是输入到层的激活数据值的全部或子集、输入到层的权重数据值的全部或子集,或者输入到层的偏置的全部或子集。借助于示例,以下中的任一者或多者可以被视为待由量化块变换的值集合:输入到层的激活数据的输入通道、输入到层的权重数据的输入通道、输入到层的权重数据的输出通道、输入到层的偏置以及/或者输入到层的权重数据的输出通道和其相关联偏置。

[0098] 每个量化块可以被配置成将特定输入类型的值的不同子集变换(例如量化或将这些子集的量化模拟)为不同输出定点数格式。例如,量化块可以将针对层的输入激活值的第一子集变换为第一输出定点数格式,并且将针对该层的输入激活值的第二子集变换为第二不同的输出定点数格式。换句话说,在示例中,一个量化块可以对输入到层的激活数据的输入通道中的每个输入通道进行变换,那些输入通道中的每个输入通道被变换为相应的(例如可能不同的)输出定点数格式。在其他情况下,每个输入类型可能有多个量化块。例如,可以有多个量化块用于对层的激活数据进行变换,其中这些量化块中的每个量化块对层的激活数据值的仅一部分(或仅子集)进行变换。换句话说,在示例中,每个量化块可以将激活数据的一个输入通道变换为输出定点数格式。

[0099] 由量化块使用的每个输出定点数格式由一个或多个量化参数定义。定义特定输出定点数格式的量化参数可以基于由将实现NN的硬件逻辑所支持的特定定点数格式。例如,每个定点数格式可以由指数exp和尾数位长度b来定义。

[0100] 在块502的第一迭代中,由量化块使用的量化参数可以从所支持的量化参数中随机选择,或者这些量化参数可以以另一方式来选择。例如,在一些情况下,尾数位长度可以被设置为高于由将用于实现NN的硬件所支持的最高位长度的值,使得信息不会因初始量化而丢失。例如,在将用于实现NN的硬件支持16位的最大位长度的情况下,则尾数位长度可以最初被设置为高于16的值(例如20)。

[0101] 一旦NN的模型响应于训练数据的输出已被确定,方法500便进行到块504。

[0102] 在块504处,根据(i)NN的量化模型响应于训练数据的输出以及(ii)基于量化参数集合的NN的实现成本来确定块502中使用的针对量化参数集合的成本度量cm。成本度量cm是量化参数集合的质量的定量测量。在本文中描述的示例中,当量化参数集合用于量化NN的值(或模拟这些值的量化)时,量化参数集合的质量基于NN的误差,并且当使用量化参数

集合时，该质量基于NN的实现成本(例如以位数或字节数来表示)。因此，在一些情况下，成本度量cm可以是误差度量em与实现度量sm的组合。实现度量可以称为实现成本度量或大小度量。在一些示例中，成本度量cm可以被计算为误差度量em和实现度量sm的加权和，如等式(3)中所示，其中α和β分别是应用于误差度量em和实现度量sm的权重。选择权重α和β以实现误差度量与实现度量之间的某种平衡。换句话说，权重用于指示是误差还是实现成本更重要。例如，如果实现度量权重β较小，则成本度量将由误差度量支配，从而产生更准确的网络。相比之下，如果实现度量权重β较大，则成本度量将由实现度量支配，从而产生具有较低准确度的较小网络。然而，在其他示例中，误差度量em与实现度量sm可以以另一合适的方式来组合以生成成本度量cm。

[0103] $cm = (\alpha * em) + (\beta * sm)$ (3)

[0104] 误差度量em可以是当使用特定量化参数集合来量化NN的值(或模拟这些值的量化)时提供NN的量化模型的输出中的误差的定量测量的任何度量。在一些示例中，NN的量化模型响应于训练数据的输出中的误差可以被计算为输出中相对于基线输出的误差。在一些情况下，如图6的600处所示出，基线输出可以是NN的浮点模型的输出(即NN的模型，其中NN的值呈浮点数格式)。因为值一般可以以浮点数格式来更准确地表示或者更精确地表示，所以NN的浮点模型表示将产生最准确输出的NN的模型。因此，由NN的浮点模型生成的输出可以用作基准或基线输出，根据该基准或基线输出来估算由NN的量化模型生成的输出数据的准确度。

[0105] 在其他示例中，如图7的700处所示出，基线输出可以是针对训练数据的基本事实输出。在这些示例中，NN的量化模型的输出中的误差可以指示NN的量化模型的输出相对于训练数据的已知结果的准确度。

[0106] 基线输出与NN的量化模型的输出之间的误差可以以任何合适的方式来确定。在NN是分类网络的情况下，NN的输出可以是分数集合。如本领域技术人员所知，分类网络确定输入数据落入多个类别中的每个类别中的概率。分类NN一般输出数据向量，其中一个元素对应于每个类别，并且这些元素中的每个元素被称为分数。例如，具有1425个潜在类别标签的分类网络可以输出1425个分数的向量。在这些情况下，基线输出与NN的量化模型的输出之间的误差可以被计算为对应分数之间的L1距离。这在等式(4)中示出，其中r是基线输出中的分数集合，并且r'是NN的量化模型的输出中的分数集合：

[0107] $em = \sum_i |r_i - r'_i|$ (4)

[0108] 在其他示例中，分类NN的输出可以替代地是应用于分数的SoftMax函数的输出。如本领域技术人员所知，SoftMax函数是应用于由NN输出的分数的变换，使得与每个分类相关联的值加起来为1。这允许SoftMax函数的输出表示类别上的概率分布。SoftMax函数的输出可以称为SoftMax归一化分数。SoftMax函数可以如等式(5)中所示来表示(有或没有额外温度参数T)，其中s_i是针对类别i的softmax输出，r_i是针对类别i的分数，并且i和j是对应于类别的向量索引。提高温度T使“SoftMax值”“更柔和”(即针对0和1的饱和度更低)，并且从而更容易进行训练抵抗。

[0109] $$s_i(r;T) = \frac{e^{r_i/T}}{\sum_j e^{r_j/T}} \quad (5)$$

[0110] 在分类NN的输出是SoftMax归一化分数集合的情况下，基线输出与NN的量化模型

的输出之间的误差可以被计算为SoftMax函数的输出之间的L1距离。

[0111] 在其他情况下,NN的量化模型响应于训练数据的输出中的误差可以是前第N个分类准确度,其中N是大于或等于一的整数。如本领域技术人员所知,前第N个分类准确度是对正确分类出现在由NN输出的前N个分类中的频繁程度的度量。频现的前N个分类准确度是前第1个和前第5个分类准确度,但可以使用任何前第N个分类准确度。

[0112] 一般来说,将根据误差度量来训练NN(即选择其权重),并且使用在训练时使用的相同误差度量来选择量化参数是有利的。

[0113] 实现度量sm是当使用特定量化参数集合时提供实现NN的硬件相关成本的定量测量的度量。实现度量表示基于一个或多个量化参数来实现NN的成本,在块502中,根据该一个或多个量化参数来对一个或多个值集合进行变换。实现度量可以称为实现成本度量或大小度量。实现NN的硬件相关成本可以包括例如将数据从存储器传输到NNA芯片的成本。当使用特定量化参数集合时,实现度量可以反映NN的性能的某种度量,例如:NN在某个硬件上运行得有多快;或者NN在某个硬件上消耗多少功率。例如,实现度量可以是硬件特定的(例如特定于将实现NN的NN加速器),使得该实现度量可以被定制以反映该硬件的特性,以便NN训练有效地优化针对该硬件的量化参数集合。实现度量可以例如以物理单位(例如焦耳(Joules))或信息单位(例如位或字节)来表示。

[0114] 在简单方法中,实现度量可以取决于用于表示NN的层中的每个层的某些值集合(例如输入激活数据集合、权重数据集合或偏置集合)的位或字节的总数。也就是说,发明人已发现,当在用于标识如本文中所描述的一个或多个量化参数的方法中使用时,可以通过考虑层(例如特别是相邻层)之间的交互来改进此简单方法。例如,考虑包括被配置成输出5个数据通道(例如使用被布置在5个输出通道中的权重数据)的第一层到被配置成输出1000个数据通道(例如使用被布置在1000个输出通道中的权重数据)的第二层的说明性网络。用于评估该网络的实现成本的简单方法可以是评估输入到每个层的权重数据的输出通道的大小(例如以位数形式)的总和。也就是说,每个层的实现成本可以根据用于对输入到该层的权重数据的输出通道中的每个输出通道进行编码的位数的总和来评估,并且网络的实现成本可以由层的实现成本的总和来表示。假设每个输出权重通道包括可比较数量的权重值,此简单方法可以确定第一层(使用被布置在5个输出通道中的权重数据)相对较小,并且第二层(使用被布置在1000个输出通道中的权重数据)相对较大。由此,基于这种实现度量的训练方法可以“瞄准”输入到第二层的权重数据的输出通道(例如基于第二层看起来更大,并且因此减小其大小显然将对NN的实现成本造成较大影响)。然而,此简单方法没有考虑到由第一层生成的输出数据的5个通道中的每个通道将与输入到第二层的权重数据的1000个输出通道进行卷积。因此,减少由第一层生成的输出数据的那5个通道中的任何一个通道的实现成本(例如通过减小输入到第一层的权重数据的输出通道的大小)可以对NN的总推断时间具有显著影响。借助于极端示例来清楚地说明此概念,将输入到第一层的权重数据的5个输出通道中的任何一个输出通道的大小减小到零位,从而使得输出数据的对应通道能够从NN中被省略,这会使将在第二层中执行的计算的量减少1000次乘加运算。用于评估网络的实现成本的简单方法没有考虑层之间的这种类型的交互。应理解,如果使用替代简单方法,将会经历类似的缺点,其中网络的实现成本根据以下各项来评估:由每个层生成的数据的输出通道的大小(例如位数);输入到每个层的权重数据的输入通道的大小(例

如以位数形式);或者输入到每个层的激活数据的输入通道的大小(例如以位数形式)。

[0115] 根据本文中描述的原理,对于NN的多个层中的每个层,实现度量取决于表示来自该层的输出的实现成本的第一贡献(例如针对该层的输出通道的数量),以及表示来自在该层之前的层的输出的实现成本的第二贡献(例如对其确定实现成本的层的输入通道的数量)。也就是说,多个层中的每个层可以提供相应的第一贡献和第二贡献。实现度量可以是根据所述第一贡献和所述第二贡献确定的多个层中的每个层的实现成本的总和。以此方式,可以更好地考虑层之间(例如特别是相邻层)的交互。基于更好地考虑了层之间的交互的这种实现度量的训练方法可以更好地"瞄准"对NN的实现成本具有更大影响的值集合——例如在更大数量的乘加运算中涉及的那些值集合。

[0116] 应理解,不需要以此方式来确定NN的每个层的实现成本以包括在实现度量中。例如,NN的最后一层的实现成本不需要取决于表示来自该层的输出的实现成本的第一贡献,并且/或者NN的第一层的实现成本不需要取决于表示来自在该层之前的层的输出的实现成本的第二贡献。替代地或另外,实现度量可以包括仅来自NN的接收权重数据和/或偏置作为输入的层(例如卷积和/或全连接层)的第一贡献和第二贡献。也就是说,多个层可以包括多个卷积和/或全连接层。换句话说,实现度量可以不包括来自不接收权重数据和/或偏置作为输入的层(例如激活层、归一化层或池化层)的贡献。

[0117] 在示例中,在对其确定实现成本的层之前的层可以是紧接在该层之前的层(例如输出数据的层,该数据是对其确定实现成本的层的输入激活数据);可以是NN中也接收权重数据和/或偏置作为输入的在前层(例如NN的先前卷积层和/或全连接层);或者可以是NN中与对其确定实现成本的层类型相同的在前层(例如,如果对其确定实现成本的层是卷积层,则是NN中的先前卷积层)。换句话说,对其确定实现成本的层与在该层之前的层可以由其他类型的层(例如激活层、归一化层或池化层)和/或中间运算(诸如该层与在该层之前的层之间的求和块)分开。换句话说,对其确定实现成本的层与在该层之前的层可以由不改变该层所接收到的输入激活数据中的数据通道的数量的其他类型的层分开,使得对其确定实现成本的层的输入激活数据以及在该层之前的层的输出数据被布置在相同数量的数据通道中。换句话说,对其确定实现成本的层与在该层之前的层可以由独立处理数据通道(例如不会导致输入数据通道与输出数据通道之间的数据值"混合")的其他类型的层分开。

[0118] 在下文中,提供了九个具体示例,其中实现度量取决于如本文中所描述的多个层中的每个层的第一贡献和第二贡献。应理解,这些具体实现仅借助于示例来提供,并且本文中描述的原理可以不同地来实现。

[0119] 示例1

[0120] 在示例1中,根据输入到层的权重数据的一个或多个输出通道的实现成本来形成第一贡献,并且根据输入到层的激活数据的一个或多个输入通道的实现成本来形成第二贡献。如本文中所描述,层的权重数据中的输出通道的数量对应于(例如等于)该层的输出数据中的通道(例如数据通道)的数量。因此,输入到层的权重数据的一个或多个输出通道的实现成本可以被视为表示来自该层的输出的实现成本。如本文中所描述,输入到层的激活数据是来自在该层之前的层的输出数据(或者直接从该输出数据中导出,例如在诸如层之间的求和块的中间运算的情况下)。因此,输入到层的激活数据的一个或多个输入通道的实现成本可以被视为表示来自在该层之前的层的输出的实现成本。

[0121] 在示例1中，在图5的块502中，由一个或多个量化块变换的一个或多个值集合中的每个值集合是输入到层的值的通道。一个或多个量化参数中的每个量化参数包括相应位宽度，根据该一个或多个量化参数对值的一个或多个通道进行变换。一个或多个量化块被配置成根据相应位宽度b_i^a来对输入到层的激活数据的输入通道i中的一个或多个输入通道中的每个输入通道进行变换(其中位宽度b_i^a可以被表示为向量$\{b_i^a\}_{i=1}^{I}$)，并且根据相应位宽度b_j^w来对输入到层的权重数据的输出通道j中的一个或多个输出通道中的每个输出通道进行变换(其中位宽度b_j^w可以被表示为向量$\{b_j^w\}_{j=1}^{O}$)。更具体地，输入激活数据x和输入权重数据w可以根据等式(6)和(7)进行变换，其中激活数据的相应位宽度b_i^a和指数e_i^a以各自具有I个元素的向量进行编码，权重数据的相应位宽度b_j^w和指数e_j^w以各自具有O个元素的向量进行编码。也就是说，b_i^a和e_i^a使用单独的一对量化参数来量化激活数据x的每个输入通道，并且b_j^w和e_j^w使用单独的一对量化参数来量化权重数据w的每个输出通道。合适的量化函数的示例q在下文描述(例如参考等式(37A)、(37B)或(37C))。

[0122] $$x' = q(x, b_i^a, e_i^a) \qquad (6)$$

[0123] $$w' = q(w, b_j^w, e_j^w) \quad (7)$$

[0124] 在示例1中，在图5的块504中，层s_l的实现成本可以根据等式(8)来定义，该等式是可微分函数。在等式(8)中，第一贡献取决于根据大于零的位宽度b_i^a进行变换的激活数据的输入通道i的数量乘以位宽度b_j^w的总和，权重数据的一个或多个输出通道j中的每个输出通道根据这些位宽度进行变换。第二贡献取决于根据大于零的位宽度b_j^w进行变换的权重数据的输出通道j的数量乘以位宽度b_i^a的总和，激活数据的一个或多个输入通道i中的每个输入通道根据这些位宽度进行变换。在等式(8)中，项$\max(0, b_j^w)$和$\max(0, b_i^a)$可用于确保在该方法的后续步骤中不将位宽度b_j^w和b_i^a分别调整为低于零，如下文进一步详细地描述。层的实现成本s_l根据第一贡献和第二贡献的总和来确定，该总和乘以输入到层的权重数据的高度H_w和宽度W_w维度的乘积。针对NN的实现度量可以通过对如根据等式(8)确定的NN的多个层的实现成本进行求和来形成。

[0125] $$s_l = H_w W_w \left(\sum_{i=1}^{I} \mathbf{1}_{b_i^a > 0} \sum_{j=1}^{O} \max(0, b_j^w) + \sum_{j=1}^{O} \mathbf{1}_{b_j^w > 0} \sum_{i=1}^{I} \max(0, b_i^a) \right) \qquad (8)$$

[0126] 示例2

[0127] 在示例2中，如同在示例1中那样，根据输入到层的权重数据的一个或多个输出通道的实现成本来形成第一贡献，并且根据输入到层的激活数据的一个或多个输入通道的实现成本来形成第二贡献。

[0128] 在示例2中，图5的块502中的一个或多个量化块对输入值集合的变换与参考示例1

描述的变换相同。换句话说，输入激活数据x和输入权重数据w可以根据等式(6)和(7)进行变换，如本文中参考示例1所描述。

[0129] 在示例2中，在图5的块504中，层s_l的实现成本可以根据等式(9)来定义，该等式是可微分函数。在等式(9)中，第一贡献取决于位宽度b_j^w的总和，权重数据的一个或多个输出通道j中的每个输出通道根据这些位宽度进行变换。第二贡献取决于位宽度b_i^a的总和，激活数据的一个或多个输入通道i中的每个输入通道根据这些位宽度进行变换。在等式(9)中，项$\max(0, b_i^a)$和$\max(0, b_j^w)$可用于确保在该方法的后续步骤中不将位宽度b_i^a和b_j^w分别调整为低于零，如下文进一步详细地描述。层的实现成本s_l根据第一贡献和第二贡献的乘积来确定，该乘积乘以输入到层的权重数据的高度H_w和宽度W_w维度的乘积。针对NN的实现度量可以通过对如根据等式(9)确定的NN的多个层的实现成本进行求和来形成。

[0130]
$$s_l = H_w W_w \sum_{i=1}^{I} \max(0, b_i^a) \sum_{j=1}^{O} \max(0, b_j^w) \quad (9)$$

[0131] 示例3

[0132] 在示例3中，根据输入到层的权重数据的一个或多个输出通道的实现成本来形成第一贡献，并且根据输入到层的权重数据的一个或多个输入通道的实现成本来形成第二贡献。如本文中所描述，层的权重数据中的输出通道的数量对应于(例如等于)该层的输出数据中的通道(例如数据通道)的数量。因此，输入到层的权重数据的一个或多个输出通道的实现成本可以被视为表示来自该层的输出的实现成本。如本文中所描述，层的权重数据中的输入通道的数量对应于(例如等于)激活数据中的输入通道的数量，该权重数据将与该激活数据进行组合。此外，如本文中所描述，输入到层的激活数据是来自在该层之前的层的输出数据(或者直接从该输出数据中导出，例如在诸如层之间的求和块的中间运算的情况下)。因此，输入到层的权重数据的一个或多个输入通道的实现成本可以被视为表示来自在该层之前的层的输出的实现成本。

[0133] 在示例3中，在图5的块502中，由一个或多个量化块变换的一个或多个值集合中的每个值集合是输入到层的值的通道。一个或多个量化参数中的每个量化参数包括相应位宽度，根据该一个或多个量化参数对值的一个或多个通道进行变换。在示例3中，为了图5的步骤504的目的，针对输入到层的权重数据的一个或多个输入通道i中的每个输入通道确定相应位宽度b_i(其中位宽度b_i可以被表示为向量$\{b_i\}_{i=1}^{I}$)，并且针对输入到层的权重数据的一个或多个输出通道j中的每个输出通道确定相应位宽度b_j(其中位宽度b_j可以被表示为向量$\{b_j\}_{j=1}^{O}$)。更具体地，输入权重数据w可以根据等式(10A)、(10B)或(10C)进行变换，其中权重数据的输入通道的位宽度b_i以具有I个元素的向量进行编码，并且权重数据的输出通道的位宽度b_j以具有O个元素的向量进行编码。在等式(10A)中，权重数据的输入通道和输出通道的指数e_{ij}以二维矩阵进行编码。换句话说，b_i和e_{ij}使用单独的一对量化参数来量化权重数据w的每个输入通道，并且b_j和e_{ij}使用单独的一对量化参数来量化权重数据w的每个输出通道。合适的量化函数的示例q在下文描述(例如参考等式(37A)、(37B)或(37C))。

[0134] $w'=q(w,min(b_i,b_j),e_{ij})$ (10A)

[0135] $w'=q(q(w,b_i,e_i),b_j,e_j)$ (10B)

[0136] $w'=q(w,min(b_i,b_j),e_j)$ (10C)

[0137] 应理解,如本文中所描述,每个权重值被权重数据的一个输入通道和一个输出通道包括。这意味着对于输入到层的每个权重值,分别确定第一位宽度b_i和第二位宽度b_j。出于图5的块502的目的,如等式(10A)中所示,输入到层的每个权重值可以根据其相应的第一位宽度或第二位宽度——以及与该位宽度相关联的指数(例如,如果选择了b_i,则为e_i,或者如果选择了b_j,则为e_j)——进行变换。可选地,可以选择其相应的第一位宽度和第二位宽度中的较小者(例如最小者)。这在等式(10A)中由项$min(b_i,b_j)$表示。替代地,如等式(10B)中所示,输入到层的每个权重值可以根据其相应的第一位宽度和第二位宽度——以及与那些位宽度相关联的指数——例如在两次传递中——进行变换。也就是说,输入权重数据w可以替代地根据(10B)进行变换。同样替代地,如等式(10C)中所示,输入到层的每个权重值可以根据其相应的第一位宽度或第二位宽度——以及与输出通道相关联的指数j,包括该权重值——进行变换。可选地,可以选择其相应的第一位宽度和第二位宽度中的较小者(例如最小者)。这在等式(10C)中由项$min(b_i,b_j)$表示。权重数据的输出通道的指数e_j可以以具有0个元素的向量进行编码。保存所述向量e_j可以比保存二维指数矩阵e_{ij}消耗更少的存储器空间,如参考等式(10A)所描述。相较于取决于选择第一位宽度和第二位宽度中的哪一者(如等式(10A)中所示的情况)而在与输入通道i相关联的指数和与输出通道j相关联的指数之间进行选择,对于每个变换,不管选择第一位宽度和第二位宽度中的哪一者,使用与输出通道j相关联的指数e_j(如等式(10C)中所示)可以更鲁棒(例如不太可能引起训练误差)。这是因为如果指数在训练期间用于量化更多值(即由于总是使用e_j,而不是e_{ij}),则该指数在训练期间不太可能"跳出范围"(例如由于"大跳跃"而变得对于量化而言太大或太小以致不能给出合理的输出)。

[0138] 在示例3中,层s_l的实现成本可以根据等式(11)来定义,该等式是可微分函数。在等式(11)中,第一贡献取决于针对权重数据的一个或多个输出通道j中的每个输出通道确定的位宽度b_j的总和。第二贡献取决于针对权重数据的一个或多个输入通道i中的每个输入通道确定的位宽度b_i的总和。在等式(11)中,项$max(0,b_i)$和$max(0,b_j)$可用于确保在该方法的后续步骤中不将位宽度b_i和b_j分别调整为低于零,如下文进一步详细地描述。层的实现成本s_l根据第一贡献和第二贡献的乘积来确定,该乘积乘以输入到层的权重数据的高度H_w和宽度W_w维度的乘积。针对NN的实现度量可以通过对如根据等式(11)确定的NN的多个层的实现成本进行求和来形成。

[0139]
$$s_l = H_w W_w \sum_{i=1}^{I} \max(0, b_i) \sum_{j=1}^{O} \max(0, b_j) \quad (11)$$

[0140] 示例4

[0141] 在示例4中,根据输入到层的权重数据的一个或多个输出通道的实现成本以及输入到层的一个或多个偏置的实现成本来形成第一贡献。根据输入到在前层的权重数据的一个或多个输出通道的实现成本以及输入到在前层的一个或多个偏置的实现成本来形成第

二贡献。如本文中所描述，层的权重数据中的输出通道的数量对应于(例如等于)该层的输出数据中的通道(例如数据通道)的数量。此外，在使用偏移偏置的层中，权重数据的输出通道中的每个输出通道与相应偏置相关联。因此，输入到层的权重数据的一个或多个输出通道的实现成本以及输入到该层的一个或多个偏置的实现成本可以被视为表示来自该层的输出的大小。出于相同原因，输入到在前层的权重数据的一个或多个输出通道的实现成本以及输入到该在前层的一个或多个偏置的实现成本可以被视为表示来自该在前层的输出的实现成本。

[0142] 在示例4中，在图5的块502中，一个或多个量化参数中的每个量化参数包括相应位宽度。由一个或多个量化块变换的一个或多个值集合包括输入到层的权重数据的一个或多个输出通道和相关联偏置，以及输入到在前层的权重数据的一个或多个输出通道和相关联偏置。一个或多个量化块被配置成根据相应位宽度b_j^w来对输入到层的权重数据的一个或多个输出通道j中的每个输出通道进行变换(其中位宽度b_j^w可以被表示为具有0个元素的向量$\{b_j^w\}_{j=1}^{O}$)，根据相应位宽度b_j^β来对输入到层的一个或多个偏置j中的每个偏置进行变换(其中位宽度b_j^β可以被表示为具有0个元素的向量$\{b_j^\beta\}_{j=1}^{O}$)，根据相应位宽度b_i^w来对输入到在前层的权重数据的一个或多个输出通道i中的每个输出通道进行变换(其中位宽度b_i^w可以被表示为具有I个元素的向量$\{b_i^w\}_{i=1}^{I}$)，并且根据相应位宽度b_i^β来对输入到在前层的一个或多个偏置i中的每个偏置进行变换(其中位宽度b_i^β可以被表示为具有I个元素的向量$\{b_i^\beta\}_{i=1}^{I}$)。可选地，可以使用相同位宽度来对权重数据的输出通道和所述输出通道的相关联偏置进行变换。也就是说，b_j^w可以等于b_j^β，并且/或者b_i^w可以等于b_i^β。更具体地，输入到层的权重数据w_j可以根据等式(12)进行变换，输入到层的偏置β_j可以根据等式(13)进行变换，输入到在前层的权重数据w_i可以根据等式(14)进行变换，输入到在前层的偏置β_i可以根据等式(15)进行变换。在等式(12)至(15)中，e_j^w、e_j^β、e_i^w和e_i^β分别是用于对w_j、β_j、w_i和β_i进行变换的指数。e_j^w、e_j^β可以以具有0个元素的向量进行编码。e_i^w、e_i^β可以以具有I个元素的向量进行编码。合适的量化函数的示例q在下文描述(例如参考等式(37A)、(37B)或(37C))。

[0143] $$w_j' = q(w_j, b_j^w, e_j^w) \quad (12)$$

[0144] $$\beta_j' = q\left(\beta_j, b_j^\beta, e_j^\beta\right) \quad (13)$$

[0145] $$w_i' = q(w_i, b_i^w, e_i^w) \quad (14)$$

[0146] $$\beta_i' = q\left(\beta_i, b_i^\beta, e_i^\beta\right) \quad (15)$$

[0147] 在示例4中，层s_1的实现成本可以根据等式(16)来定义，该等式是可微分函数。在等式(16)中，第一贡献取决于输入到在前层的权重数据的输出通道以及其输入到在前层的相关联偏置中的一者或两者根据大于零的位宽度进行变换的实例的数量，乘以输入到层的

权重数据的一个或多个输出通道中的每个输出通道进行变换所根据的位宽度以及输入到层的一个或多个相关联偏置中的每个相关联偏置进行变换所根据的位宽度的加权和的总和。在等式(16)中，加权和由项α加权。第二贡献取决于输入到层的权重数据的输出通道以及其输入到层的相关联偏置中的一者或两者根据大于零的位宽度进行变换的实例的数量，乘以输入到在前层的权重数据的一个或多个输出通道中的每个输出通道进行变换所根据的位宽度以及输入到在前层的一个或多个相关联偏置中的每个相关联偏置进行变换所根据的位宽度的加权和的总和。在等式(16)中，加权和由项α加权。在等式(16)中，项$\max\left(0,b_j^w\right)$、$\max\left(0,b_j^\beta\right)$、$\max(0,b_i^w)$和$\max\left(0,b_i^\beta\right)$可用于确保在该方法的后续步骤中不将位宽度$b_j^w$、$b_j^\beta$、$b_i^w$和$b_i^\beta$分别调整为低于零，如下文进一步详细地描述。层的实现成本s_l根据第一贡献和第二贡献的总和来确定，该总和乘以输入到层的权重数据的高度H_w和宽度W_w维度的乘积。针对NN的实现度量可以通过对如根据等式(16)确定的NN的多个层的实现成本进行求和来形成。

[0148]
$$s_l = H_w W_w \left(\sum_{i=1}^{I} \mathbf{1}_{\left(b_i^w>0 \; OR \; b_i^\beta>0\right)} \sum_{j=1}^{O} \left(\max\left(0,b_j^w\right) + \alpha \max\left(0,b_j^\beta\right) \right) + \sum_{j=1}^{O} \mathbf{1}_{\left(b_j^w>0 \; OR \; b_j^\beta>0\right)} \sum_{i=1}^{I} \left(\max(0,b_i^w) + \alpha \max\left(0,b_i^\beta\right) \right) \right) \tag{16}$$

[0149] 示例5

[0150] 在示例5中，根据输入到层的权重数据的一个或多个输出通道的实现成本来形成第一贡献，并且根据输入到在前层的权重数据的一个或多个输出通道的实现成本来形成第二贡献。如本文中所描述，层的权重数据中的输出通道的数量对应于(例如等于)该层的输出数据中的通道(例如数据通道)的数量。因此，输入到层的权重数据的一个或多个输出通道的实现成本可以被视为表示来自该层的输出的实现成本。响应于确定该层和在前层没有接收到偏置，可以优先于示例4使用示例5。

[0151] 在示例5中，在图5的块502中，一个或多个量化参数中的每个量化参数包括相应位宽度。由一个或多个量化块变换的一个或多个值集合包括输入到层的权重数据的一个或多个输出通道，以及输入到在前层的权重数据的一个或多个输出通道。一个或多个量化块被配置成根据相应位宽度b_j来对输入到层的权重数据的输出通道j中的一个或多个输出通道中的每个输出通道进行变换(其中位宽度b_j可以被表示为具有0个元素的向量$\{b_j\}_{j=1}^{O}$)，并且根据相应位宽度b'_i来对输入到在前层的权重数据的一个或多个输出通道i中的每个输出通道进行变换(其中位宽度b'_i可以被表示为具有I个元素的向量$\{b'_i\}_{i=1}^{I}$)。更具体地，输入到层的权重数据w_j可以根据等式(17)进行变换，并且输入到在前层的权重数据w'_i可以根据等式(18)进行变换。在等式(17)和(18)中，e_j和e'_i分别是用于对w_j和w'_i进行变换的指数。e_j可以以具有0个元素的向量进行编码。e'_i可以以具有I个元素的向量进行编码。合适的量化函数的示例q在下文描述(例如参考等式(37A)、(37B)或(37C))。

[0152] $$\dot{w}_j = q(w_j, b_j, e_j) \quad (17)$$

[0153] $$\dot{w}'_i = q(w'_i, b'_i, e'_i) \quad (18)$$

[0154] 在示例5中，层s_1的实现成本可以根据等式(19)来定义，该等式是可微分函数。在等式(19)中，第一贡献取决于输入到在前层的权重数据的输出通道根据大于零的位宽度进行变换的实例的数量，乘以输入到层的权重数据的一个或多个输出通道中的每个输出通道进行变换所根据的位宽度的总和。第二贡献取决于输入到层的权重数据的输出通道根据大于零的位宽度进行变换的实例的数量，乘以输入到在前层的权重数据的一个或多个输出通道中的每个输出通道进行变换所根据的位宽度的总和。在等式(19)中，项$\max(0,b_j)$和$\max(0,b'_i)$可用于确保在该方法的后续步骤中不将位宽度b_j和b'_i分别调整为低于零，如下文进一步详细地描述。层的实现成本s_1根据第一贡献和第二贡献的总和来确定，该总和乘以输入到层的权重数据的高度H_w和宽度W_w维度的乘积。针对NN的实现度量可以通过对如根据等式(19)确定的NN的多个层的实现成本进行求和来形成。

[0155] $$s_l = H_w W_w \left(\sum_{i=1}^{I} \mathbf{1}_{(b'_i>0)} \sum_{j=1}^{O} \max(0, b_j) + \sum_{j=1}^{O} \mathbf{1}_{b_j>0} \sum_{i=1}^{I} \max(0, b'_i) \right) \quad (19)$$

[0156] 示例6

[0157] 在示例6中，第一贡献和第二贡献与针对示例5描述的第一贡献和第二贡献相同。也就是说，相对于示例5，在示例6中，层s_1的实现成本进一步取决于表示输入到在前层的偏置(β'_i)的实现成本的额外贡献。响应于确定在前层接收到偏置，可以优先于示例5使用示例6。

[0158] 在示例6中，图5的块502中的一个或多个量化块对输入值集合的变换与参考示例5描述的变换相同。换句话说，输入到层的权重数据w_j的一个或多个输出通道以及输入到在前层的权重数据w'_i的一个或多个输出通道可以根据等式(17)和(18)进行变换，如本文中参考示例6所描述。

[0159] 在示例6中，层s_1的实现成本可以根据等式(20)来定义，该等式是可微分函数。在等式(20)中，第一贡献和第二贡献与等式(19)中所示的第一贡献和第二贡献相同。在等式(20)中，第一贡献和第二贡献的总和乘以输入到层的权重数据的高度H_w维度和宽度W_w维度的乘积。在等式(20)中，额外贡献取决于输入到在前层的权重数据的输出通道根据零或小于零的位宽度进行变换的实例的数量，乘以输入到层的权重数据的输出通道根据大于零的位宽度进行变换的实例的数量，乘以输入到在前层的偏置(β'_i)的绝对值。应理解，输入到在前层的偏置(β'_i)可以被量化，也可以不被量化。如等式(20)中所示，可选地，此额外贡献乘以输入到层的权重数据的高度H_w维度和宽度W_w维度的乘积。如等式(20)中所示，可选地，此额外贡献由项α加权。针对NN的实现度量可以通过对如根据等式(20)确定的NN的多个层的实现成本进行求和来形成。

[0160]
$$s_l = H_w W_w \left(\sum_{i=1}^{I} \mathbf{1}_{(b_i'>0)} \sum_{j=1}^{O} \max(0, b_j) + \sum_{j=1}^{O} \mathbf{1}_{b_j>0} \sum_{i=1}^{I} \max(0, b_i') \right) + \alpha H_w W_w \sum_{i=1}^{I} \mathbf{1}_{b_i' \le 0} \sum_{j=1}^{O} \mathbf{1}_{b_j>0} |\beta_i'| \quad (20)$$

[0161] 示例7

[0162] 在许多NN结构中，每个层的激活输入从仅一个在前层的激活输出中导出。也就是说，在某些NN结构中，层的激活输入可以从多于一个在前层的激活输出中导出。这种NN的一个示例在图10C中示出，该图是示出包括剩余层的NN的示意图。在图10C中，求和运算1020从层E 1012和层F 1016两者接收输入。求和运算1020的输出被输入到层G 1018。也就是说，层G 1018的激活输入是从两个在前层——层E 1012和层F 1016——的激活输出中导出的。示例7涉及确定接收激活输入数据的层的实现成本，该激活输入数据已从多于一个在前层的激活输出中导出。

[0163] 在示例7中，层(例如层G 1018)的实现度量取决于：表示来自该层(例如层G 1018)的输出的实现成本的第一贡献；表示来自在该层之前的第一层(例如层E 1012)的输出的实现成本的第二贡献；以及表示来自在该层之前的第二层(例如层F 1016)的输出的实现成本的第三贡献。第一贡献可以根据与参考示例1至6中的任何示例描述的第一贡献相同的因素来形成。第二贡献可以根据与参考示例1至6中的任何示例描述的第二贡献相同的因素来形成。第三贡献可以根据与参考示例1至6中的任何示例描述的第二贡献相同的因素来形成。在示例7中，根据本文中参考示例6描述的原理，针对层的实现度量可以进一步取决于表示输入到第一在前层和第二在前层的偏置的实现成本的额外贡献。

[0164] 为了给出其中对实现度量的贡献是基于与参考示例5描述的因素相同的因素的一个具体示例，在示例7中，层的实现成本s_1可以根据等式(21)来定义，该等式是可微分函数。在等式(21)中，上标E、F和G用于指与第一在前层(例如层E 1012)、第二在前层(例如层F 1016)以及对其确定实现成本的层(例如层G 1018)相关联的项。一个或多个量化块被配置成根据相应位宽度b_j^G来对输入到层(例如层G)的权重数据的一个或多个输出通道j中的每个输出通道进行变换(其中位宽度b_j^G可以被表示为具有O个元素的向量$\{b_j^G\}_{j=1}^{O}$)，根据相应位宽度b_i^E来对输入到第一在前层(例如层E)的权重数据的一个或多个输出通道i中的每个输出通道进行变换(其中位宽度b_i^E可以被表示为具有I个元素的向量$\{b_i^E\}_{i=1}^{I}$)，并且根据相应位宽度b_i^F来对输入到第二在前层(例如层F)的权重数据的一个或多个输出通道i中的每个输出通道进行变换(其中位宽度b_i^F可以被表示为具有I个元素的向量$\{b_i^F\}_{i=1}^{I}$)。参考如本文中参考示例5所描述的等式(17)和(18)，本领域技术人员将理解可以如何执行这些变换。层的实现成本s_1根据第一贡献、第二贡献和第三贡献的总和来确定，该总和乘以输入到对其确定实现成本的层(例如层G 1018)的权重数据的高度H_w维度和宽度W_w维度的乘积。针对NN的实现度量可以通过对如根据等式(21)确定的NN的多个层的实现成本进行求和来形成。

[0165]
$$s_l = H_w W_w \left(\sum_{j=1}^{O} \mathbf{1}_{b_j^G>0} \sum_{i=1}^{I} \max(0, b_i^E) + \sum_{j=1}^{O} \mathbf{1}_{b_j^G>0} \sum_{i=1}^{I} \max(0, b_i^F) + \sum_{i=1}^{I} \mathbf{1}_{(b_i^E>0 \ OR \ b_i^F>0)} \sum_{j=1}^{O} \max(0, b_j^G) \right) \quad (21)$$

[0166] 示例8

[0167] 在许多NN结构中，每个层的输出被输入到仅一个其他层(或从NN输出)。也就是说，在某些NN结构中，层的输出可以被输入到多于一个后续层。这种NN的一个示例在图10D中示出，该图是示出包括剩余层的NN的示意图。在图10D中，层T 1032的输出被输入到层U 1038和层V 1036两者。

[0168] 参考图10D，可能不需要确定针对例如层V 1036的实现成本，该实现成本取决于表示来自层T 1032的输出的实现成本的第二贡献。这是因为，在该方法的后续阶段(下文进一步详细地描述)中，至少部分地基于此第二贡献来调整一个或多个量化参数，并且可选地根据经调整的量化参数从NN的模型中移除值集合。调整用于对输入到层T 1032的权重数据进行变换的量化参数、调整用于对从层T 1032输出的激活数据进行变换的量化参数，或者甚至从层T 1032的输入/输出中移除值集合都可能影响在层U 1038处执行的计算。

[0169] 可以使用示例8以便防止针对层V 1036形成的实现度量潜在地影响在层U 1038处执行的计算。参考图10E，在示例8中，向层T 1032与层V 1036之间的NN添加新的层X 1034。层X 1034可以被配置成接收由层T 1032输出的激活数据，并且将该激活数据输出到层V 1036。也就是说，层X 1034不需要对由层T 1032输出的激活数据执行任何计算。换句话说，层X 1034不接收任何权重数据或偏置。可以向NN的量化模型添加一个或多个量化块，以根据相应量化参数来对输入到新的层X的值集合进行变换。针对层V 1036的实现度量随后可以使用作为在前层的层X 1034(即，而不是层T 1032)来形成。所述实现度量可以使用本文中参考示例1至3中的任何示例描述的原理来形成。由于层X 1034的输出仅被提供给层V 1036(即，并且不提供给层U 1038)，因而针对用于对从层X 1034输出的激活数据进行变换的量化参数的任何后续调整或者从由层X 1034输出的激活数据中对值集合的任何移除都不会影响在层U 1038处执行的计算。

[0170] 尽管图10E中未示出，但可以执行相同步骤以便形成层U 1038的实现度量。也就是说，可以在层T 1032与层U 1038之间添加新的层。为了计算层U 1038的实现成本的目的，该新的层可以被视为在前层。

[0171] 示例9

[0172] 在一些NN结构中，本文中参考示例7和8描述的方法可以进行组合。这种NN结构的一个示例在图10A中示出，该图是示出包括剩余层的NN的示意图。在图10A中，层A 1002的输出被输入到层B 1004和求和运算1010。层B 1004的输出被输入到层C 1006。求和运算1010从层A 1002和层C 1006两者接收输入。求和运算1010的输出被输入到层D 1008。

[0173] 这意味着层D 1008的激活输入是从两个在前层——层A 1002和层C 1006——的激活输出中导出的。也就是说，层A 1002的输出也被输入到层B 1004。因此，使用示例7执行

本文中描述的方法以形成层D 1008的取决于表示来自层A 1002的输出的实现成本的贡献的实现度量可能影响在层B 1004处执行的计算。

[0174] 因此，在示例9中，根据参考示例8描述的原理，可以在层A 1002与求和运算1010之间添加新的层(图10A中未示出)。随后，根据参考示例7描述的原理，可以形成针对层D 1008的实现度量，该实现度量取决于：表示来自该层(例如层D 1008)的输出的实现成本的第一贡献；表示来自在该层之前的第一层(例如新添加的层——图10A中未示出)的输出的实现成本的第二贡献；以及表示来自在该层之前的第二层(例如层C 1006)的输出的实现成本的第三贡献。

[0175] 应理解，在实现度量中，多个层中的不同层的实现成本不需要以相同方式来计算。例如，多个层中的第一层的实现成本可以根据示例1来计算，而多个层中的第二层的实现成本可以根据示例4来计算，等等。返回图5，一旦已经确定了针对量化参数集合的成本度量cm，方法500便进行到块506。

[0176] 在块506处，成本度量cm的导数被反向传播到一个或多个量化参数，以生成成本度量相对于一个或多个量化参数中的每个量化参数的梯度。

[0177] 如本领域技术人员所知，函数在特定点处的导数是该函数在该点处变化的速率或速度。导数是可分解的，并且因此可以被反向传播到NN的参数，以生成成本度量相对于那些参数的导数或梯度。如上文所描述，反向传播(其也可以称为误差的反向传播)是一种用于训练NN以计算误差度量相对于NN的权重的梯度的方法。反向传播也可用于确定成本度量cm相对于量化参数(例如位宽度b和指数exp)的导数($\frac{\partial cm}{\partial qp_i}$)。成本度量cm的导数到量化参数的反向传播可以例如使用任何合适的工具来执行，该工具用于使用诸如但不限于TensorFlow™或PyTorch™的反向传播来训练NN。

[0178] 成本度量相对于特定量化参数($\frac{\partial cm}{\partial qp_i}$)的梯度指示向哪个方向移动量化参数以减小成本度量cm。具体地，正梯度指示成本度量cm可以通过减小量化参数来减小；并且负梯度指示成本度量cm可以通过增大量化参数来减小。例如，图8示出了相对于特定位宽度b_i的示例成本度量cm的曲线图800。曲线图800示出了当位宽度b_i具有第一值x_1时实现最低成本度量。从曲线图800中可以看出，当位宽度b_i小于x_1(例如当其具有第二值x_2)时，该位宽度具有负梯度802，并且成本度量cm可以通过增大位宽度b_i来减小。类似地，当位宽度b_i大于x_1(例如当其具有第三值x_3)时，该位宽度具有正梯度804和成本度量cm。成本度量cm相对于特定量化参数的梯度在本文中可以称为针对量化参数的梯度。

[0179] 一旦成本度量的导数被反向传播到一个或多个量化参数以生成成本度量针对那些量化参数中的每个量化参数的梯度，方法500便进行到块508。

[0180] 在块508处，基于梯度来调整量化参数中的一个或多个量化参数(例如位宽度b_i和指数exp_i)。方法500的目标是标识将产生‘最佳’成本度量的量化参数集合。构成‘最佳’成本度量的内容将取决于如何计算成本度量。例如，在一些情况下，成本度量越低，成本度量越好，而在其他情况下，成本度量越高，成本度量越好。

[0181] 如上文所描述，针对量化参数的梯度的符号指示成本度量是否将通过增大或减小量化参数而减小。具体地，如果针对量化参数的梯度是正的，则量化参数的减小将减小成本度量；并且如果针对量化参数的梯度是负的，则量化参数的增大将减小成本度量。因此，调

整量化参数可以包括根据梯度的符号来增大或减小量化参数，以便增大或减小成本度量(取决于是期望增大还是减小成本度量)。例如，如果较低成本度量是期望的，并且针对量化参数的梯度是负的，则量化参数可以被增大以试图减小成本度量。类似地，如果较低成本度量是期望的，并且针对量化参数的梯度是正的，则量化参数可以被减小以试图减小成本度量。

[0182] 在一些情况下，量化参数增大或减小的量可以基于梯度的量值。特别地，在一些情况下，量化参数可以增大或减小达梯度的量值。例如，如果梯度的量值是0.4，则量化参数可以增大或减小0.4。在其他情况下，量化参数可以增大或减小达梯度的量值的因子。

[0183] 更一般地，当目标是减小成本度量cm时，经调整的量化参数(qp_{adj})可以通过从量化参数(qp)中减去针对该量化参数的梯度(g_{qp})来生成，如等式(22)中所示。在一些情况下，有可能通过将梯度乘以学习速率l来对调整不同量化参数的速率进行调整，如等式(23)中所示。学习速率越高，量化参数将被调整得越快。对于不同量化参数，学习速率可以不同。

[0184] $qp_{adj}=qp-g_{qp}$ (22)

[0185] $qp_{adj}=qp-l*g_{qp}$ (23)

[0186] 通常，用于实现NN的硬件可以仅支持整数位宽度b_i和指数exp_i，并且在一些情况下可以仅支持针对位宽度和/或指数的特定整数值集合。例如，用于实现NN的硬件逻辑可以仅支持4、5、6、7、8、10、12和16的位宽度。因此，在量化参数被用于在硬件中实现NN之前，量化参数被舍入到最接近的整数或者所支持的整数集合中最接近的整数。例如，如果根据该方法将最佳位宽度确定为4.4，则在位宽度被用于在硬件中实现NN之前，该位宽度可以被量化(例如舍入)为最接近的(RTN)整数(在此情况下为4)。

[0187] 因此，在一些情况下，为了考虑在硬件中实现NN时发生的量化参数的量化(例如舍入)，当标识‘最佳’量化参数时，在于下一迭代中使用增大/减小的量化参数之前，可以将增大/减小的量化参数舍入到最接近的整数或者整数集合的最接近的整数，如等式(24)中所示，其中RTN是被舍入到最接近的整数函数，并且qp_{adj}^{r}是在其已被舍入到最接近的整数之后增大/减小的量化参数。例如，在根据与特定位宽度相关联的梯度来增大或减小该位宽度之后，在用于下一迭代中之前，可以将增大或减小的位宽度舍入到最接近的整数或者集合{4,5,6,7,8,10,12,16}中的最接近者。

[0188] $$qp_{adj}^{r}=RTN\left(qp_{adj}\right) \quad (24)$$

[0189] 在其他情况下，不是在量化参数已被增大/减小之后对量化参数进行实际量化(例如舍入)，而是量化参数的量化(例如舍入)所表示的变换可以仅被模拟。例如，在一些情况下，不是将增大/减小的量化参数舍入到最接近的整数或者集合中最接近的整数，而是可以通过对增大/减小的量化参数执行随机量化来模拟量化。对增大/减小的量化参数执行随机量化可以包括向增大/减小的量化参数添加-a与+a之间的随机值u以生成随机化量化参数，其中a是以下两者之间的距离的一半：(i)集合中与增大/减小的量化参数最接近的小于增大/减小的量化参数的整数，以及(ii)集合中与增大/减小的量化参数最接近的大于增大/减小的量化参数的整数；以及随后将随机化量化参数设置为这两个最接近的整数中的最接近者。当随机量化用于模拟舍入到最接近的整数时，则a等于0.5，并且随机量化可以如等式(25)中所示来实现，其中RTN是舍入到最接近的整数函数，并且qp_{adj}^{s}是在随机量化之后增

大/减小的量化参数。

[0190] $$qp_{adj}^{s} = RTN(qp_{adj} + u) \quad \text{其中} u \leftarrow \mathcal{U}(-0.5,0.5) \quad (25)$$

[0191] 例如，如果在硬件实现中，位宽度可以是集合{4,5,6,7,8,10,12,16}中的任何整数，如果位宽度b_i增大/减小到4.4，则向增大/减小的位宽度b_i添加-0.5与+0.5之间的随机值，因为集合中最接近的较低整数与较高整数(4与5)之间的距离为1；并且随后将随机化位宽度设置为那两个最接近的整数(4和5)中的最接近者。类似地，如果位宽度b_i增大/减小到10.4，则向增大/减小的位宽度b_i添加-1与+1之间的随机值，因为集合中最接近的较低整数与较高整数(10、12)之间的距离为2；并且随后将随机化位宽度设置为那两个最接近的整数(10、12)中的最接近者。以此方式，增大/减小的量化参数以与到该整数的距离成比例的概率被向上或向下舍入到整数。例如，4.2将以20％的概率被舍入到4，并且以80％的概率被舍入到5。类似地，7.9将以10％的概率被舍入到7，并且以90％的概率被舍入到8。测试已表明，在一些情况下，通过向增大/减小的量化参数添加随机值并且随后舍入，而不是仅对增大/减小的量化参数进行舍入，可以更高效且有效地标识量化参数。

[0192] 在其他情况下，不是将增大/减小的量化参数舍入到最接近的整数或者集合中最接近的整数，而是可以通过对增大/减小的量化参数执行均匀噪声量化来模拟对量化参数的量化。对增大/减小的量化参数执行均匀噪声量化可以包括向增大/减小的量化参数添加-a与+a之间的随机值u，其中如上文所描述，a是以下两者之间的距离的一半：(i)集合中与增大/减小的量化参数最接近的小于增大/减小的量化参数的整数，以及(ii)集合中与增大/减小的量化参数最接近的大于增大/减小的量化参数的整数。当使用均匀噪声量化来模拟舍入到最接近的整数时，则a等于0.5，并且均匀噪声量化可以如等式(26)中所示来实现，其中qp_{adj}^{u}是在均匀噪声量化之后增大/减小的参数。通过仅向增大/减小的量化参数添加随机值，增大/减小的量化参数以与舍入增大/减小的量化参数类似的方式失真。

[0193] $$qp_{adj}^{u} = qp_{adj} + u \quad \text{其中} u \leftarrow \mathcal{U}(-0.5,0.5) \quad (26)$$

[0194] 在又其他情况下，不是将增大/减小的量化参数舍入到最接近的整数或者集合中最接近的整数，而是可以通过对增大/减小的量化参数执行梯度平均量化来模拟对量化参数的量化。执行梯度平均量化可以包括取小于或等于增大/减小的量化参数的可允许整数中的最高可允许整数，以及随后添加0与c之间的随机值h，其中c是以下两者之间的距离：(i)集合中与增大/减小的量化参数最接近的小于增大/减小量的化参数的整数，以及(ii)集合中与增大/减小的量化参数最接近的大于增大/减小的量化参数的整数(或者通过在数学上与上文等效的任何运算)。当使用梯度平均量化来模拟舍入到最接近的整数时，则c等于1，并且梯度平均量化可以如等式(27)中所示来实现，其中RTNI是向负无穷舍入函数(其也可以称为地板函数)，并且qp_{adj}^{a}是在梯度平均量化之后增大/减小的量化参数。

[0195] $$qp_{adj}^{a} = RTNI(qp_{adj}) + h \quad \text{其中} h \leftarrow H(0,1) \quad (27)$$

[0196] 例如，如果位宽度b_i可以是集合{4,5,6,7,8,10,12,16}中的任何整数，并且特定位宽度b_i根据梯度增大/减小到4.4，则选择集合中小于或等于增大/减小的量化参数的最高整数(即4)，并且向该最高整数添加0与1之间的均匀随机值，因为集合中最接近的较低整数与较高整数(4与5)之间的距离为1。类似地，如果位宽度b_i根据梯度增大/减小到10.4，则

选择集合中小于或等于该值的最高整数(即10),并且向该最高整数添加0与2之间的随机值,因为集合中最接近的较低整数与较高整数(10与12)之间的距离为2。

[0197] 测试已表明,对于所量化的参数在很大程度上独立的问题,梯度平均量化方法工作良好,但在对高度相关的参数进行优化时并不太良好。

[0198] 在又其他情况下,不是将增大/减小的量化参数舍入到最接近的整数或者集合中最接近的整数,而是可以通过执行双峰量化来模拟对量化参数的量化,该双峰量化是舍入到最接近的整数量化(例如等式(24))与梯度平均量化(例如等式(27))的组合。具体地,在双峰量化中,以概率p对增大/减小的量化参数执行梯度平均量化,并且以其他方式对增大/减小的量化参数执行舍入量化。当使用双峰量化来模拟舍入到最接近的整数时,p是到最接近的整数的距离的两倍,并且双峰量化可以如等式(28)中所示来实现,其中qp_{adj}^{b}是在其双峰量化之后增大/减小的量化参数。

[0199] $$qp_{adj}^{b}=\begin{cases}qp_{adj}^{r} & 如果 1-2\left|qp_{adj}-RND\left(qp_{adj}\right)\right|>u \quad 其中 u\leftarrow\mathcal{U}(0,1)\\ qp_{adj}^{a} & 否则\end{cases} \tag{28}$$

[0200] 其中集合中的连续整数之间的差不是恒定的有序整数集合被称为非均匀整数集合。例如,有序整数集合{4,5,6,7,8,10,12,16}是非均匀整数集合,因为整数4与5之间的差是一,但整数12与16之间的差是四。相比之下,有序整数集合{1,2,3,4,5}是均匀整数集合,因为任何两个连续整数之间的差是一。

[0201] 如上文所描述,为了模拟将增大/减小的量化参数舍入到非均匀整数集合中最接近的整数,基于集合中低于增大/减小的量化参数的最接近的整数与集合中高于如上文所描述的增大/减小的量化参数的最接近的整数之间的差,量化参数(例如a或c)可以被选择用于上述量化模拟方法中的一种量化模拟方法(例如随机量化、均匀噪声量化、梯度平均量化或双峰量化),并且根据期望的模拟方法来对增大/减小的量化参数进行量化。在其他情况下,将增大/减小的量化参数舍入到非均匀整数集合中最接近的整数可以通过以下操作来模拟:(1)基于非均匀整数集合中最接近的较低整数与非均匀整数集合中最接近的较高整数之间的距离/差(其可以被描述为值的局部“密度”)来对增大/减小的量化参数进行缩放,以生成经变换或经缩放的增大/减小的量化参数;(2)使用上文描述的模拟方法中的一种模拟方法(例如等式(25)、(26)、(27)或(28))来模拟将经变换的增大/减小的量化参数舍入到最接近的整数;以及(3)将步骤(1)中执行的变换或缩放进行反转,以获得最终量化的增大/减小的量化参数。

[0202] 这将借助于示例来进一步描述。在此示例中,非均匀整数集合是{4,5,6,7,8,10,12,16}。在步骤(1)中,基于非均匀整数集合中最接近的较低整数与非均匀整数集合中最接近的较高值之间的距离/差来对增大/减小的量化参数进行缩放。具体地,经变换或经缩放的增大/减小的量化参数等于增大/减小的量化参数除以集合中最接近的较低整数与集合中最接近的较高整数之间的距离。例如,当集合中最接近的较低整数(即8或10)与集合中最接近的较高整数(即10或12)之间的距离为2时,8与12之间的增大/减小的量化参数被缩放(乘以)1/2;当集合中最接近的较低整数(即12或14)与集合中最接近的较高整数(即14或16)之间的距离为4时,12与16之间的增大/减小的量化参数被缩放1/4;并且当集合中最接近的较低整数(即4、5、6、7)与集合中最接近的较高整数(即5、6、7、8)之间的距离为1时,4与

8之间的增大/减小的量化参数被缩放1。例如，13被变换为3.25；5.4被变换为5.4；8.9被变换为4.45；并且11.5被变换为5.75。此变换可以由等式(29)表示，其中qp_{adj}是增大/减小的量化参数，qp_{adj}^{t}是经变换的增大/减小的量化参数，并且s如等式(30)中所示，其中当$qp_{adj}>8$时，$I_{qp_{adj}>8}$是1，否则是0，当$qp_{adj}>12$时，$I_{qp_{adj}>82}$是1，否则是0，使得对于$qp_{adj}<8$，s＝1，对于$8<qp_{adj}<12$，s＝2，并且对于$qp_{adj}>12$，s＝4。

[0203] $$qp_{adj}^{t}=\frac{qp_{adj}}{s} \quad (29)$$

[0204] $$s=\left(1+I_{qp_{adj}>8}\right)\left(1+I_{qp_{adj}>12}\right) \quad (30)$$

[0205] 在步骤(2)中，使用上文描述的用于模拟舍入到最接近的整数的方法中的一种方法(例如等式(25)、(26)、(27)或(28))来模拟将经变换的值舍入到最接近的整数。在步骤(3)中，将步骤(1)中执行的变换进行反转以生成最终量化值。这由等式(31)表示，其中qp_{adj}^{t-q}是步骤(2)中生成的量化变换值，并且qp_{adj}^{q}是最终量化的增大/减小的量化参数。

[0206] $$qp_{adj}^{q}=qp_{adj}^{t-q}*s \quad (31)$$

[0207] 例如，如果步骤(2)的输出是3并且s＝4，则这被变换回12；如果步骤(2)的输出是5并且s＝1，则这被变换回5；如果步骤(2)的输出是4并且s＝2，则这被变换回8；并且如果步骤(2)的输出是6并且s＝2，则这被变换回12。这在表2中进行概述。

[0208] 表2

[0209]

qp_{adj}	13	5.4	8.9	11.5
s	4	1	2	2
qp_{adj}^{t}	3.25	5.4	4.45	5.75
qp_{adj}^{t-q}	3	5	4	6
qp_{adj}^{q}	12	5	8	12

[0210] 对于本领域技术人员将显而易见的是，这些是可用于对量化参数进行量化或者模拟其量化的函数的示例，并且可以使用其他函数来对量化参数进行量化或者模拟其量化。然而，为了能够将成本度量cm的导数反向传播到量化参数，定义量化函数q(例如qp_{adj}^{r}、qp_{adj}^{s}、qp_{adj}^{u}、qp_{adj}^{g}、qp_{adj}^{b})，使得成本度量的导数可以根据量化参数来定义。发明人已认识到，如果量化函数q(例如qp_{adj}^{r}、qp_{adj}^{s}、qp_{adj}^{u}、qp_{adj}^{g}、qp_{adj}^{b})相对于所量化的量化参数的导数被定义为一，则机器学习框架可以生成成本函数相对于量化参数的有用梯度。

[0211] 在一些情况下，增大/减小的量化参数的量化(例如舍入)可以由相关量化块来执行。例如，在一些情况下(如下文更详细地描述)，增大/减小的量化参数可以被提供给量化块，并且每个量化块可以被配置成在使用量化参数以对输入值进行量化之前对该量化块的量化参数进行量化(例如舍入)，或者模拟其量化(例如舍入)。

[0212] 在调整量化参数包括(根据梯度)对增大/减小的量化参数进行量化(例如舍入)或者模拟其量化的情况下，通过上文描述的方法中的任何方法，可以维持量化参数的较高精度(例如浮点)型式，并且在块508的后续迭代中，根据梯度增大/减小的就是量化参数的较

高精度型式。在一些情况下，可以维持增大/减小的量化参数的随机量化型式，并且在后续迭代中增大/减小的就是量化参数的随机量化型式。

[0213] 在基于梯度来调整量化参数(例如位宽度b_i和指数exp_i)中的一个或多个量化参数之后，该方法移动到块509，其中可以可选地从NN的模型中移除值集合。在块508中，可以取决于值集合的量化参数(例如位宽度)或者被调整为零的相关联值集合而从NN的模型中移除值集合。这是因为，在某些情况下，相对于保留包括零值的值集合，从NN的模型中移除可以使用零位宽度来量化的值集合(即其中该值集合中的每个值可以被量化为零)可能不会影响NN的模型的输出。也就是说，移除该值集合可以减少NN的推断时间(并且从而提高其效率)，因为移除那些值减少了将在层中执行的乘法运算的数量(即使在那些乘法是乘零的乘法的情况下)。

[0214] 提供了响应于将量化参数调整为零而从NN的模型中移除值集合的六个具体示例。这些示例返回参阅参考块504描述的示例1至6。应理解，这些具体实现仅借助于示例来提供，并且本文中描述的原理可以不同地来实现。

[0215] 这些示例可以参考图9来理解，该图示出了NN的两个相邻层之间的交互。图9示出了层904以及在该层之前的层902。在图9中，层902和904都是卷积层。激活数据906-1、权重数据908-1和偏置912-1被输入到在前层902。激活数据906-2(例如在前层902的输出)、权重数据908-2和偏置912-2被输入到层904。为了易于理解，中间输出数据910-1和910-2被示出为分别用于层902和在前层904，但应理解，所述中间数据不需要由那些层物理地形成，并且可以仅表示方便地描述由那些层在它们的输入与输出之间执行的处理的逻辑值。

[0216] 在示例1和2中，当针对输入到层的激活数据的对应输入通道的经调整的位宽度为零时，输入到在前层的权重数据的输出通道(以及(如果存在)其相关联偏置)可以从NN的模型中移除。例如，在图9中，当针对输入到层904的激活数据的对应输入通道922的经调整的位宽度为零时，输入到在前层902的权重数据的输出通道920可以从NN的模型中移除。使用交叉影线示出了这些通道之间的对应关系(如参考图2B可以理解)。使用参考示例1或2定义的实现度量，可以确定移除输出通道920而不影响NN的模型的输出是“安全的”(相对于保留包括零值的输出通道920)。这是因为与输出通道920卷积以生成中间输出通道924以及随后与偏置926求和的结果生成了方法500确定可以使用零位宽度来量化的输入通道922。因此，可以理解，不需要执行卷积和求和。由此，权重数据908-1的输出通道920、偏置912-1中的偏置926、激活数据906-2的输入通道922和权重数据908-2的输入通道928可以从NN的模型中移除，而不影响NN的模型的输出(相对于保留包括零值的输出通道920)。

[0217] 在示例3中，当针对输入到层的权重数据的对应输入通道的经调整的位宽度为零时，输入到在前层的权重数据的输出通道(以及(如果存在)其相关联偏置)可以从NN的模型中移除。例如，在图9中，当针对输入到层904的权重数据的对应输入通道928的经调整的位宽度为零时，输入到在前层902的权重数据的输出通道920可以从NN的模型中移除。使用交叉影线示出了这些通道之间的对应关系(如参考图2B可以理解)。使用参考示例3定义的实现度量，可以确定移除输出通道920而不影响NN的模型的输出是“安全的”(相对于保留包括零值的输出通道920)。这是因为与输出通道920卷积以生成中间输出通道924以及随后与偏置926求和的结果生成了方法500确定可以使用零位宽度来量化的待与输入通道928进行卷积的输入通道922。因此，可以理解，不需要执行卷积和求和。由此，权重数据908-1的输出通

道920、偏置912-1中的偏置926、激活数据906-2的输入通道922和权重数据908-2的输入通道928可以从NN的模型中移除，而不影响NN的模型的输出(相对于保留包括零值的输出通道920)。

[0218] 当已知在前层没有接收到偏置(图9中未示出)时，示例5可用于移除输入到在前层的权重数据的输出通道。在示例5中，当针对输入到在前层的权重数据的输出通道的经调整的位宽度为零时，该输出通道可以从NN的模型中移除。输入到对其形成实现成本的层的激活数据的对应输入通道以及输入到对其形成实现成本的层的权重数据的对应输入通道也可以从NN的模型中移除，而不影响NN的模型的输出(相对于保留输入到在前层的权重数据的包括零值的输出通道)。

[0219] 应理解，响应于仅确定输入到层的权重数据的输出通道可以使用零位宽度来编码而移除权重数据的该输出通道不一定是“安全的”。如本文中所描述，这是因为权重数据的输出通道可能与偏置相关联。参考图9，即使针对输入到层904的权重数据的输出通道930的经调整的位宽度为零，其相关联偏置932仍可以为非零(例如具有非零位宽度)。在此情况下，如果权重数据908-2的输出通道930将从NN的模型中移除，则不会形成中间输出通道934，这意味着偏置932将不具有用于求和的值。这是使用诸如示例1至3中定义的实现度量的优点，该实现度量考虑了两个相邻层之间的交互(例如凭借取决于表示来自在前层的输出的实现成本的第二贡献)。

[0220] 在示例4中，当输入到层的权重数据的输出通道的经调整的位宽度和其相关联偏置为零时，该输出通道可以从NN的模型中移除。例如，在图9中，当针对输入到在前层902的权重数据的输出通道920的经调整的位宽度和其相关联偏置926为零时，该输出通道920可以从NN的模型中移除。使用交叉影线示出了这些通道和偏置之间的对应关系(如参考图2B可以理解)。由此，权重数据908-1的输出通道920、偏置912-1中的偏置926、激活数据906-2的输入通道922和权重数据908-2的输入通道928可以从NN的模型中移除，而不影响NN的模型的输出(相对于保留包括零值的输出通道920)。替代地或另外，当针对输入到层904的权重数据的输出通道的经调整的位宽度和其相关联偏置为零时，该输出通道可以从NN的模型中移除。

[0221] 在示例6中，当针对输入到在前层的权重数据的输出通道的经调整的位宽度和其相关联偏置的绝对值(例如，如在反向传播期间所调整——如参考图11所描述)为零时，可以从NN的模型中移除该输出通道。例如，在图9中，当针对输入到在前层902的权重数据的输出通道920的经调整的位宽度和其相关联偏置926的经调整的绝对值为零时，该输出通道920可以从NN的模型中移除。使用交叉影线示出了这些通道和偏置之间的对应关系(如参考图2B可以理解)。由此，权重数据908-1的输出通道920、偏置912-1中的偏置926、激活数据906-2的输入通道922和权重数据908-2的输入通道928可以从NN的模型中移除，而不影响NN的模型的输出(相对于保留包括零值的输出通道920)。

[0222] 在块509中移除一个或多个值集合的额外优点是，NN的训练随后将在块502至508的后续迭代中“加速”，如下文进一步详细地描述。这是因为从NN的模型中移除一个或多个值集合减少了NN的模型的实现成本，并且因此提高了其推断速度。因此，可以更快速地执行块502至508的后续迭代。

[0223] 在其中每个层的输出被输入到仅一个其他层(或从NN输出)的许多NN结构中，可以

执行在块509中对一个或多个值集合(例如输出权重通道)的移除,而不需要对NN进行任何进一步的修改。也就是说,在某些NN结构中,层的输出可以输入到多于一个后续层,或者NN的运算可以从多于一个在前层接收输入。这种NN的一个示例在图10A中示出,该图是示出包括剩余层的NN的示意图。在图10A中,层A 1002的输出被输入到层B 1004和求和运算1010。层B 1004的输出被输入到层C 1006。求和运算1010从层A 1002和层C 1006两者接收输入。求和运算1010的输出被输入到层D 1008。

[0224] 在图10A中,求和运算1010可能需要接收具有相同结构的两个输入(例如具有相同数量的数据通道的两个输入激活数据集合)。因此,例如,如果输入到层A 1002的权重数据的输出通道在块509中被移除,从而导致其输出的对应数据通道未被形成(参考图2B可以理解),则有可能需要在求和运算1010之前在层A 1002的输出中提供替换通道。换句话讲,如果输入到层C 1006的权重数据的输出通道在块509中被移除,从而导致其输出的对应数据通道未被形成(参考图2B可以理解),则有可能需要在求和运算1010之前在层C 1006的输出中提供替换通道。参考图10B来描述将替换通道插入到这种NN中的层的输出数据中的方法1020。在示例中,图10B的方法1020可用于将替换通道插入到深度神经网络(DNN)-其为一种类型的NN-中的层的输出数据中。

[0225] 在块1022中,对于层的输出数据的标识通道,对输入到该层的激活数据进行运算,使得层的输出数据不包括标识通道。例如,这可以通过不包括负责形成标识通道的权重数据的输出通道,使得层的输出数据不包括标识通道来实现。如本文中所描述,可以在NN的训练阶段中标识出,负责形成标识通道的权重数据的输出通道(以及可选地,对应偏置)在位宽度为零的情况下是可量化的(例如该输出通道相对于保留输入到在前层的权重数据的包括零值的输出通道可以从NN的模型中移除,而不影响NN的模型的输出——如参考块508和509所描述)。换句话说,在NN的训练阶段中确定了权重数据(以及可选的对应偏置)的输出通道可以使用零位宽度进行量化后,输出数据的标识通道可以被标识为权重数据(以及可选的对应偏置)的该输出通道负责形成的输出数据的通道。在图10A中,此步骤的效果可以是层A 1002的输出数据不包括标识通道。层A 1002的所述输出(即不包括标识通道)可以由层B 1004来运算。在图10A中,此步骤的效果可以是层C 1006的输出数据不包括标识通道。

[0226] 在块1024中,在被配置成对层的输出数据进行运算的NN的运算(例如求和运算1010)之前,可以将替换通道插入到层的输出数据中,以代替(例如替代)标识通道。例如,替换通道可以是包括多个零值的通道。标识通道可以是数据值阵列,并且替换通道可以是与该数据值阵列具有相同维数的零(例如零值)阵列。NN的所述运算(例如求和运算1010)随后可以根据替换通道来执行。应理解,如果标识通道包括多个零值,则相对于通过保留包括多个零值的标识通道来执行NN的运算,插入如本文中所描述的包括多个零值的替换通道不会改变NN的运算的结果。

[0227] 如果已包括标识通道,则可以根据指示层的输出数据的结构的信息来插入替换通道。也就是说,所述信息可以指示在该输出数据已形成为包括标识通道的情况下,层的输出数据的结构会是什么。换句话说,如果已包括标识通道,则可以根据指示层的输出数据的结构的信息来插入替换通道。该信息可以在NN的训练阶段中生成,该信息指示包括标识通道的层的输出数据的结构。例如,所述信息可以包括位掩码。位掩码的每个位可以表示数据通道,第一位值(例如1或0)指示包括在输出数据中的数据通道,第二位值(例如0或1)指示不

包括在输出数据中的数据通道。替换通道可以被插入到位掩码的第二位值所指示的层的输出数据中。例如，如果位掩码包括一连串的位值…1、0、1…，则可以在由位值0指示的位置，在由位值1表示的输出数据中包括的两个数据通道之间插入替换通道。应理解，本文中描述的插入替换通道的方法可用于插入多个替换通道以代替多个相应标识通道。例如，位掩码可以包括多个第二位值，该多个第二位值各自指示不包括在输出数据中的数据通道，使得多个替换通道可以被插入到由那些第二位值指示的层的输出数据中。

[0228] 插入替换通道的此方法可以在NN的训练阶段期间执行(例如当在较早迭代中确定权重数据的输出通道在位宽度为零的情况下是可量化的之后执行块502至509的后续迭代时，如下文进一步详细地描述)以及/或者在随后实现NN以在使用阶段中处理数据时执行(例如在块514中，也在下文进一步详细地描述)。

[0229] 一旦已在块508中基于梯度而调整了量化参数中的一个或多个量化参数(并且可选地在块509中移除了一个或多个值集合)，方法500便可以结束，或者方法500可以前进到块510，其中可以重复块502至509。

[0230] 在块510处，关于是否将重复块502至509作出确定。在一些情况下，关于是否将重复块502至509的确定是基于是否已完成块502至509的预定次数的迭代，或者是否已过去预定量的训练时间。预定次数的迭代或预定量的训练可能已在经验上被确定为足以产生良好结果。在其他情况下，关于是否将重复块502至509的确定可以基于成本度量是否已收敛。可以使用任何合适的标准来确定成本度量何时收敛。例如，在一些情况下，如果成本度量在预定次数的迭代内没有显著改变(例如大于预定阈值)，则可以确定成本度量已收敛。

[0231] 如果确定将不重复块502至509，则方法500可以结束，或者方法500可以前进到块512。然而，如果确定将重复块502至509，则方法500返回到块502，其中使用如在块508中调整的量化参数重复块502至509(并且可选地不包括在块509中移除的值集合)。例如，如果在第一迭代中，值集合由量化块变换为由尾数位宽度6和指数4定义的定点数形式，并且尾数位宽度被调整为位宽度5，并且指数未被调整，则在下一迭代中，该值集合将由量化块变换为由位宽度5和指数4定义的定点数格式。

[0232] 在块512处，输出如在块508中调整的量化参数以及可选地指示在块509中移除的值集合的信息，以用于将硬件逻辑配置成实现NN。在一些情况下，输出量化参数的浮点型式。在其他情况下，输出的就是可由硬件逻辑使用的量化参数的型式(即量化参数在其已被量化为整数或整数集合之后的浮点型式)。量化参数可以以任何合适的方式输出。一旦已输出如在块508中调整的量化参数，方法500便可以结束，或者方法500可以前进到块514。

[0233] 在块514处，能够实现NN的硬件逻辑被配置成使用在块512中输出的量化参数来实现NN。在块512中输出的量化参数呈浮点数格式的情况下，在量化参数被用于将硬件逻辑配置成实现NN之前，量化参数可以被量化为整数或整数集合。将硬件逻辑配置成实现NN一般可以包括将硬件逻辑配置成根据该层来处理针对NN的每个层的输入，以及将该层的输出提供给后续层，或者提供该输出作为NN的输出。例如，如果NN包括第一卷积层和第二归一化层，则将硬件逻辑配置成实现这种NN包括将硬件逻辑配置成接收针对NN的输入，以及根据卷积层的权重数据将输入作为输入激活数据进行处理，根据归一化层来处理卷积层的输出，并且随后输出归一化层的输出作为NN的输出。将硬件逻辑配置成使用在块512中输出的量化参数来实现NN可以包括将硬件逻辑配置成根据该层的量化参数(即根据由量化参数定

义的定点数格式)来接收和处理针对每个层的输入。例如,如果量化参数指示由指数4和位宽度6定义的定点数格式将被用于NN的层的输入数据值,则用于实现NN的硬件逻辑可以被配置成基于该层的输入数据值呈由指数4和位宽度6定义的定点数格式来解释该层的输入数据值。

[0234] 当在块514处实现NN时,在块509处从NN的模型中移除的值集合可以不包括在NN的运行时实现中。例如,在输入到层的权重数据的输出通道在块509处被移除的情况下,该输出通道的权重值可以不被写入到存储器以供NN的运行时实现使用,并且/或者实现NN的运行时实现的硬件可以不被配置成使用那些权重值来执行乘法。

[0235] 在图5的方法500中,计算完整成本度量(例如根据等式(3)),并且将成本度量的导数反向传播到量化参数以计算针对每个量化参数的梯度。针对特定量化参数的梯度随后被用于调整量化参数。然而,在其他示例中,计算成本度量可以包括计算误差度量和实现度量,以及确定针对每个量化参数的每个度量的单独梯度。换句话说,生成误差度量相对于每个量化参数的梯度,并且生成实现度量相对于每个量化参数的梯度。以与将成本度量的导数反向传播到量化参数的方式相同的方式,误差度量相对于量化参数的梯度可以通过将误差度量的导数反向传播到量化参数来生成。实现度量相对于量化参数的梯度可以通过反向传播来生成,或者可以直接从实现度量中生成。针对每个量化参数的最终梯度可以以相同方式从两个梯度中生成,即对应成本度量被组合以形成成本度量。例如,最终梯度可以生成为两个梯度的加权和。通过改变与两个梯度相关联的权重,可以在实现成本与误差之间找到平衡。随后可以以与上文描述的方式相同的方式根据最终梯度来调整量化参数。

[0236] 标识量化参数和权重

[0237] 尽管图5的方法500已被描述为用于标识NN的量化参数,但在其他示例中,NN的权重值(例如权重)和可选的偏置可以与量化参数同时被标识。在这些情况下,成本度量的导数也可以被反向传播到权重(以及可选的偏置),以生成成本度量相对于权重(以及可选的偏置)的梯度,并且权重(以及可选的偏置)可以基于对应梯度以与量化参数类似的方式进行调整。

[0238] 现在参考图11,该图示出了标识NN的量化参数和权重(以及可选的偏置)的方法1100。在示例中,图11的方法1100可用于经由反向传播来标识深度神经网络(DNN)——其为NN的一种类型——的量化参数和权重(以及可选的偏置)。方法1100可用于重新训练网络以考虑NN的值的量化(例如在初始训练项目之后更新权重,诸如对NN的浮点模型执行的初始训练项目),或者可用于执行网络的初始训练(例如根据未训练的权重集合来训练网络)。方法1100包括图5的方法500的块502至512,但还包括块1102和1104(以及可选的块1106和1108)。块502至512以与上文描述的方式相同的方式来操作。当方法1100被用于重新训练NN时,用于NN的量化模型中的初始权重集合可以是经训练的权重集合。然而,在方法1100被用于训练NN的情况下,用于NN的模型中的初始权重集合可以是随机权重集合或者被设计成用于训练NN的另一权重集合。

[0239] 在块1102处,在已确定NN的量化模型响应于训练数据的输出(块502)并且已根据NN的量化模型的输出和量化参数确定成本度量(块504)之后,将成本度量的导数反向传播到一个或多个权重(以及可选的偏置),以便生成成本度量相对于那些权重(以及可选的偏置)中的每个权重的梯度。成本度量相对于权重的梯度在本文中称为针对权重的梯度。如同

针对量化参数的梯度那样，针对权重的正梯度指示可通过减小该权重来减小成本度量，并且针对权重的负梯度指示可通过增大该权重来减小成本度量。一旦已生成针对一个或多个权重(以及可选的偏置)的梯度，该方法便前进到块1104。

[0240] 在块1104处，基于针对权重(以及可选的偏置)的梯度来调整权重(以及可选的偏置)中的一个或多个权重。权重(以及可选的偏置)可以以与量化参数类似的方式进行调整。例如，如上文所描述，针对权重的梯度的符号指示成本度量是否将通过增大或减小权重而减小。具体地，如果针对权重的梯度是正的，则权重的减小将减小成本度量；并且如果针对权重的梯度是负的，则量化参数的增大将减小成本度量。因此，调整权重可以包括根据梯度的符号来增大或减小权重以便增大或减小成本度量(取决于是期望增大还是减小成本度量)。例如，如果较低成本度量是期望的，并且针对权重的梯度是负的，则权重可以被增大以试图减小成本度量。类似地，如果较低成本度量是期望的，并且针对权重的梯度是正的，则权重可以被减小以试图减小成本度量。

[0241] 在一些情况下，权重增大或减小的量可以基于针对该权重的梯度的量值。特别地，在一些情况下，权重可以增大或减小达针对该权重的梯度的量值。例如，如果梯度的量值是0.6，则权重可以增大或减小0.6。在其他情况下，权重可以增大或减小达针对该权重的梯度的量值的因子。特别地，在一些情况下，通过利用所谓的学习速率来调整权重，权重可以更快地收敛。

[0242] 一旦已基于对应梯度来调整权重(以及可选的偏置)，方法1100便可以结束，或者方法1100可以前进到块509，其中可以可选地从NN的模型中移除一个或多个值集合。此后，可以重复块502至508和1102至1104。类似于块512和514，方法1100还可以包括输出经调整的权重(以及可选的偏置)(在1106处)以及/或者将硬件配置成使用经调整的权重(以及可选的偏置)来实现NN(在1108处)。

[0243] 尽管在图11的方法1100中，每次迭代都调整权重(以及可选的偏置)和量化参数，但在其他示例中，在每次迭代中，可以选择权重(以及可选的偏置)和量化参数中的一者或两者进行调整。例如，量化参数可以被调整达预定次数的迭代，并且随后权重(以及可选的偏置)可以被调整达预定次数的迭代。在其他情况下，可以在交替迭代中调整权重(以及可选的偏置)和量化参数。例如，权重(以及可选的偏置)调整可以在奇数次迭代中执行，量化参数调整可以在偶数次迭代中执行。这将允许在量化参数被舍入(或其舍入被模拟)时调整权重(以及可选的偏置)，并且在权重(以及可选的偏置)被舍入时调整量化参数。

[0244] 量化块

[0245] 现在将描述NN的量化模型的量化块的示例具体实现。如上文所描述，每个量化块被配置成将输入到NN的层的一个或多个值集合变换为由一个或多个量化参数定义的定点数格式。在这些示例中，每个定点数格式由尾数位长度b和指数exp定义，其中指数exp是由以定点数格式表示的值集合共享的整数，使得呈定点数格式的输入数据值集合的大小是基于尾数位长度b。

[0246] 为了能够将成本度量的导数反向传播到量化参数，不仅定义了由每个量化块执行的量化函数，还定义了其导数。在实践中，等式的导数由诸如但不限于TensorFlow™或PyTorch™的机器学习框架自动定义。

[0247] 将值x量化为定点数格式的过程可以被描述为包括两个步骤——(i)将值x阈值化

为可由定点数格式表示的数字范围(例如图12的线1202针对指数-1和位宽度3);以及(ii)通过将阈值化值x舍入到最接近的2的exp^{th}次幂来选择呈定点数格式的可表示数字以表示值x(例如图12的线1204针对指数-1和位宽度3)。

[0248] 值x到由尾数位长度b和指数exp定义的定点数格式的量化运算的阈值化步骤——即,将值x阈值化为可由定点数格式表示的范围——可以通过等式(32)来实现,其中clamp(x,low,high)如等式(33)中所定义,并且low是呈由b和exp定义的定点数格式的最小或最低可表示数字(例如low=$-2^{b-exp-1}$),并且high是呈由b和exp定义的定点数格式的最大或最高可表示数字(例如high=$2^{b+exp-1}-2^{exp}$):

[0249] $thresh(x,b,exp)=clamp(x,-2^{exp+b-1},2^{exp+b-1}-2^{exp})$ (32)

[0250] $clamp(x,low,high)=min(max(x,low),high)$ (33)

[0251] 为了能够通过阈值化运算来执行反向传播,定义阈值化运算的导数。等式(32)中定义的阈值化函数相对于x的导数对于落入可表示范围内的值是1,否则是0。然而,在一些情况下,更有用的导数对于落入量化区间内的所有值是1,否则是0。这可以通过使用等式(34)中阐明的阈值化函数而不是等式(32)中阐明的阈值化函数来实现:

[0252] $thresh(x,b,exp)=clamp(x,-2^{exp+b-1}-2^{exp-1},2^{exp+b-1}-2^{exp-1})$ (34)

[0253] 量化运算的舍入步骤——即,将值舍入到最接近的2的exp^{th}次幂——可以通过等式(35A)或等式(35B)中的任一者来实现,其中⌊ ⌋是RTNI(向负无穷舍入)函数(也称为地板函数)。

[0254] $$round(x,exp)=2^{exp}\left\lfloor 2^{-exp}x+\frac{-1^{sign(x)}}{2}\right\rfloor \quad (35A)$$

[0255] $$round(x,exp)=2^{exp}\left\lfloor 2^{-exp}x-\frac{-1^{sign(x)}}{2}\right\rfloor \quad (35B)$$

[0256] 等式(35A)或等式(35B)中的任一者中定义的舍入函数相对于x的导数在标识NN参数(例如权重和/或量化参数)时可能并不是有用的,因为该导数几乎在任何地方都是零,因此该导数可以被设置为1。

[0257] 因此,值x到由位宽度b和指数exp定义的定点数格式的总量化quant(x,b,exp)可以使用如等式(36)中所示的阈值化等式(等式(32)或等式(34))与舍入等式(等式(35A)或等式(35B))的组合来实现:

[0258] $quant(x,b,exp)=round(thresh(x,b,exp),exp)$ (36)

[0259] 在量化块没有被配置成在使用量化参数来量化输入值之前量化(例如舍入)所接收到的量化参数的情况下,组合公式可以如等式(37A)中所示来书写。在训练阶段期间(例如,如本文中参考图5的块502至510所描述),量化块不被配置成量化(例如舍入)所接收到的量化参数,使得量化块在该训练阶段期间用于量化输入值的量化参数不被约束为具有整数值可能是有利的——这可以实现那些量化参数的较高分辨率(例如更高精度)训练。

[0260] $quant(x,b,exp)=2^{exp}round(min(max(2^{-exp}x,-2^{(b-1)}),2^{(b-1)}-1))$ (37A)

[0261] 在替代示例中,在量化块没有被配置成在使用量化参数来量化输入值之前量化(例如舍入)所接收到的量化参数的情况下,组合公式可以如等式(37B)中所示来书写。等式(37A)与等式(37B)之间的主要差异是引入了作为缩放因子(例如移位参数)的α。

[0262] $quant(x,b,exp,\alpha)=2^{exp}round(min(max(2^{-exp}x,(\alpha-1)2^{(b-1)}),(\alpha+$

[0263] $1)2^{(b-1)}-1))$ (37B)

[0264] 在量化块被配置成接收增大/减小的量化参数并且在使用量化参数来量化输入值之前量化(例如舍入)所接收到的量化参数的情况下,组合公式可以如等式(37C)中所示来书写,其中q是用于对量化参数进行量化或者模拟其量化的舍入函数或者量化函数。上文关于块508描述了用于对量化参数进行量化或者用于模拟其量化的示例舍入函数。换句话说,量化函数q可以实现(i)上文描述的舍入方法以舍入到最接近的整数或集合中最接近的整数,或者(ii)上文描述的模拟舍入到最接近的整数或集合中的整数的方法中的任何方法(例如随机量化方法、均匀量化方法、梯度平均量化方法或双峰量化方法中的一者)。如上文所描述,为了能够将成本度量cm的导数反向传播到量化参数,定义量化函数q,使得成本度量的导数可以根据量化参数来定义。在训练阶段期间(例如,如本文中参考图5的块502至510所描述),量化块被配置成在使用量化参数来量化输入值之前量化(例如舍入)所接收到的量化参数可能是有利的,因为这可以使得训练能够考虑当NN随后在硬件中实现时将发生的量化参数的量化(例如舍入)——尤其是在量化块被配置成使用那些量化参数来量化输入激活值的情况下。

[0265] $quant(x,b,exp)=2^{q(exp)}round(clamp(2^{-q(exp)}x,-2^{q(b-1)},2^{q(b-1)}-1))$

[0266] (37C)

[0267] 发明人已认识到,如果量化函数q相对于其正在量化的量化参数的导数被定义为一,则机器学习框架可以生成成本函数相对于量化参数的有用梯度(例如可用于调整量化参数的梯度)。例如,测试已表明,如果量化函数q相对于其正在量化的量化参数的导数被设置为一,则机器学习框架可以生成:(i)主量化函数quant相对于量化参数b的导数$d_b(x)$,如等式(38)中所示,其中low是呈由b和exp定义的定点数格式的最小或最低可表示数字,并且high是呈由b和exp定义的定点数格式的最大或最高可表示数字;以及(ii)主量化函数quant相对于量化参数exp的导数$d_{exp}(x)$,如等式(39)中所示。

[0268] $$d_b(x)=\begin{cases} x<low,\ \log(2)\,(low) \\ x>high,\ \log(2)\,(-low) \\ 否则,\ 0 \end{cases} \quad (38)$$

[0269] $$d_{exp}(x)=\begin{cases} x<low,\ 2^{exp}\log(2)\,round(-2^{b-1}) \\ x>high,\ 2^{exp}\log(2)\,round(2^{b-1}-1) \\ 否则,\ \log(2)\,(2^{exp}round(2^{-exp}x)-x) \end{cases} \quad (39)$$

[0270] 可以看出,机器学习框架可以针对由量化块量化的每个输入值计算针对该量化块的每个量化参数(例如b、exp)的成本函数的导数。机器学习框架随后可以基于每个量化参数的各个导数来计算成本函数针对每个量化参数(例如b、exp)的最终导数。例如,在一些情况下,机器学习框架可以通过将量化块的每个量化参数的各个导数相加或求和来计算成本函数针对该量化参数的最终导数。

[0271] 在使用Q8A定点数格式的可变位长度变型来表示针对NN的层的输入值并且零点z为0的情况下,由量化块执行的量化函数可以由等式(40)表示,其中b、exp和α是可训练量化参数:

[0272] $quant(x,b,exp,\alpha)=$

[0273] $2^{exp}round(clamp(2^{-exp}x,(\alpha-1)2^{q(b-1)},(\alpha+1)2^{q(b-1)}-1))$ (40)

[0274] 等式(40)与等式(37C)之间的主要差异是引入了作为缩放因子的α以及exp没有被量化的事实。如等式(1)中所示,Q8A格式的可变位长度变型的量化参数可以从如等式(41)、(42)和(43)中所示的经训练的量化参数exp、b和α中生成:

[0275] $r_{min}=2^{exp}RND(2^{RND(b)-1}(\alpha-1))$ (41)

[0276] $r_{max}=2^{exp}RND(2^{RND(b)-1}(\alpha+1)-1)$ (42)

[0277] $z=0$ (43)

[0278] 在使用Q8A定点数格式的可变位长度变型来表示针对NN的层的输入值的情况下,其中零点z可以不为零,由量化块执行的量化函数可以由等式(44)表示。

[0279] $quant(x,b,exp,\alpha)=2^{exp}(round(clamp(2^{-exp}x-2^{q(b-1)}\alpha,-2^{q(b-1)},(\alpha+1)2^{q(b-1)}-1))+2^{q(b-1)}\alpha)$ (44)

[0280] 针对等式(40)和(44),虽然Q8A定点数格式的位长度变型的量化参数是r_{min}、r_{max}、z和b,但测试已表明,训练b、exp和α以及根据其计算r_{min}、r_{max}和z已展现出更好地训练。

[0281] 在一些情况下,不是量化块将输入到该量化块的值量化为由一个或多个量化参数定义的输出定点数格式(例如根据等式(36)、(37A)、(37B)、(37C)、(40)或(44)),而是量化块可以被配置成仅模拟输入值的量化所表示的变换。应理解,在量化块在本文中被描述为将值集合变换为由一个或多个量化参数定义的定点数格式的情况下,所述变换可以涉及根据一个或多个量化参数来量化该值集合,或者可以涉及通过一个或多个量化参数来模拟该值集合的量化。

[0282] 例如,在一些情况下,不是量化块被配置成将权重或输入/激活值阈值化为定点数格式的可表示范围,并且随后将阈值化激活/权重/偏置值舍入到呈定点数格式的最接近的可表示数字,而是可以通过将权重/激活值阈值化,并且将-a与+a之间的随机值u添加到阈值化激活/权重/偏置值,并且随后进行舍入来模拟量化,其中a是定点数格式的可表示数字之间的距离的一半(即$\frac{2^{exp}}{2}$)。例如,如果定点数格式的指数exp为0,则在舍入激活/权重/偏置值之前,将-0.5与+0.5之间的随机值添加到阈值化激活/权重/偏置值,因为可表示数字之间的距离为1。类似地,如果定点数格式的指数为1,则将-1与+1之间的随机值添加到阈值化激活/权重/偏置,因为可表示数字之间的距离为2。以此方式,阈值化激活/权重/偏置值以与到该可表示数字的距离成比例的概率被向上或向下舍入到可表示数字。例如,在指数exp为0的情况下,阈值化激活/权重/偏置值4.2将以80%的概率被舍入到4,并且以20%的概率被舍入到5。类似地,7.9将以10%的概率被舍入到7,并且以90%的概率被舍入到8。在其他示例中,随机化和阈值化的排序可以反转。例如,不是将激活/权重/偏置值阈值化、将随机值添加到阈值激活/权重/偏置值并且随后舍入,而是可以将随机值添加到激活/权重/偏置值以生成随机化权重,随机化激活/权重/偏置值可以被阈值化,随后被舍入。

[0283] 在其他情况下,不是量化块被配置成将阈值化激活/权重/偏置值舍入到最接近的可表示数字,而是量化块可以被配置成通过将-a与+a之间的随机值u添加到阈值化激活/权重/偏置值来模拟激活/权重/偏置值的量化,其中如上文所描述,a是呈定点数格式的可表示数字之间距离的一半。通过仅将这种随机值添加到阈值化激活/权重/偏置值,阈值化激活/权重/偏置值以与对阈值化激活/权重/偏置值进行舍入类似的方式失真。在其他示例中,随机化和阈值化的排序可以反转。例如,不是将激活/权重/偏置值阈值化,并且将随机

值添加到阈值权重，而是可以将随机值添加到激活/权重/偏置值以生成随机化激活/权重/偏置值，并且随机化激活/权重/偏置值可以被阈值化。

[0284] 在又其他情况下，不是量化块将阈值化激活/权重/偏置值舍入到最接近的可表示数字，而是量化块可以被配置成通过对阈值化激活/权重/偏置值执行梯度平均量化来模拟量化。对阈值化激活/权重/偏置值执行梯度平均量化可以包括取阈值化激活/权重/偏置值的底，并且随后添加0与c之间的随机值h，其中c是呈定点数格式的可表示数字之间的距离。例如，如果定点数格式的指数exp是0，则在取阈值化激活/权重/偏置值的底之后，将0与1之间的随机值添加到该底上，因为呈定点数格式的可表示数字之间的距离为1。类似地，如果定点数的指数exp是1，则在取阈值化激活/权重/偏置值的底之后，将0与2之间的随机值添加到该底上，因为可表示数字之间的距离为2。

[0285] 在又其他情况下，不是量化块将阈值化激活/权重/偏置值舍入到最接近的可表示数字，而是量化块可以被配置成通过对阈值化激活/权重/偏置值执行双峰量化来模拟量化，如上文所描述，该双峰量化是舍入到最接近的量化与梯度平均量化的组合。具体地，在双峰量化中，以概率p对阈值化激活/权重/偏置值执行梯度平均量化，否则对阈值化激活/权重/偏置值执行舍入量化，其中p是到最接近的可表示值的距离的两倍除以呈定点数格式的可表示数字之间的距离。在其他示例中，双峰量化和阈值化的排序可以反转。例如，不是对激活/权重/偏置值进行阈值化，并且对阈值化激活/权重/偏置执行双峰量化，而是可以对激活/权重/偏置值执行双峰量化，并且可以对双峰量化的结果执行阈值化。

[0286] 换句话说，等式(36)、(37A)、(37B)、(37C)、(40)和(44)中的任一者中的舍入函数(round)可以使用实现上文描述的任何模拟舍入方法(例如随机量化方法、均匀噪声量化方法、梯度平均量化方法或双峰量化方法)的函数来替换。

[0287] 示例NN加速器

[0288] 现在参考图13，该图示出了示例硬件逻辑，该示例硬件逻辑可以被配置成使用根据图5的方法500或图11的方法1100标识的量化参数来实现NN。具体地，图13示出了示例NN加速器1300。在示例中，NN加速器1300可以被配置成使用根据图5的方法500或图11的方法1100标识的量化参数来实现深度神经网络(DNN)——其为一种类型的NN。

[0289] 图13的NN加速器1300被配置成通过一系列硬件传递(其也可以称为处理传递)来计算NN的输出，其中在每次传递期间，NN加速器接收NN的层的输入数据的至少一部分，并且根据该层(以及可选地根据一个或多个后续层)来处理所接收到的输入数据以产生处理后的数据。处理后的数据要么输出到存储器用作随后的硬件传递的输入数据，要么输出为NN的输出。NN加速器在单个硬件传递期间可处理的层数可基于数据的大小、NN加速器的配置，以及层的顺序。例如，在NN加速器包括执行每个可能的层类型的硬件逻辑的情况下，包括第一卷积层、第一激活层、第二卷积层、第二激活层和池化层的NN可能能够接收初始NN输入数据，并且根据第一硬件传递中的第一卷积层和第一激活层来处理该输入数据，然后将激活层的输出输出到存储器中，接着在第二硬件传递中从存储器中接收该数据作为输入，并且根据第二卷积层、第二激活层和池化层处理该数据，以产生NN的输出数据。

[0290] 图13的示例NN加速器1300包括输入模块1301、卷积引擎1302、累积缓冲器1304、逐元素运算模块1306、激活模块1308、归一化模块1310、池化模块1312、输出交织模块1314以及输出模块1315。每个模块或引擎实现或处理一种或多种类型的层的全部或一部分。具体

地，卷积引擎1302和累积缓冲器1304一起实现或处理卷积层或全连接层。激活模块1308处理或实现激活层。归一化模块1310处理或实现归一化层。池化模块1312实现池化层，并且输出交织模块1314处理或实现交织层。

[0291] 输入模块1301被配置成接收待处理的输入数据，并且将其提供给下游模块以进行处理。

[0292] 卷积引擎1302被配置成使用所接收到的与特定卷积层相关联的输入权重数据对所接收到的输入激活数据执行卷积运算。如图13中所示出，针对NN的每个卷积层(其可以通过图11的方法1100来生成)的权重可以存储在系数缓冲器1316中，并且当卷积引擎1302正在处理特定卷积层时，针对该特定卷积层的权重可以被提供给卷积引擎1302。在NN加速器支持可变权重格式的情况下，卷积引擎1302可以被配置成接收指示正在处理的当前卷积层的权重的一种或多种格式的信息，以允许卷积引擎正确地解释和处理所接收到的权重。

[0293] 卷积引擎1302可包括多个乘法器(例如，128)和多个加法器，所述多个加法器将乘法器的结果相加以产生单个和。尽管在图13中示出了单个卷积引擎1302，但在其他示例中，可存在多个(例如，8个)卷积引擎，使得可以同时处理多个窗口。卷积引擎1302的输出被馈送到累积缓冲器1304。

[0294] 累积缓冲器1304被配置成接收卷积引擎的输出，并且向累积缓冲器1304的当前内容添加该输出。以这种方式，累积缓冲器1304在卷积引擎1302的若干硬件传递上累积卷积引擎1302的结果。尽管在图13中示出了单个累积缓冲器1304，但在其他示例中，可存在多个(例如，8个，每个卷积引擎一个)累积缓冲器。累积缓冲器1304将累积结果输出到逐元素运算模块1306，该逐元素运算模块根据在当前硬件传递期间是否要处理逐元素层，可能或可能不对累积结果进行运算。

[0295] 逐元素运算模块1306被配置成接收当前硬件传递的输入数据(例如在当前硬件传递中未处理卷积层时)或者来自累积缓冲器1304的累积结果(例如在当前硬件传递中处理卷积层时)。逐元素运算模块1306可根据在当前硬件传递中是否处理逐元素层以及/或者根据是否要在逐元素层之前对激活层进行处理，处理所接收的输入数据或将所接收的输入数据传递到另一个模块(例如，激活模块1308和/或归一化模块1310)。当逐元素运算模块1306被配置成处理所接收到的输入数据时，逐元素运算模块1306对所接收到的数据执行逐元素运算(可选地与另一数据集合(其可以从外部存储器中获得)一起)。逐元素运算模块1306可以被配置成执行任何合适的逐元素运算，诸如但不限于相加、相乘、最大值和最小值。随后，根据是否将在逐元素层之后处理激活层，将逐元素运算的结果提供给激活模块1308或归一化模块1310。

[0296] 激活模块1308被配置成接收以下各项中的一者作为输入数据：(例如当在当前硬件传递中未处理卷积层时)(经由逐元素运算模块1306)针对硬件传递的原始输入；(例如当在当前硬件传递中处理卷积层并且在当前硬件传递中未处理逐元素层或者在当前硬件传递中处理逐元素层但在激活层之后时)(经由逐元素运算模块1306)累积数据。激活模块1308被配置成将激活函数应用于输入数据，并且将输出数据提供回逐元素运算模块1306，其中该输出数据被直接或者在逐元素运算模块1306对其进行处理之后转发到归一化模块1310。在一些情况下，应用于由激活模块1308接收的数据的激活函数可在每个激活层变化。在这些情况下，可以(例如，在存储器中)存储指定要应用于每个激活层的激活函数的一个

或多个属性的信息，并且可以将在特定硬件传递中处理的激活层的相关信息在该硬件传递期间提供给激活模块1308。

[0297] 在一些情况下，激活模块1308可以被配置成在查找表的条目中存储表示激活函数的数据。在这些情况下，输入数据可用于查找查找表中的一个或多个条目以及表示激活函数的输出的输出值。例如，激活模块1308可以被配置成通过在从查找表中读取的两个或更多个条目之间进行内插来计算输出值。

[0298] 在一些示例中，激活模块1308可以被配置成通过实现ReLU函数而作为整流线性单元(ReLU)来操作。在ReLU函数中，输出元素$y_{i,j,k}$通过标识等式(45)中列出的最大值来计算，其中对于小于0的x值，y＝0：

[0299] $$y_{i,j,k}=f\ (x_{i,j,k})=\max\{0,x_{i,j,k}\} \quad (45)$$

[0300] 在其他示例中，激活模块1308可以被配置成通过实现PReLU函数而作为参数整流线性单元(PReLU)来操作。PReLU函数执行与ReLU函数类似的运算。具体地，在$w_1, w_2, b_1, b_2 \in \mathbb{R}$为常数的情况下，PReLU被配置成生成输出元素$y_{i,j,k}$，如等式(46)中所示：

[0301] $$y_{i,j,k}=f\ (x_{i,j,k};w_1,w_2,b_1,b_2)=\max\{(w_1*x_{i,j,k}+b_1),(w_2*x_{i,j,k}+b_2)\} \quad (46)$$

[0302] 归一化模块1310被配置成接收以下各项中的一者作为输入数据：(例如当在当前硬件传递中未处理卷积层并且在当前硬件传递中未处理逐层元素层和激活层时)(经由逐元素运算模块1306)针对硬件传递的原始输入数据；(例如当在当前硬件传递中处理卷积层并且在当前硬件传递中未处理逐元素层和激活层时)(经由逐元素运算模块1306)累积输出；以及逐元素运算模块和/或激活模块的输出数据。归一化模块1310随后对所接收到的输入数据执行归一化函数以产生归一化数据。在一些情况下，归一化模块1310可以被配置成执行局部响应归一化(LRN)函数和/或局部对比度归一化(LCN)函数。然而，对于本领域技术人员将显而易见的是，这些仅是示例，并且归一化模块1310可以被配置成实现任何合适的一个或多个归一化函数。不同的归一化层可以被配置成应用不同的归一化函数。

[0303] 池化模块1312可接收来自归一化模块1310的归一化数据，或者可经由归一化模块1310接收到归一化模块1310的输入数据。在一些情况下，可以经由XBar(或“交叉开关(crossbar)”)1318在归一化模块1310与池化模块1312之间传输数据。在本文中，术语“XBar”用于指包含路由逻辑的简单硬件模块，该路由逻辑以动态方式将多个模块连接在一起。在此示例中，XBar可以取决于在当前硬件传递中将处理哪些层而动态地连接归一化模块1310、池化模块1312和/或输出交织模块1314。因此，XBar可以在每次传递中接收指示将连接哪些模块1310、1312、1314的信息。

[0304] 池化模块1312被配置成对所接收到的数据执行池化函数，诸如但不限于max函数或mean函数，以产生池化数据。池化层的目的是减小表示的空间大小，以减少网络中参数和计算的数量，并且因此也控制过度拟合。在一些示例中，在每个池化层定义的滑动窗口上执行池化运算。

[0305] 输出交织模块1314可接收来自归一化模块1310的归一化数据、(经由归一化模块1310)对归一化函数的输入数据，或者来自池化模块1312的池化数据。在一些情况下，可以经由XBar 1318在归一化模块1310、池化模块1312和输出交织模块1314之间传输数据。输出交织模块1314被配置成执行重新排列运算以产生处于预定顺序的数据。这可包括对所接收的数据进行排序和/或转置。由层中的最后一层生成的数据被提供给输出模块1315，其中该

数据被转换为针对当前硬件传递的期望输出格式。

[0306] 归一化模块1310、池化模块1312和输出交织模块1314可各自访问共享缓冲器1320，这些模块1310、1312和1314可使用该共享缓冲器向其中写入数据并且从中检索数据。例如，这些模块1310、1312、1314可使用共享缓冲器1320来重新排列所接收的数据或所生成的数据的顺序。例如，这些模块1310、1312、1314中的一个或多个模块可以被配置成将数据写入共享缓冲器1320，并且以不同的顺序读出相同的数据。在一些情况下，尽管归一化模块1310、池化模块1312和输出交织模块1314中的每一者都可以访问共享缓冲器1320，但是归一化模块1310、池化模块1312和输出交织模块1314中的每一者可被分配共享缓冲器1320的仅它们自身可访问的一部分。在这些情况下，归一化模块1310、池化模块1312和输出交织模块1314中的每一者可能仅能够从共享缓冲器1320读取它们已经写入共享缓冲器1320中的数据。

[0307] 在任何硬件传递期间使用的或活动的NN加速器1300的模块是基于在该硬件传递期间处理的层。特别地，仅与在当前硬件传递期间处理的层相关的模块或部件被使用或者是活动的。如上文所描述，基于NN中的层的顺序以及可选地一个或多个其他因素(诸如数据的大小)(通常预先通过例如软件工具)来确定在特定硬件传递期间处理的层。例如，在一些情况下，除非可以在不将数据写入到层之间的存储器的情况下处理多个层，否则NN加速器可以被配置成每硬件传递执行单个层的处理。例如，如果在第一卷积层之后紧随第二卷积层，则卷积层中的每个卷积层将必须在单独的硬件传递中执行，因为来自第一硬件卷积的输出数据需要首先被写出到存储器，随后才可以被用作针对第二硬件卷积的输入。在这些硬件传递中的每个硬件传递中，仅与卷积层相关的模块、部件或引擎(诸如卷积引擎1302和累积缓冲器1304)可以被使用或者是活动的。

[0308] 尽管图13的NN加速器1300示出了模块、引擎等的特定的布置顺序，并且因此示出了数据的处理如何流经NN加速器，但应当理解，这仅是示例，并且在其他示例中，模块、引擎可以以不同的方式布置。此外，其他硬件逻辑(例如，其他NN加速器)可实现NN层的另外的或另选的类型，并且因此可包括不同的模块、引擎等。

[0309] 替代成本度量

[0310] 在根据等式(33)中的clamp(x,low,high)的定义来实现本文中描述的阈值化步骤的示例中，被钳制到low或high的输入x(例如其中输入x取决于权重值w，诸如$x=w$或$x=2^{e}w$)可以生成不取决于x(例如并且因此在其中输入x取决于权重值w的示例中，不取决于w，)的输出。例如，其中low是呈由b和exp定义的定点数格式的最小或最低可表示数字(例如$low=-2^{b-exp-1}$)，并且high是呈由b和exp定义的定点数格式的最大或最高可表示数字(例如$high=2^{b+exp-1}-2^{exp}$)。这是因为，在此示例中，low和high都不取决于x(例如并且因此也不取决于w)。在这些示例中，不可能经由在阈值化步骤中使用的等式将成本度量相对于x(例如或w)的非零梯度反向传播到那些钳制值，这意味着可能并不可能有效地调整在阈值化步骤期间钳制的输入权重值。换句话说，在这些示例中，当执行本文中参考图11描述的方法时，可能只有在阈值化步骤期间没有被钳制的权重值可以具有经由在阈值化步骤中使用的等式反向传播到这些权重值的非零梯度，并且因此只有那些权重值可以分别在块1102和1104中被有效地调整。

[0311] 为了解决此情况，以及其中在阈值化步骤中使用的low和high的定义不取决于x

(例如并且因此也不取决于w)(这意味着不可能经由在该阈值化步骤中使用的等式将成本度量相对于x(例如或w)的非零梯度反向传播到钳制值)的其他示例,可以将替代成本度量(例如损失函数)用于块504中。等式(47)中示出了替代成本度量的示例。等式(3)与等式(47)的主要差异是引入了另一项——(γ*tm)。该另一项包括"阈值化度量"tm以及应用于该阈值化度量的权重γ。也就是说,成本度量可以是误差度量em、实现度量sm和阈值化度量tm的组合(例如加权和)。

[0312] cm=(α*em)+(β*sm)+(γ*tm) (47)

[0313] 阈值化度量tm的目的可以是在量化期间向输入值的阈值化分配成本。这意味着,当作为成本度量cm的一部分被最小化时,阈值化度量tm用于减少在阈值化步骤期间被钳制的输入值的数量——例如通过调整在阈值化步骤期间使用的钳制输入值以及/或者low和/或high阈值。例如,针对NN的阈值化度量tm可以通过对如根据等式(48)确定的NN的多个层l的"阈值化成本"t_l进行求和来形成——其中x_i取决于第i个权重w_i,例如$x_i=2^{-e}w_i$,并且N是层中的权重的数量。

[0314] $$t_l=\frac{1}{N}\sum_{i=1}^{N}(\max(0,low-x_i)+\max(0,x_i-high))$$

[0315] (48)

[0316] 在等式(48)中,权重值w_i对阈值化成本t_l的贡献仅对于呈定点数格式的可表示范围之外的权重值(即小于low或大于high并且因此在阈值化步骤中将被钳制到low或high的权重值)为非零。这是因为,例如,如果权重值在可表示范围内(例如大于low且小于high),则等式(48)中的"max"函数中的两者都将返回"0"。因此,将阈值化度量最小化用于将在阈值化步骤期间被钳制的权重值w_i朝向可由定点数格式表示的数字范围"推动",并且将那些权重值w_i被钳制到的相应low或high阈值朝向那些权重值w_i"拉动"。换句话说,将阈值化度量最小化会驱使权重值w_i以及low和high阈值朝向使得等式(48)中的"max"函数更经常地返回"0"的值(即凭借更多的权重值w_i在可表示范围内)。换句话说,这意味着在反向传播和调整期间,权重值w_i受到误差度量em和实现度量sm的影响(例如,如果该权重值w_i在可表示范围内,并且因此没有被钳制到low或high),或者受到阈值化度量tm的影响(例如,如果该权重值w_i在可表示范围之外,并且因此被钳制到low或high)。当权重值w_i受到阈值化度量tm的影响时,该权重值朝向可表示值的范围被"推"回,在该范围中,该权重值可能受到误差度量em和实现度量sm的影响。

[0317] 图14示出了示例性通用基于计算的设备1400的各种部件,该基于计算的设备可以被实现为任何形式的计算和/或电子设备,并且可以在其中实现上文描述的图5和图10A至图10E的方法500、1100的实施方案。

[0318] 基于计算的设备1400包括一个或多个处理器1402,该一个或多个处理器可以是微处理器、控制器或任何其他合适类型的处理器,用于处理计算机可执行指令以控制设备的操作,以便评估由硬件设计定义的集成电路在完成任务时的性能。在一些示例中,例如在使用片上系统架构的情况下,处理器1402可包括一个或多个固定功能块(也被称为加速器),该一个或多个固定功能块实现用于确定定点数格式的方法的一部分,该定点数格式用于表示在硬件(而不是软件或固件)中输入到NN的层或从其中输出的值集合。可以在基于计算的

设备处提供包括操作系统1404的平台软件或任何其他合适的平台软件，以使得诸如用于实现图5和图10A至图10E的方法500、1100中的一种或多种方法的计算机可执行代码1405的应用程序软件能够在设备上执行。

[0319] 可以使用可由基于计算的设备1400访问的任何计算机可读介质来提供计算机可执行指令。计算机可读介质可以包括例如计算机存储介质，诸如存储器1406和通信介质。计算机存储介质(即非暂态机器可读介质)诸如存储器1406，包括用于存储信息诸如计算机可读指令、数据结构或程序模块或其他数据的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括但不限于RAM、ROM、EPROM、EEPROM、闪存存储器或其他存储器技术、CD-ROM、数字多功能盘(DVD)或其他光学存储设备、磁带盒、磁带、磁盘存储或其他磁性存储设备，或可用于存储信息以供计算设备访问的任何其他非传输介质。相反，通信介质可以在调制数据信号例如载波或者其他传输机制中体现计算机可读指令、数据结构、程序模块或其他数据。如本文所定义，计算机存储介质不包括通信介质。尽管在基于计算的设备1400内示出了计算机存储介质(即非暂态机器可读介质，例如存储器1406)，但应当理解，存储器可以是分布式的或位于远程位置的，并且可经由网络或其他通信链路(例如，使用通信接口1408)进行访问。

[0320] 基于计算的设备1400还包括输入/输出控制器1410，所述输入/输出控制器被布置为将显示信息输出到显示设备1412，所述显示设备可以与基于计算的设备1400分离或成一体。显示信息可以提供图形用户界面。输入/输出控制器1410还被布置为接收和处理来自一个或多个设备(诸如用户输入设备1414(例如，鼠标或键盘))的输入。在一个实施方案中，如果显示设备1412是触敏显示设备，则其也可以充当用户输入设备1414。输入/输出控制器1410还可以将数据输出到除显示设备之外的设备，例如本地连接的打印设备(图14中未示出)。

[0321] 图15示出了其中可以实现可配置成实现本文中描述的NN的硬件逻辑(例如NN加速器)的计算机系统。计算机系统包括CPU 1502、GPU 1504、存储器1506和其他设备1514，诸如显示器1516、扬声器1518和相机1520。如图15中所示出，可配置成实现NN 1510的硬件逻辑(例如图13的NN加速器1300)可以在GPU 1504上实现。计算机系统的部件可经由通信总线1522彼此通信。在其他示例中，可配置成实现NN 1510的硬件逻辑可以独立于CPU或GPU来实现，并且可以具有与通信总线1522的单独连接。在一些示例中，可以不存在GPU，并且CPU可以将控制信息提供给可配置成实现NN 1510的硬件逻辑。

[0322] 图13的NN加速器1300被示出为包括多个功能块。这仅仅是示意性的，并且不旨在限定这类实体的不同逻辑元件之间的严格划分。每个功能块可以任何合适的方式提供。应当理解，本文中描述为由NN加速器或处理模块形成的中间值不需要在任何时候由NN加速器或处理模块物理生成，并且仅表示逻辑值，这些逻辑值方便地描述NN加速器或处理模块在其输入和输出之间执行的处理。

[0323] 可配置成实现本文中描述的NN的硬件逻辑(例如图13的NN加速器1300)可以在集成电路上的硬件中体现。一般来讲，上文所述的功能、方法、技术或部件中的任一者可在软件、固件、硬件(例如，固定逻辑电路系统)或它们的任何组合中实施。本文中可以使用术语“模块”、“功能性”、“部件”、“元件”、“单元”、“块”和“逻辑”来概括地表示软件、固件、硬件或它们的任何组合。在软件具体实施的情况下，模块、功能性、部件、元件、单元、块或逻辑表示

程序代码，该程序代码当在处理器上被执行时执行指定任务。本文中所描述的算法和方法可由执行代码的一个或多个处理器执行，所述代码促使处理器执行算法/方法。计算机可读存储介质的示例包括随机访问存储器(RAM)、只读存储器(ROM)、光盘、闪存存储器、硬盘存储器，以及可使用磁性、光学和其他技术来存储指令或其他数据并且可由机器访问的其他存储器设备。

[0324] 如本文中所使用的术语计算机程序代码和计算机可读指令是指用于处理器的任何种类的可执行代码，包括以机器语言、解译语言或脚本语言表达的代码。可执行代码包括二进制代码、机器代码、字节代码、定义集成电路的代码(诸如硬件描述语言或网表)，以及用诸如C、Java或OpenCL等编程语言代码表达的代码。可执行代码可以是例如任何种类的软件、固件、脚本、模块或库，当在虚拟机或其他软件环境中被适当地执行、处理、解译、编译、运行时，这些软件、固件、脚本、模块或库使得支持可执行代码的计算机系统的处理器执行由所述代码指定的任务。

[0325] 处理器、计算机或计算机系统可以是任何种类的设备、机器或专用电路，或其集合或一部分，它具有处理能力使得可以执行指令。处理器可以是任何种类的通用或专用处理器，诸如CPU、GPU、片上系统、状态机、媒体处理器、专用集成电路(ASIC)、可编程逻辑阵列、现场可编程门阵列(FPGA)等。计算机或计算机系统可以包括一个或多个处理器。

[0326] 本发明还意图涵盖限定如本文中所描述的硬件的配置的软件，诸如HDL(硬件描述语言)软件，如用于设计集成电路，或者用于配置可编程芯片以实施所需功能。即，可以提供一种计算机可读存储介质，其上编码有呈集成电路定义数据集形式的计算机可读程序代码，当在集成电路制造系统中被处理(即运行)时，该计算机可读程序代码将系统配置成制造可配置成实现本文中描述的NN的硬件逻辑(例如NN加速器)。集成电路定义数据集可以是例如集成电路描述。

[0327] 因此，可以提供一种在集成电路制造系统处制造可配置成实现如本文中所描述的NN的硬件逻辑(例如图13的NN加速器1300)的方法。此外，可以提供一种集成电路定义数据集，当在集成电路制造系统中被处理时，该集成电路定义数据集使得制造可配置成实现NN的硬件逻辑(例如图13的NN加速器1300)的方法得以执行。

[0328] 集成电路定义数据集可以是计算机代码的形式，例如作为网表，用于配置可编程芯片的代码，作为定义适合于在集成电路中以任何层级制造的硬件描述语言，包括作为寄存器传输级(RTL)代码，作为高级电路表示法(诸如Verilog或VHDL)，以及作为低级电路表示法(诸如，OASIS(RTM)和GDSII)。在逻辑上定义适合于在集成电路中制造的硬件的更高级表示法(诸如RTL)可在计算机系统处进行处理，该计算机系统被配置用于在软件环境的上下文中产生集成电路的制造定义，该软件环境包括电路元件的定义以及用于组合这些元件以便产生由表示法如此定义的集成电路的制造定义的规则。如通常软件在计算机系统处执行以便定义机器的情况一样，可能需要一个或多个中间用户步骤(例如提供命令、变量等)，以便将计算机系统配置成生成集成电路的制造定义，以执行定义集成电路以便生成该集成电路的制造定义的代码。

[0329] 现在将针对图16来描述在集成电路制造系统处处理集成电路定义数据集以便将系统配置成制造可配置成实现NN的硬件逻辑(例如NN加速器)的示例。

[0330] 图16示出了集成电路(IC)制造系统1602的示例，该集成电路制造系统被配置成制

造可配置成实现如本文中的示例中的任何示例中所描述的NN的硬件逻辑(例如NN加速器)。具体地,IC制造系统1602包括布局处理系统1604和集成电路生成系统1606。IC制造系统1602被配置成接收IC定义数据集(例如定义可配置成实现本文中的示例中的任何示例中所描述的NN的硬件逻辑(例如NN加速器)),处理IC定义数据集,并且根据IC定义数据集来生成IC(例如其体现可配置成实现本文中的示例中的任何示例中所描述的NN的硬件逻辑(例如NN加速器))。对IC定义数据集的处理将IC制造系统1602配置成制造集成电路,该集成电路体现可配置成实现如本文中的示例中的任何示例中所描述的NN的硬件逻辑(例如NN加速器)。

[0331] 布局处理系统1604配置成接收并处理IC定义数据集以确定电路布局。根据IC定义数据集确定电路布局的方法在本领域中是已知的,并且例如可以涉及合成RTL代码以确定要生成的电路的门级表示,例如就逻辑部件(例如NAND、NOR、AND、OR、MUX和FLIP-FLOP部件)而言。通过确定逻辑部件的位置信息,可以根据电路的门级表示来确定电路布局。这可以自动完成或者在用户参与下完成,以便优化电路布局。当布局处理系统1604已经确定电路布局时,其可将电路布局定义输出到IC生成系统1606。电路布局定义可以是例如电路布局描述。

[0332] 如所属领域中已知,IC生成系统1606根据电路布局定义来生成IC。例如,IC生成系统1606可实施生成IC的半导体设备制造工艺,其可涉及光刻和化学处理步骤的多步骤序列,在此期间,在由半导体材料制成的晶片上逐渐形成电子电路。电路布局定义可呈掩模的形式,其可在光刻工艺中用于根据电路定义来生成IC。替代地,提供给IC生成系统1606的电路布局定义可呈计算机可读代码的形式,IC生成系统1606可使用所述计算机可读代码来形成用于生成IC的合适掩码。

[0333] 由IC制造系统1602执行的不同过程可全部在一个位置例如由一方来实施。替代地,IC制造系统1602可以是分布式系统,使得一些过程可以在不同位置执行,并且可以由不同方来执行。例如,以下阶段中的一些阶段可以在不同位置和/或由不同方来执行:(i)合成表示IC定义数据集的RTL代码,以形成待生成的电路的门级表示;(ii)基于门级表示来生成电路布局;(iii)根据电路布局来形成掩模;以及(iv)使用掩模来制造集成电路。

[0334] 在其他示例中,在集成电路制造系统处对集成电路定义数据集的处理可以将系统配置成制造可配置成实现NN的硬件逻辑(例如NN加速器),而无需处理IC定义数据集以便确定电路布局。例如,集成电路定义数据集可以定义可重新配置的处理器(诸如FPGA)的配置,并且对该数据集的处理可以将IC制造系统配置成(例如通过将配置数据加载到FPGA)生成具有该定义配置的可重新配置的处理器。

[0335] 在一些实施方案中,当在集成电路制造系统中被处理时,集成电路制造定义数据集可以使得集成电路制造系统生成如本文中所描述的设备。例如,通过集成电路制造定义数据集,以上文关于图16所描述的方式对集成电路制造系统的配置,可使得制造出如本文所描述的设备。

[0336] 在一些示例中,集成电路定义数据集可包括在数据集处定义的硬件上运行的软件,或者与在数据集处定义的硬件组合运行的软件。在图16中所示的示例中,IC生成系统可以由集成电路定义数据集进一步配置,以在制造集成电路时根据在集成电路定义数据集处定义的程序代码将固件加载到所述集成电路上,或者以其他方式向集成电路提供与集成电

路一起使用的程序代码。

[0337] 与已知的具体实施相比，在本申请中阐述的概念在设备、装置、模块和/或系统中(以及在本文中所实现的方法中)的具体实施可引起性能改进。性能改进可包括计算性能提高、等待时间减少、吞吐量增大和/或功耗减小中的一者或多者。在制造这类设备、装置、模块和系统(例如在集成电路中)期间，可在性能提高与物理具体实现之间进行权衡，从而改进制造方法。例如，可在性能改进与布局面积之间进行权衡，从而匹配已知具体实施的性能，但使用更少的硅。例如，这可以通过以串行方式重复使用功能块或在设备、装置、模块和/或系统的元件之间共享功能块来完成。相反，本申请中所阐述的带来设备、装置、模块和系统的物理实现的改进(例如，硅面积减小)的概念可与性能提高进行权衡。这可以例如通过在预定义面积预算内制造模块的多个实例来完成。

[0338] 申请人据此独立地公开了本文中所描述的每个单独特征以及两个或更多个这类特征的任何组合，到达的程度使得这类特征或组合能够根据本领域的技术人员的普通常识基于本说明书整体来实行，而不管这类特征或特征的组合是否解决本文中所公开的任何问题。鉴于前文描述，本领域技术人员将清楚，可以在本发明的范围内进行各种修改。

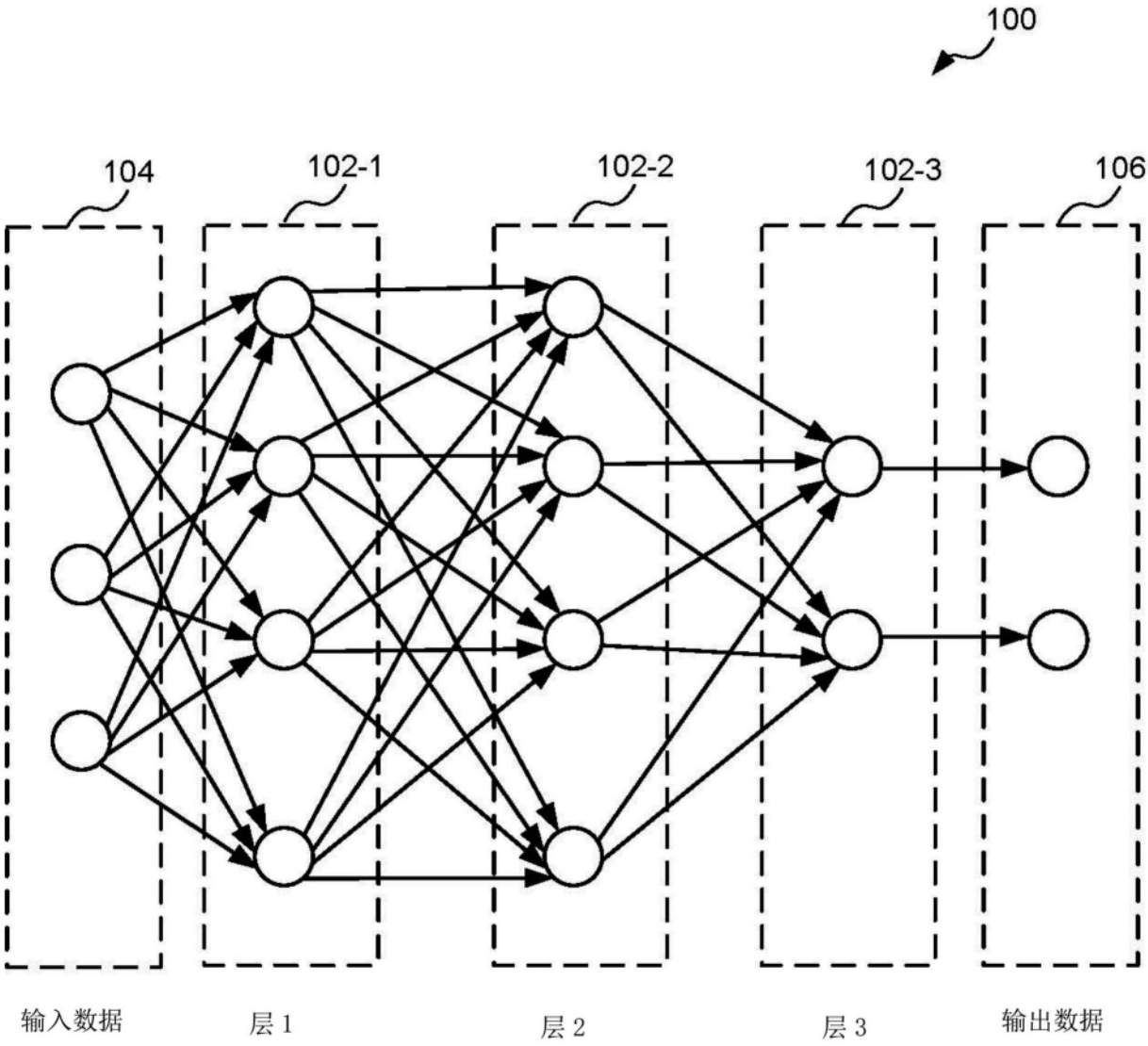


图1

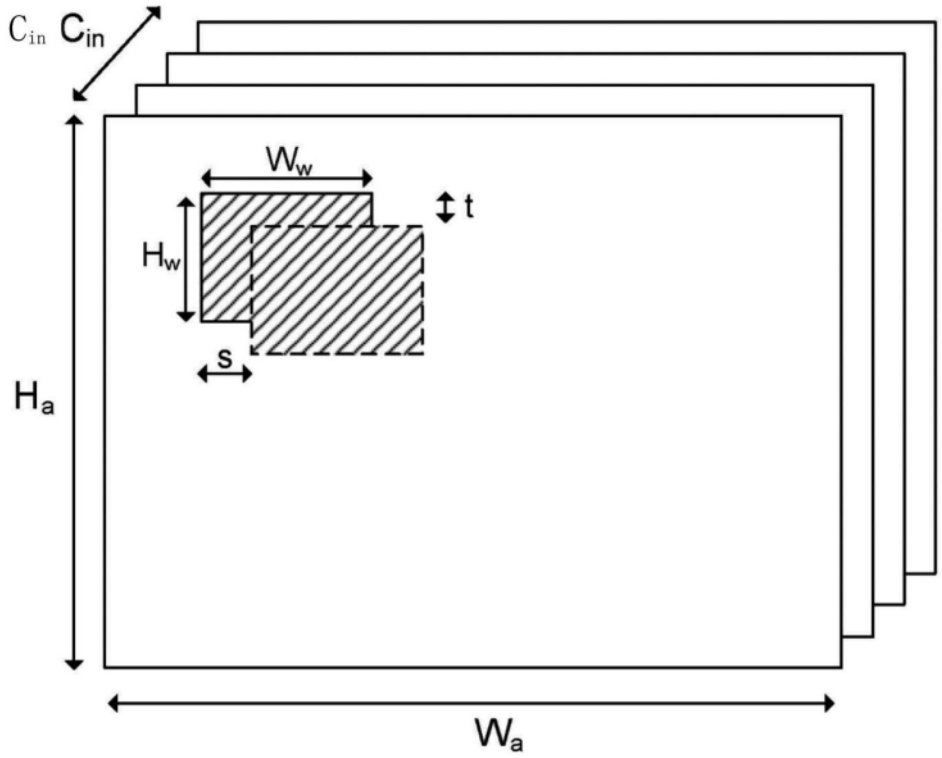


图2A

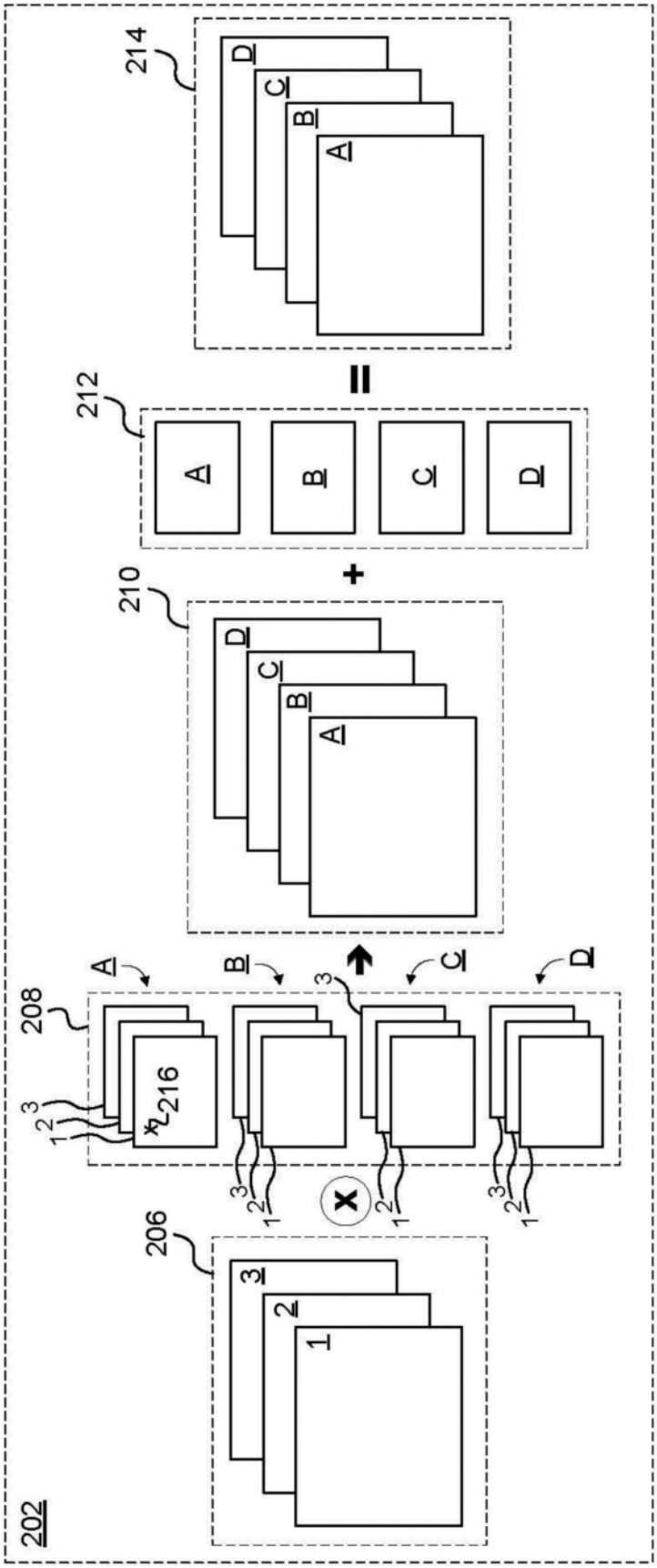


图2B

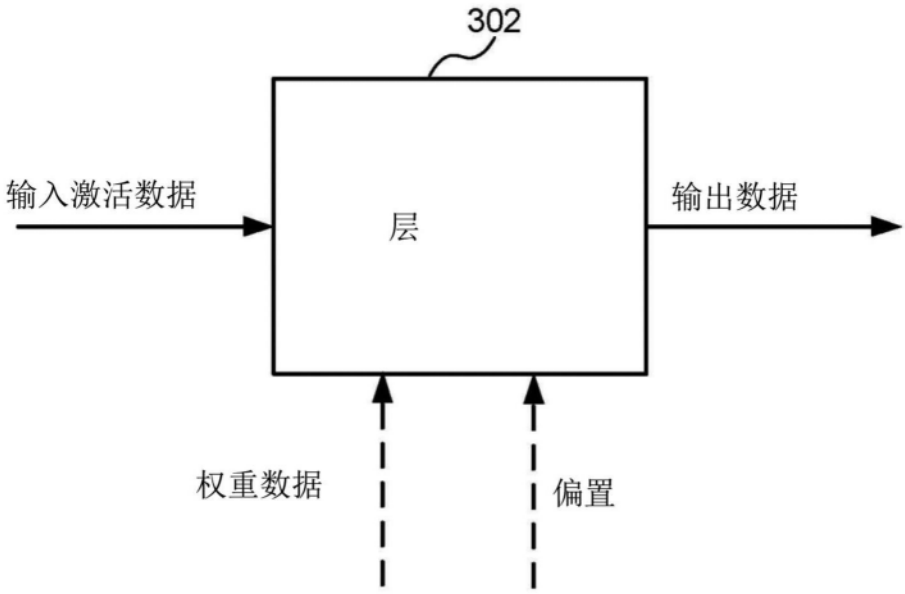


图3

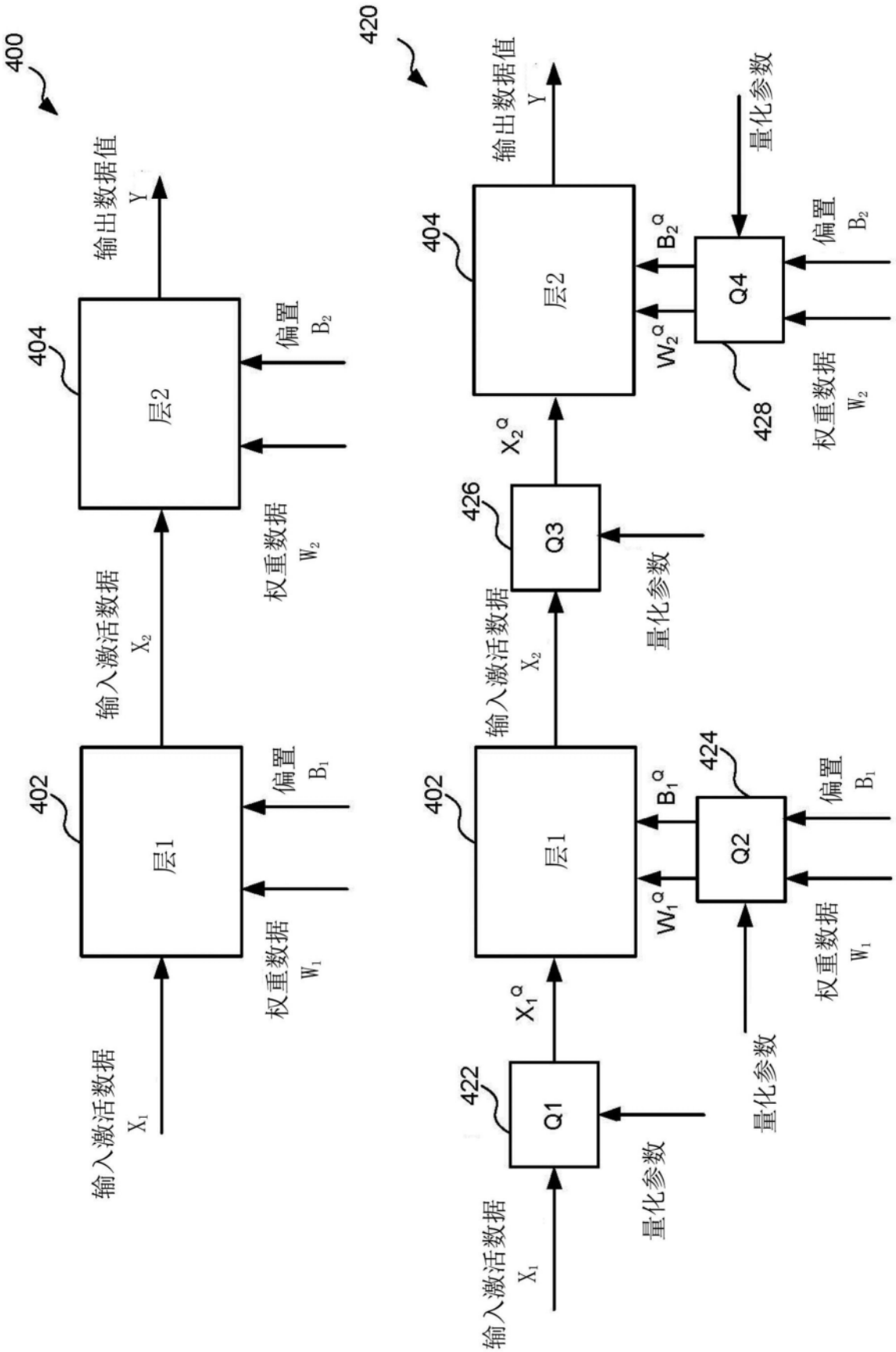

400
输入激活数据
X_1
层1
402
权重数据
W_1
偏置
B_1
输入激活数据
X_2
层2
404
权重数据
W_2
偏置
B_2
输出数据值
Y
420
输入激活数据
X_1
量化参数
Q1
422
X_1^Q
层1
402
W_1^Q
B_1^Q
量化参数
Q2
424
权重数据
W_1
偏置
B_1
输入激活数据
X_2
量化参数
Q3
426
X_2^Q
层2
404
W_2^Q
B_2^Q
量化参数
Q4
428
权重数据
W_2
偏置
B_2
输出数据值
Y

图4

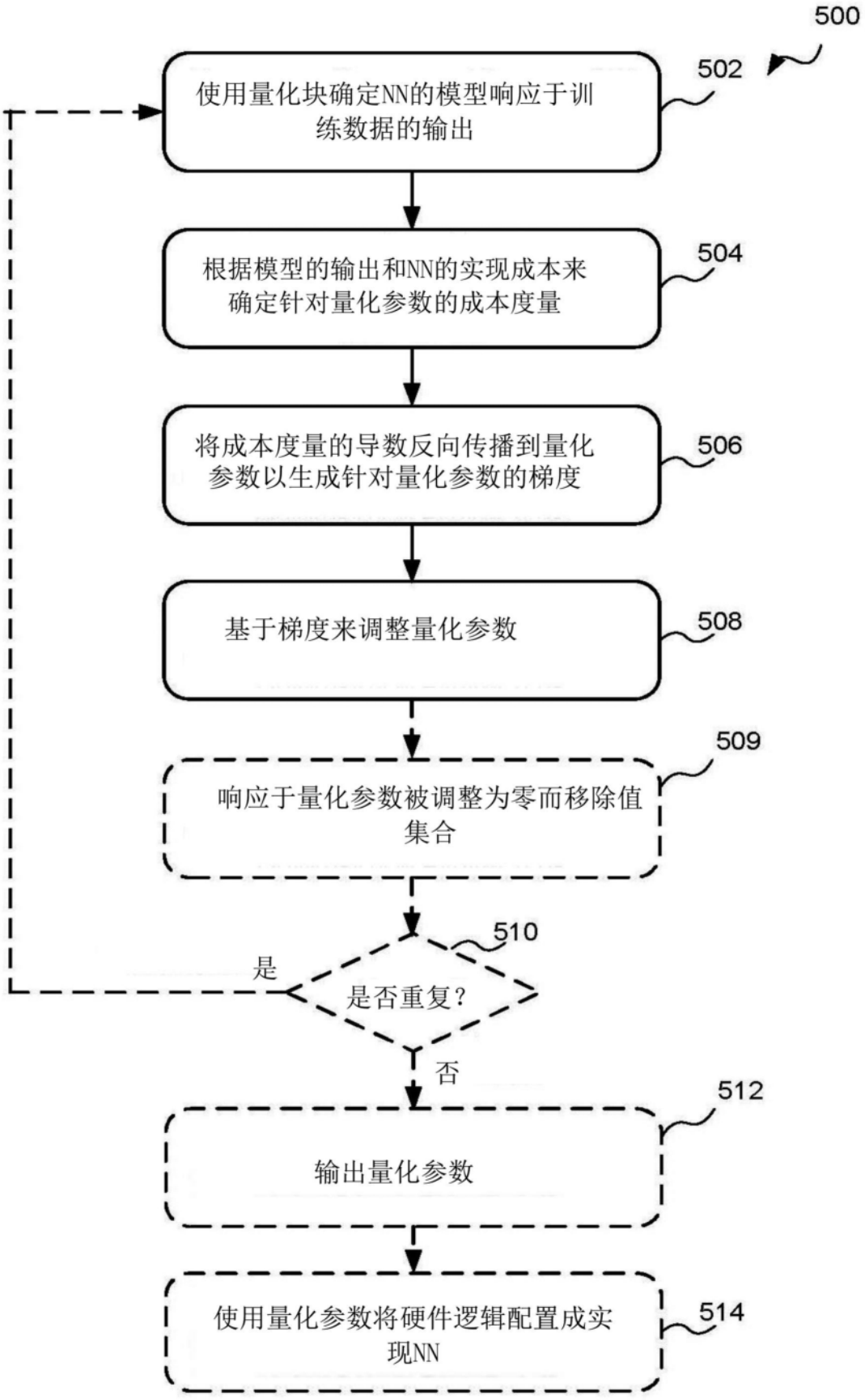

500
使用量化块确定NN的模型响应于训练数据的输出
502
根据模型的输出和NN的实现成本来确定针对量化参数的成本度量
504
将成本度量的导数反向传播到量化参数以生成针对量化参数的梯度
506
基于梯度来调整量化参数
508
响应于量化参数被调整为零而移除值集合
509
是否重复？
510
是
否
输出量化参数
512
使用量化参数将硬件逻辑配置成实现NN
514

图5

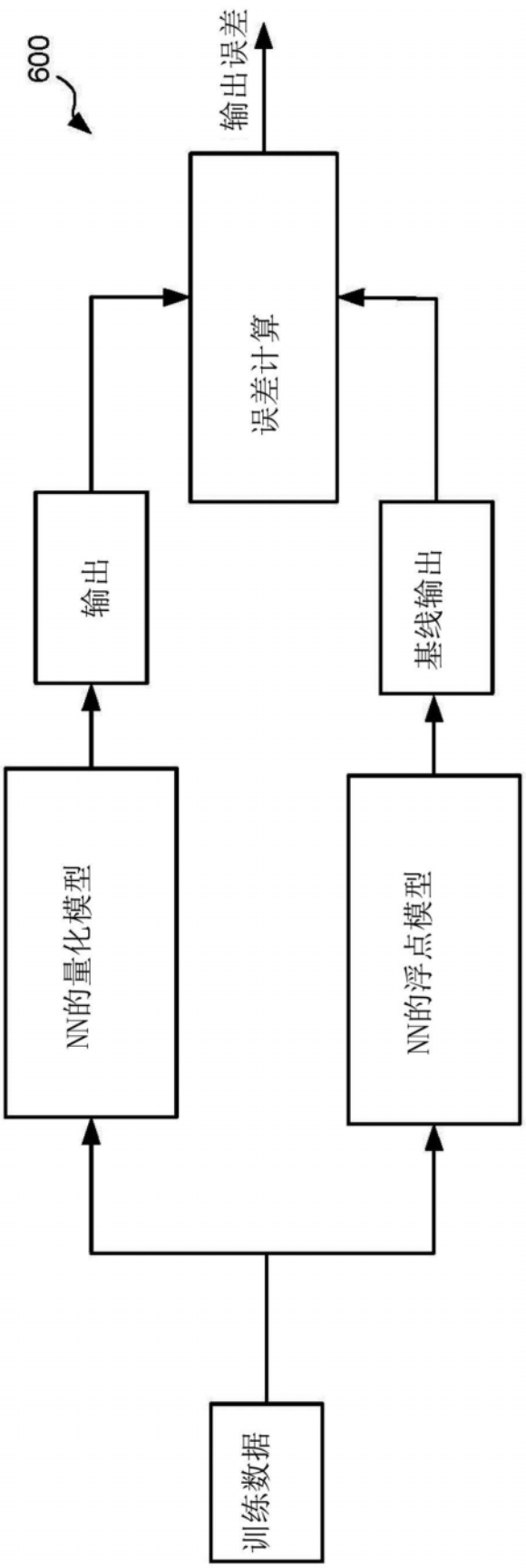


图6

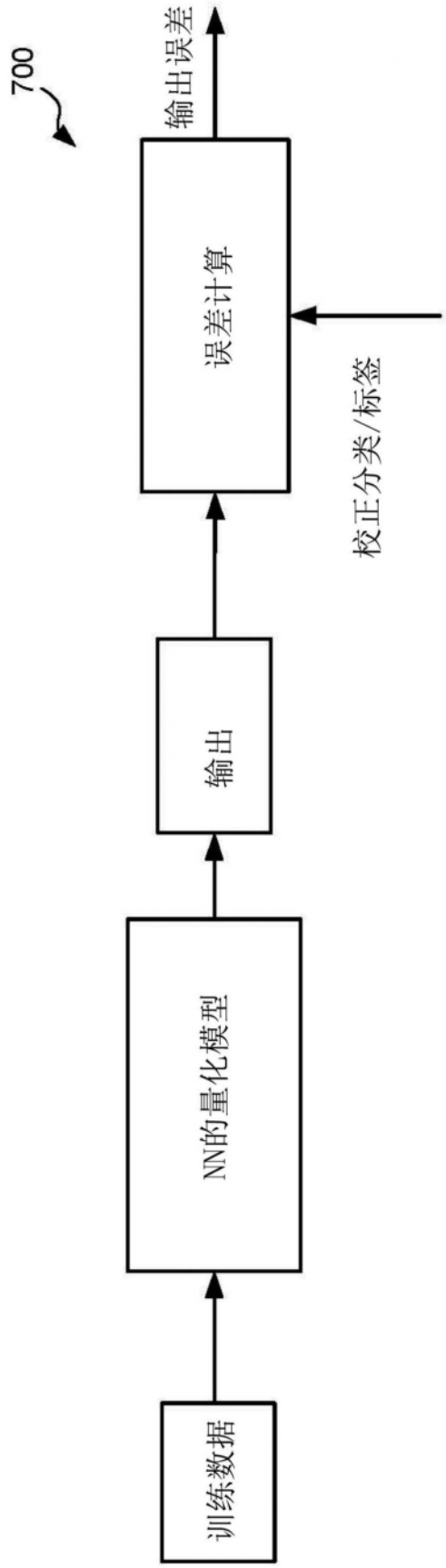


图7

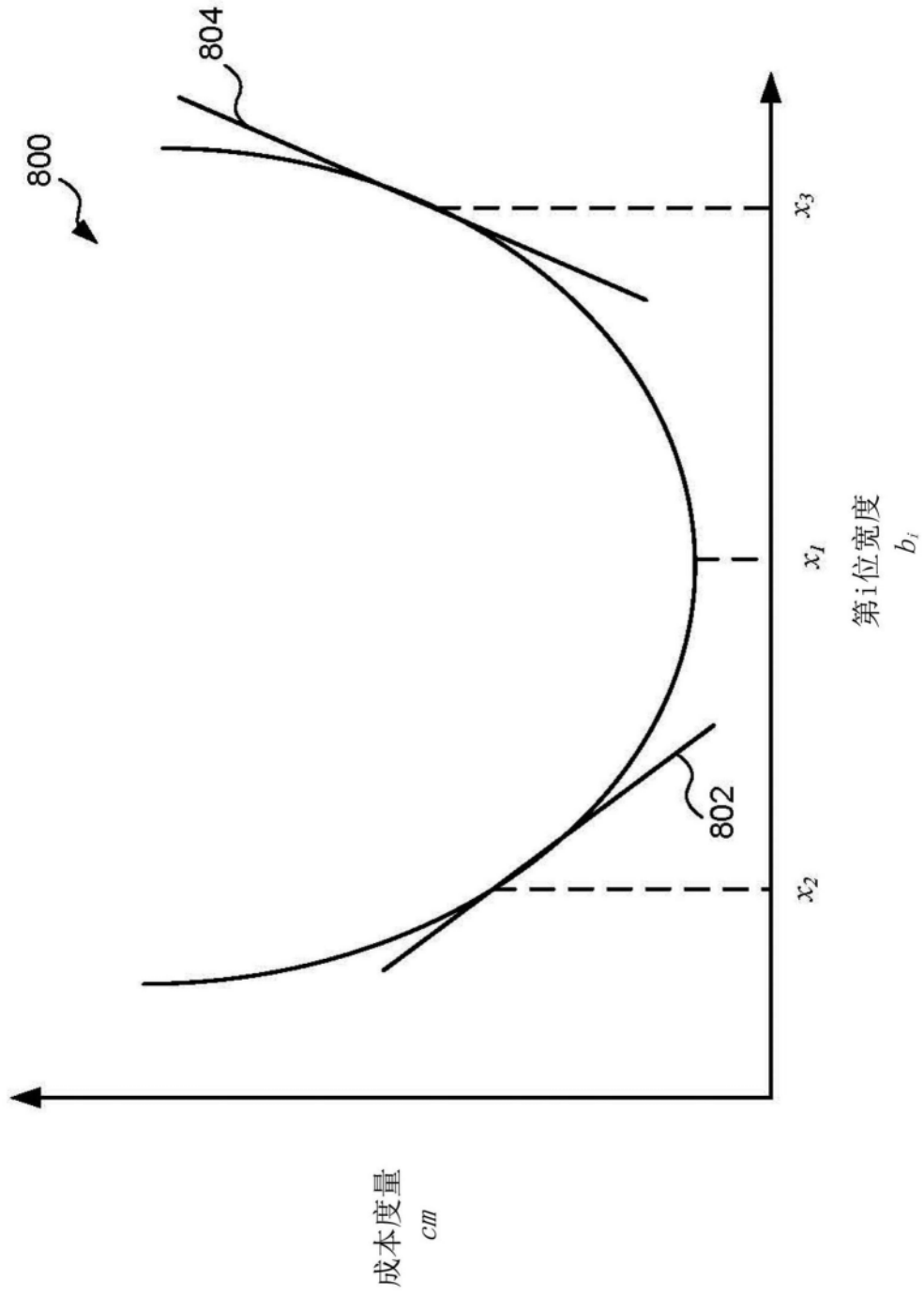


图8

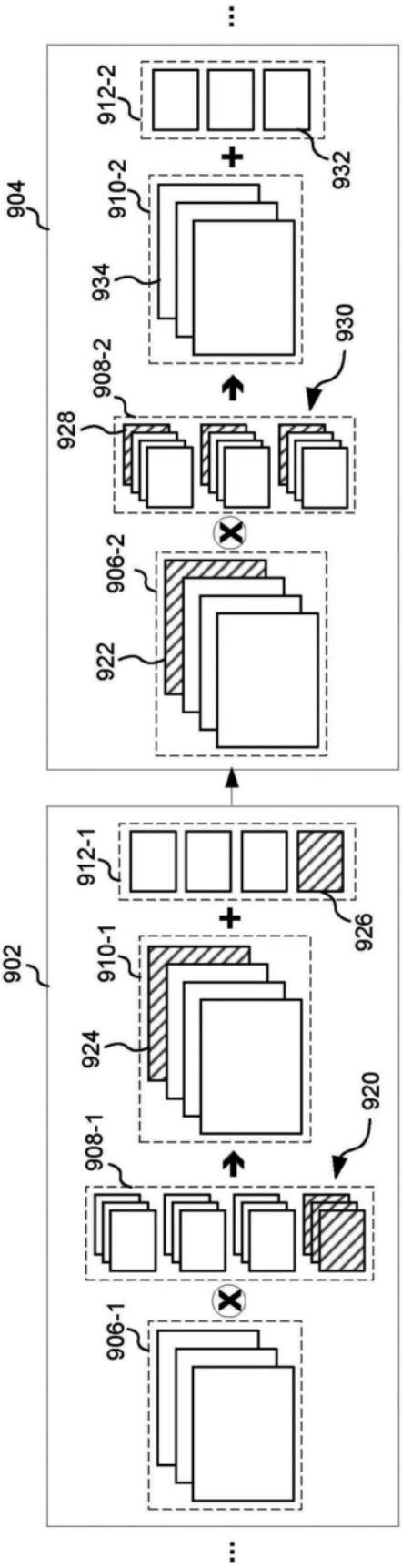

912-2
932
910-2
904
934
930
908-2
928
906-2
922
912-1
926
910-1
902
924
920
908-1
906-1

图9

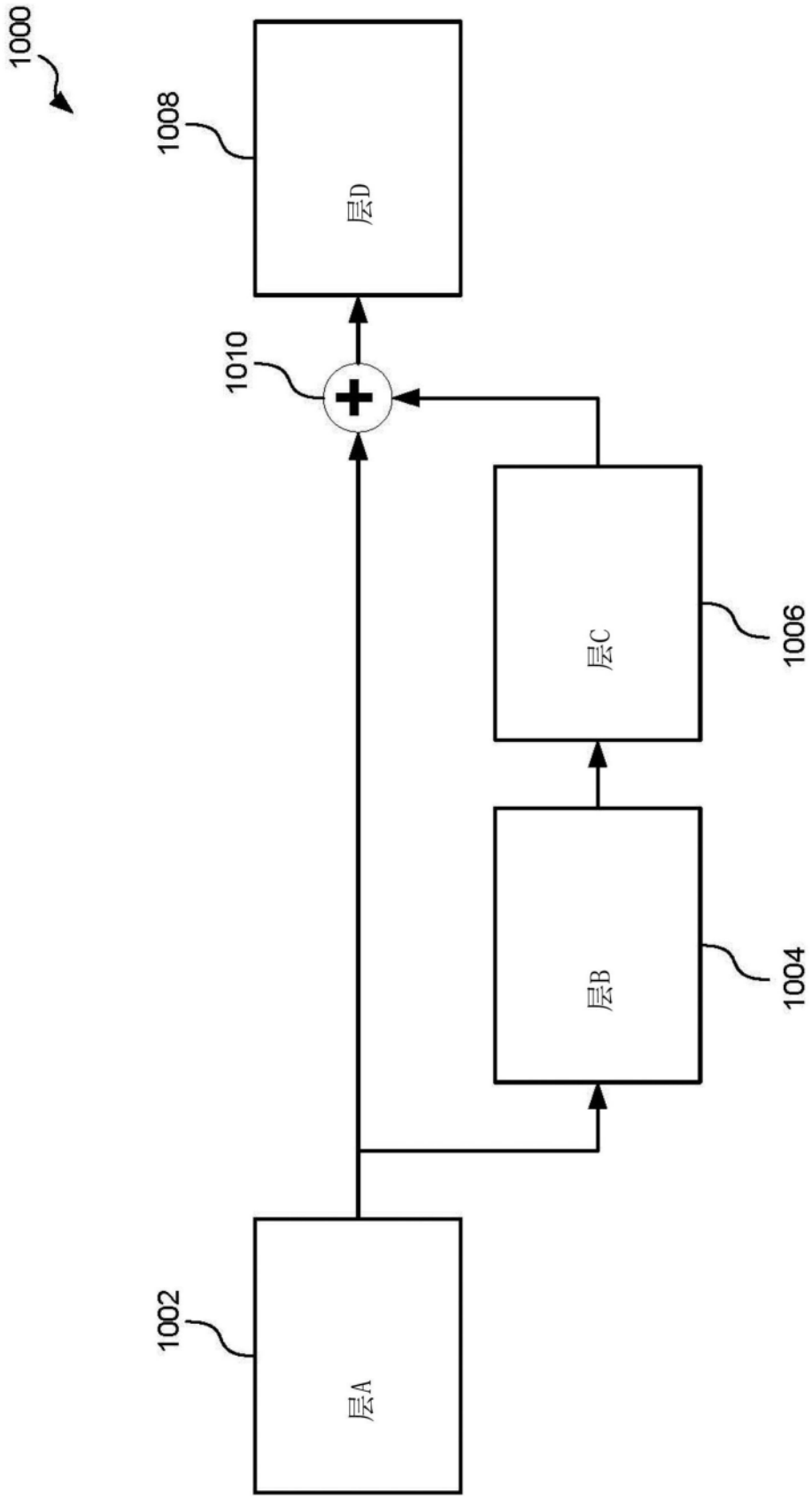


图10A

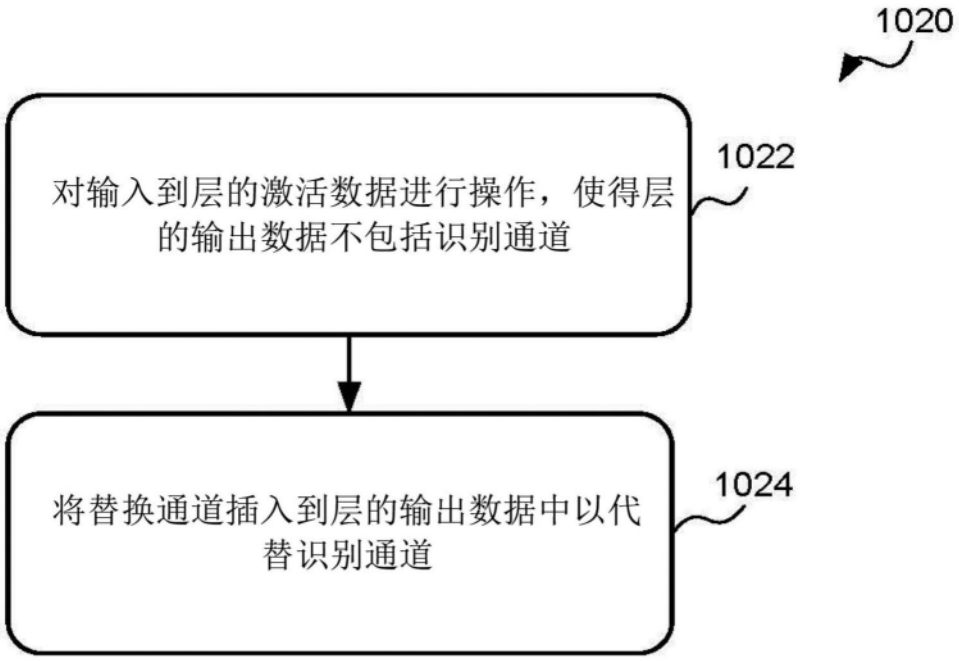

1020
对输入到层的激活数据进行操作，使得层的输出数据不包括识别通道
1022
将替换通道插入到层的输出数据中以代替识别通道
1024

图10B

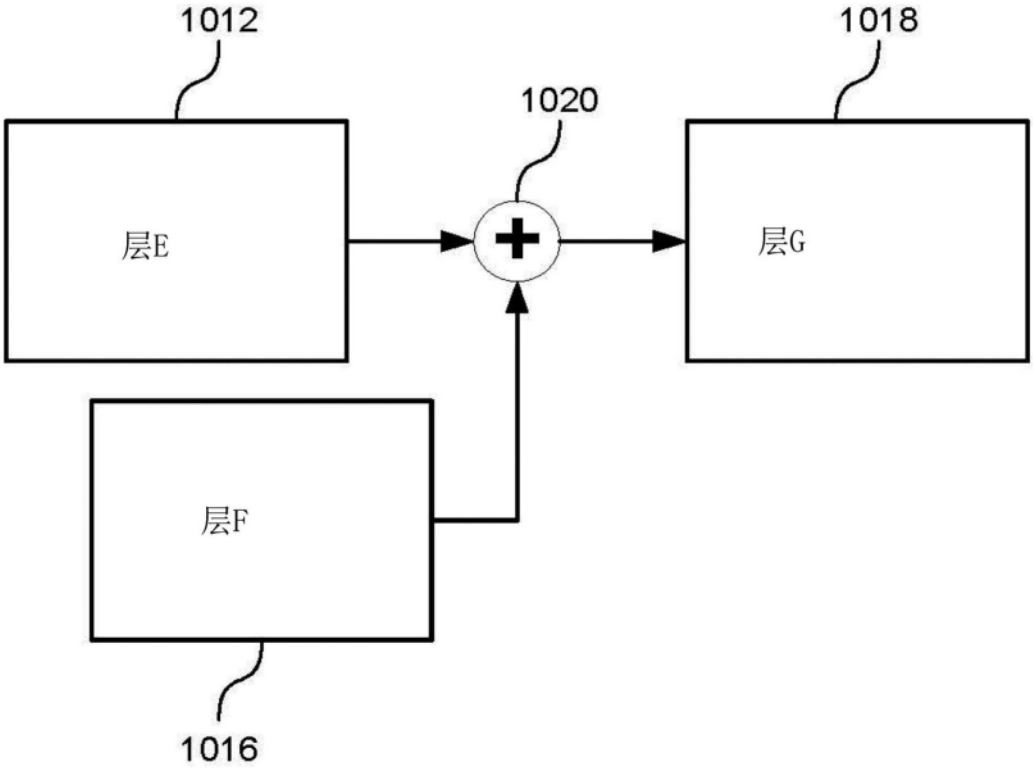

1012
1018
1020
层E
层G
层F
1016

图10C

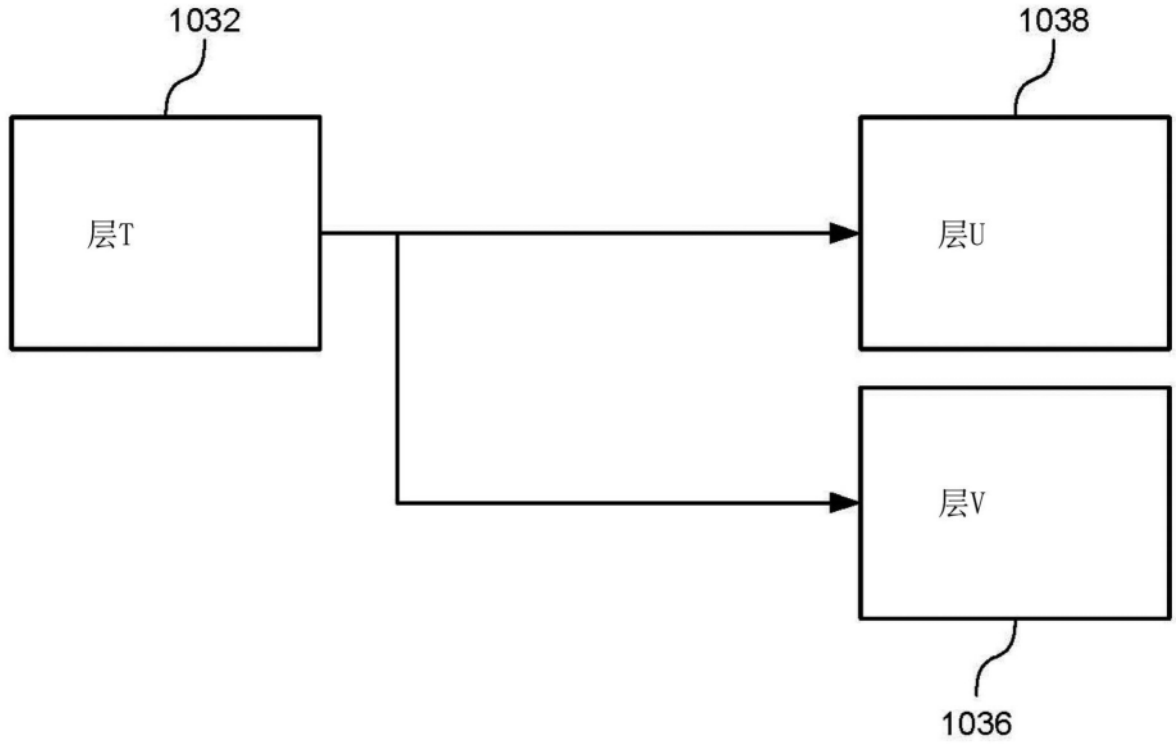


图10D

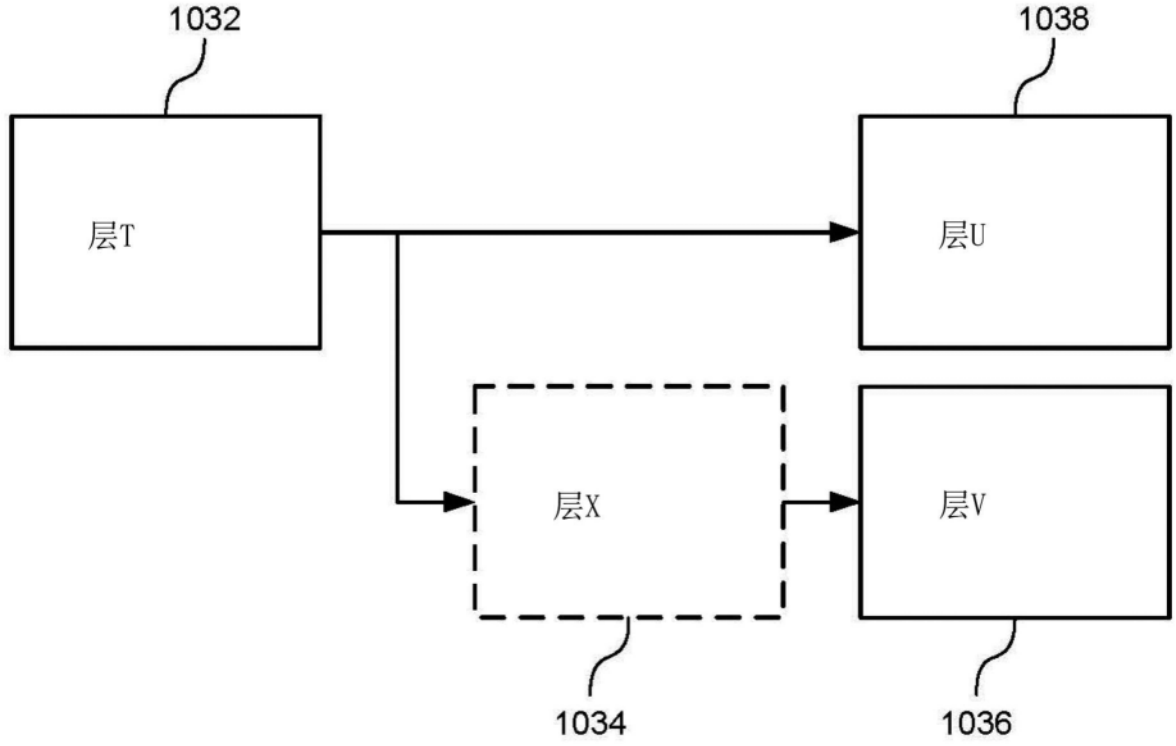


图10E

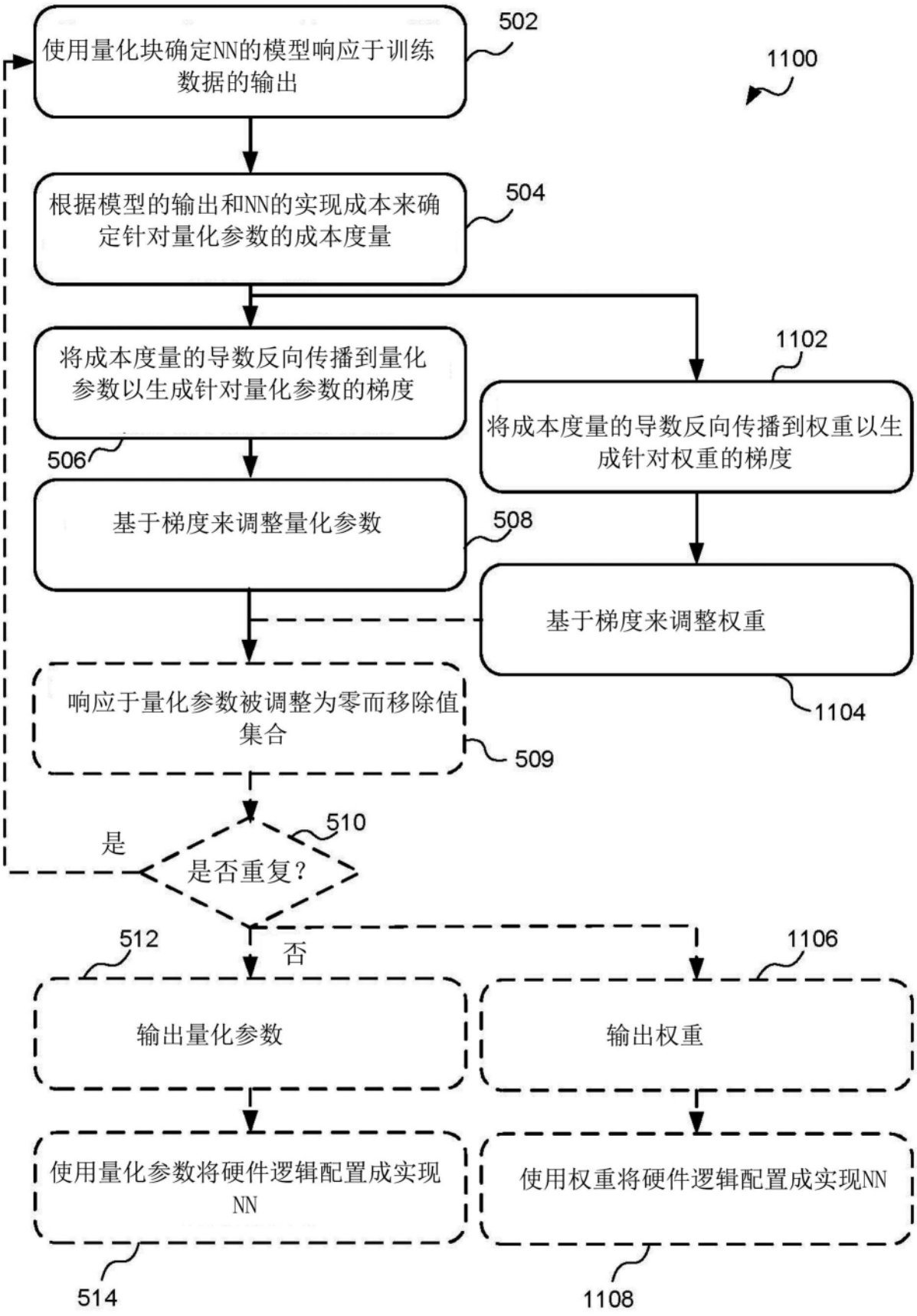

使用量化块确定NN的模型响应于训练数据的输出
502
1100
根据模型的输出和NN的实现成本来确定针对量化参数的成本度量
504
将成本度量的导数反向传播到量化参数以生成针对量化参数的梯度
506
1102
将成本度量的导数反向传播到权重以生成针对权重的梯度
基于梯度来调整量化参数
508
基于梯度来调整权重
1104
响应于量化参数被调整为零而移除值集合
509
是
是否重复?
510
否
512
1106
输出量化参数
输出权重
使用量化参数将硬件逻辑配置成实现NN
使用权重将硬件逻辑配置成实现NN
514
1108

图11

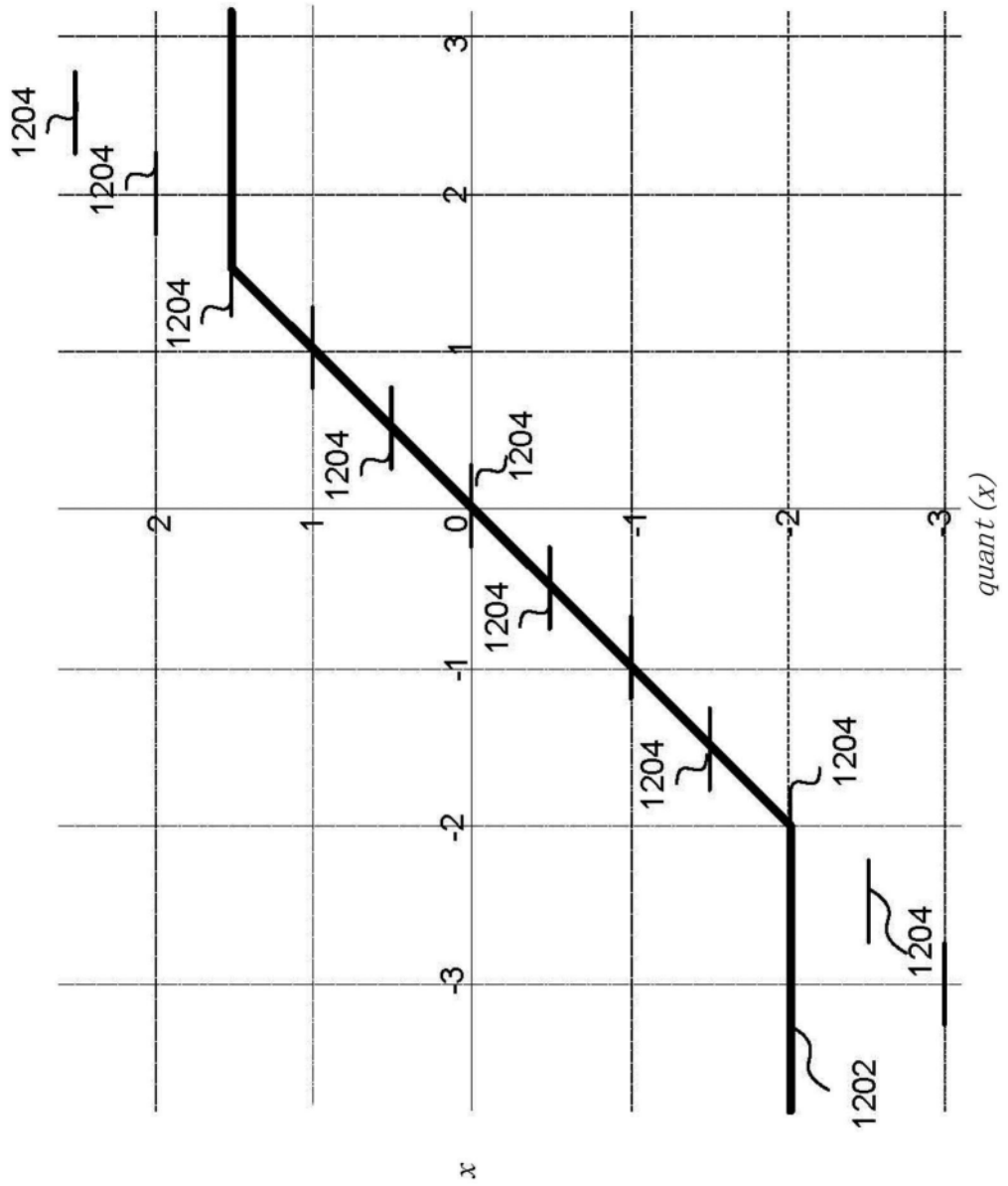

1204
1204
1204
1204
1204
1204
1204
1204
1204
1202
3
2
1
0
-1
-2
-3
2
1
0
-1
-2
-3
x
quant (x)

图12

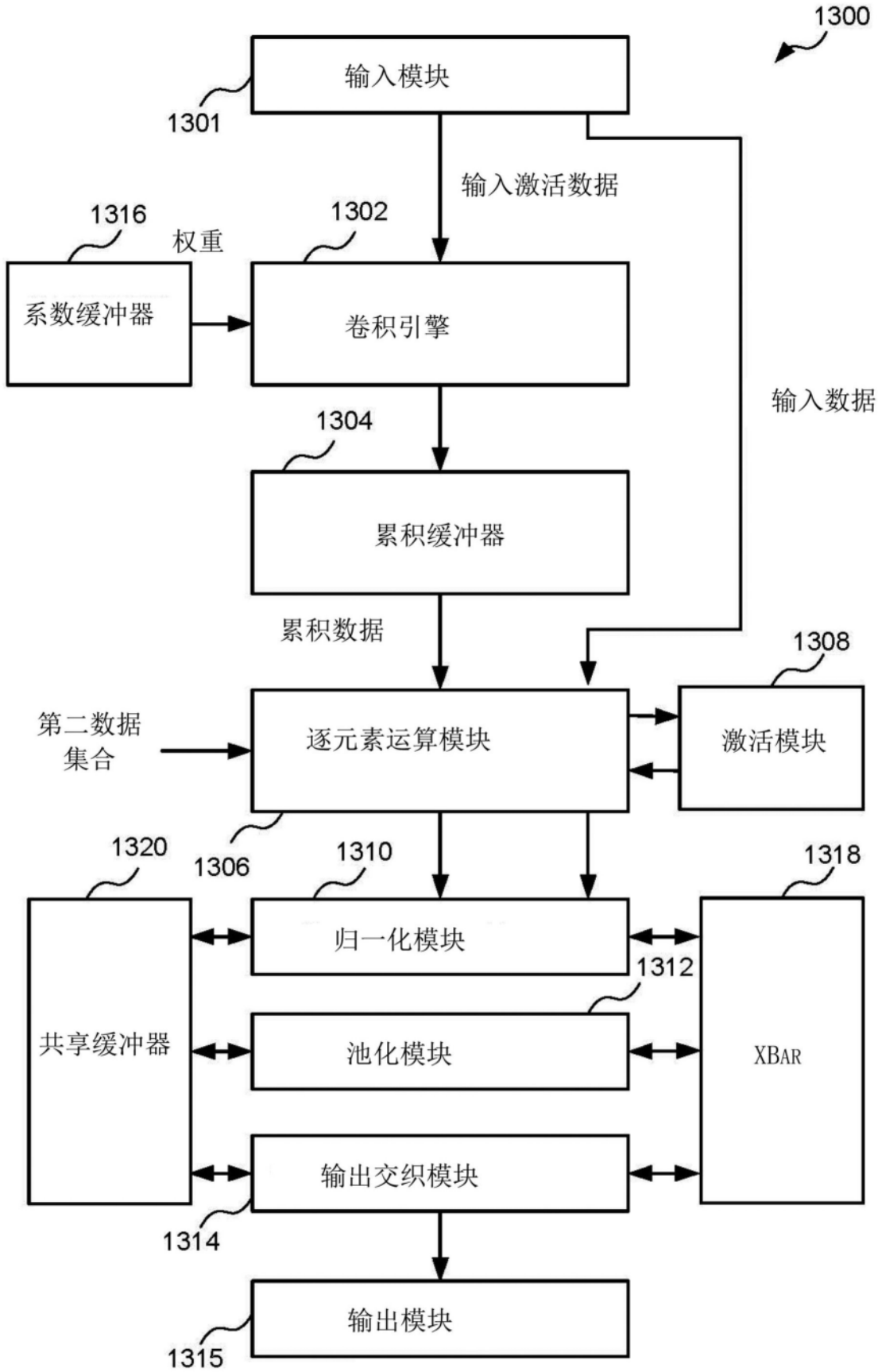

1300
输入模块
1301
输入激活数据
1316
权重
1302
系数缓冲器
卷积引擎
输入数据
1304
累积缓冲器
累积数据
1308
第二数据集合
逐元素运算模块
激活模块
1320
1306
1310
1318
归一化模块
1312
共享缓冲器
池化模块
XBAR
输出交织模块
1314
输出模块
1315

图13

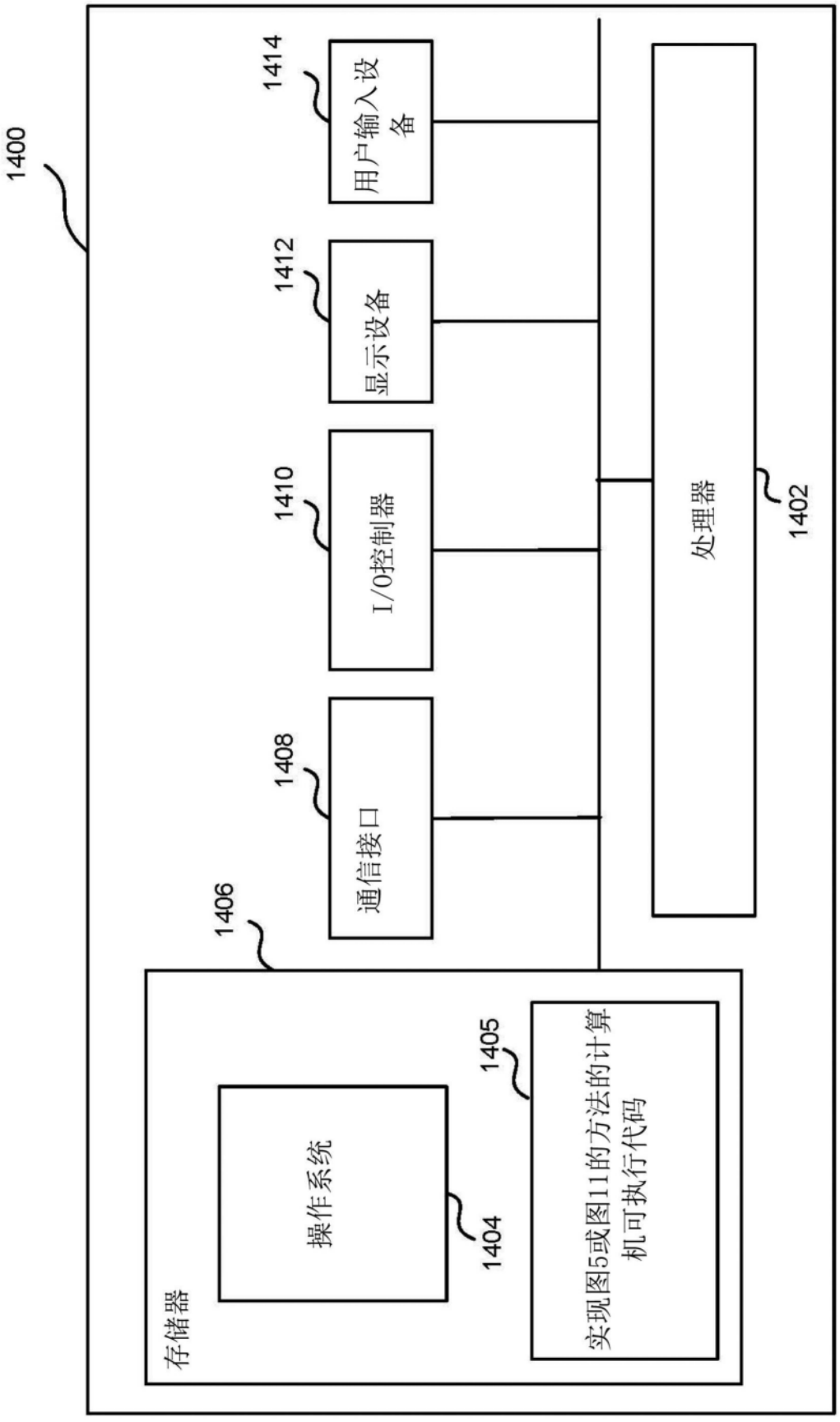

1400
1414
用户输入设备
1412
显示设备
1410
I/O控制器
1408
通信接口
1406
存储器
操作系统
1404
1405
实现图5或图11的方法的计算机可执行代码
处理器
1402

图14

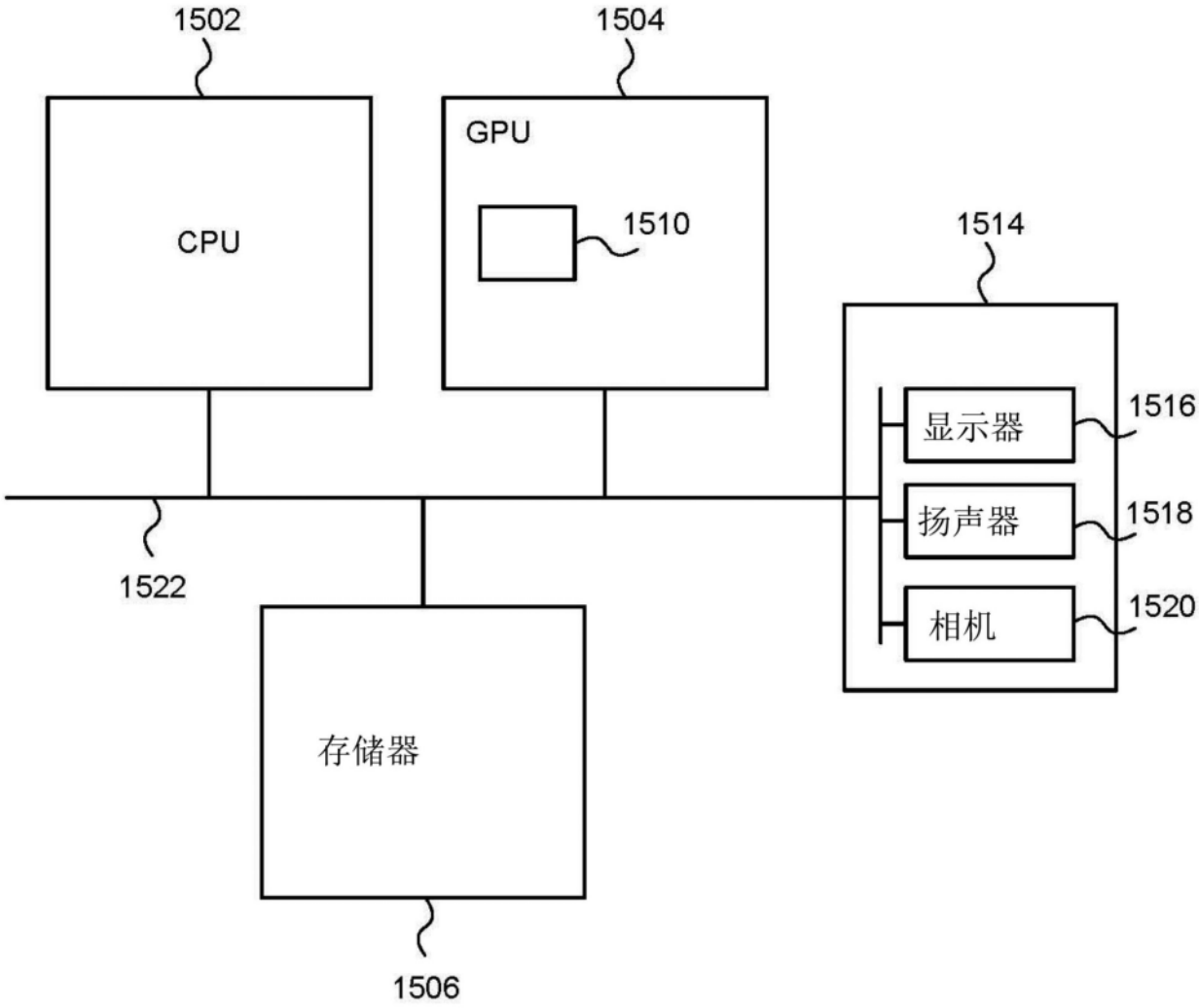

1502
1504
CPU
GPU
1510
1514
显示器
1516
扬声器
1518
相机
1520
1522
存储器
1506

图15

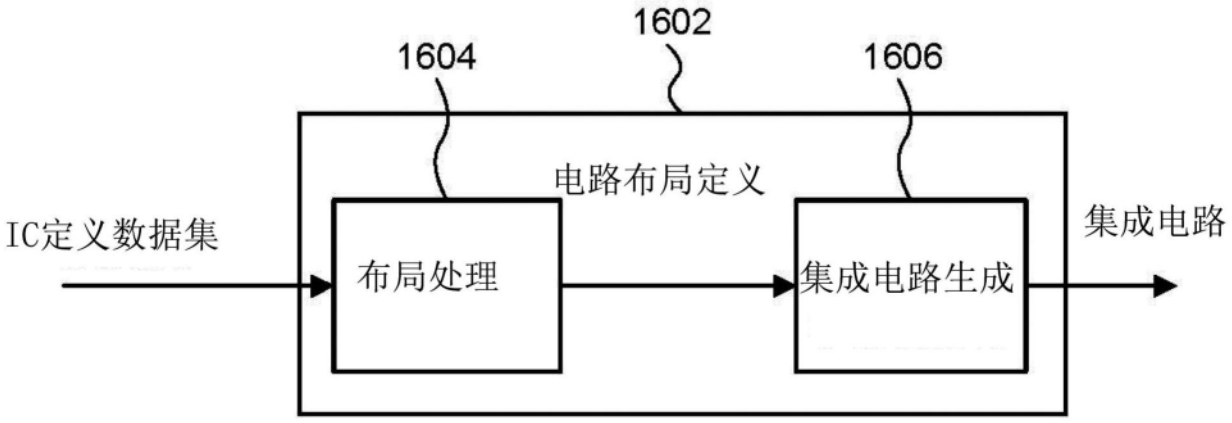

1602
1604
1606
电路布局定义
IC定义数据集
布局处理
集成电路生成
集成电路

图16