(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 113902114 A
(43)申请公布日 2022.01.07

(21)申请号 202111153963.1

(22)申请日 2021.09.29

(71)申请人 南京后摩智能科技有限公司
地址 210046 江苏省南京市栖霞区经济技术开发区兴智路6号兴智科技园C栋第18层1807室

(72)发明人 袁之航 陈亮 赵亦彤 王辉 吴强

(74)专利代理机构 北京思源智汇知识产权代理有限公司 11657
代理人 毛丽琴

(51)Int.Cl.
G06N 3/08(2006.01)
G06N 3/063(2006.01)

权利要求书4页 说明书17页 附图3页

(54)发明名称

神经网络模型的量化方法、装置和系统、电子设备和存储介质

(57)摘要

本公开实施例公开了一种神经网络模型的量化方法、装置和系统、电子设备和介质，其中，方法包括：分别针对待量化神经网络模型中的任一待量化网络层，获取所述任一待量化网络层的权重矩阵；对所述任一待量化网络层的权重矩阵进行矩阵变换，得到所述任一待量化网络层的待量化权重矩阵；对所述任一待量化网络层的待量化权重矩阵进行量化，得到所述任一待量化网络层的量化后权重矩阵；基于所述待量化神经网络模型中的任一待量化网络层的量化后权重矩阵，得到量化后神经网络模型。本公开实施例可以降低权重矩阵中各通道的权重数据的分布差异，能够减少量化误差，有助于提升量化后神经网络的精度。

CN 113902114 A

1.一种神经网络模型的量化方法,其特征在于,包括:

分别针对待量化神经网络模型中的任一待量化网络层,获取所述任一待量化网络层的权重矩阵;

对所述任一待量化网络层的权重矩阵进行矩阵变换,得到所述任一待量化网络层的待量化权重矩阵;

对所述任一待量化网络层的待量化权重矩阵进行量化,得到所述任一待量化网络层的量化后权重矩阵;

基于所述待量化神经网络模型中的任一待量化网络层的量化后权重矩阵,得到量化后神经网络模型。

2.根据权利要求1所述的方法,其特征在于,所述对所述任一待量化网络层的待量化权重矩阵进行量化,得到所述任一待量化网络层的量化后权重矩阵,包括:

根据存算一体加速器中的存算单元阵列的大小,将所述任一待量化网络层的待量化权重矩阵进行划分,得到多个待量化子权重矩阵;

分别对所述多个待量化子权重矩阵中的任一待量化子权重矩阵进行量化,得到多个量化后的子权重矩阵;

将所述多个量化后的子权重矩阵分别映射到所述存算一体加速器中的多个存算单元阵列,由所述多个存算单元阵列存储所述多个量化后的子权重矩阵。

3.根据权利要求1或2所述的方法,其特征在于,所述基于所述待量化神经网络模型中的任一待量化网络层的量化后权重矩阵,得到量化后神经网络模型之后,还包括:

从校准样本集合中获取多个校准样本,将所述多个校准样本分别作为输入提供给所述量化后神经网络模型,经由所述量化后神经网络模型对输入的多个校准样本进行处理,得到所述量化后神经网络模型的输出;

根据所述量化后神经网络模型的输出与所述待量化神经网络模型的输出之间的余弦距离,调整所述任一待量化网络层的待量化权重矩阵;其中所述待量化神经网络模型的输出为将所述多个校准样本分别作为输入,提供给所述待量化神经网络模型,经由所述待量化神经网络模型对输入的多个校准样本进行处理得到的输出。

4.根据权利要求1-3任一所述的方法,其特征在于,所述对所述任一待量化网络层的权重矩阵进行矩阵变换,得到所述任一待量化网络层的待量化权重矩阵之后,还包括:

将矩阵变换次数初始化为预设值;

随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换,得到所述任一待量化网络层的新的待量化权重矩阵;对所述任一待量化网络层的新的待量化权重矩阵进行量化,得到所述任一待量化网络层的新的量化后权重矩阵,并基于所述任一待量化网络层的新的量化后权重矩阵,得到新的量化后神经网络模型;

确定所述新的量化后神经网络模型的输出与所述待量化神经网络模型的输出之间的余弦距离,并确定所述余弦距离是否小于预设余弦距离,以及确定所述矩阵变换次数是否小于预设阈值;

若所述余弦距离小于所述预设余弦距离且所述矩阵变换次数小于所述预设阈值,基于所述余弦距离和所述预设余弦距离确定目标概率,并确定所述目标概率是否大于预设概率;

若所述目标概率大于所述预设概率，将所述余弦距离作为预设余弦距离，所述矩阵变换次数加1，并执行所述随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的操作；

若所述目标概率不大于所述预设概率将所述任一待量化网络层的新的待量化权重矩阵作为所述任一待量化网络层的待量化权重矩阵。

5.根据权利要求4所述的方法，其特征在于，还包括：

若所述余弦距离小于所述预设余弦距离且所述矩阵变换次数不小于所述预设阈值，将所述任一待量化网络层的新的待量化权重矩阵作为所述任一待量化网络层的待量化权重矩阵；

若所述余弦距离不小于所述预设余弦距离且所述矩阵变换次数小于所述预设阈值，将所述余弦距离作为预设余弦距离，所述矩阵交换次数加1，并执行所述随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的操作。

6.根据权利要求4或5所述的方法，其特征在于，所述对所述任一待量化网络层的权重矩阵进行矩阵变换，得到所述任一待量化网络层的待量化权重矩阵，包括：

确定所述任一待量化网络层的联合网络层，所述联合网络层为所述待量化神经网络模型中的除所述任一待量化网络层以外的其它任一待量化网络层，所述联合网络层以所述任一待量化网络层的输出作为输入；

确定所述任一待量化网络层的权重矩阵的第一变换矩阵，并利用所述第一变换矩阵对所述任一待量化网络层的权重矩阵进行矩阵变换得到所述任一待量化网络层的待量化权重矩阵；

将所述第一变换矩阵的转置矩阵作为所述联合网络层的权重矩阵的第二变换矩阵，并利用所述第二变换矩阵对所述联合网络层的权重矩阵进行矩阵变换得到所述联合网络层的待量化权重矩阵。

7.根据权利要求6所述的方法，其特征在于，所述确定所述任一待量化网络层的权重矩阵的第一变换矩阵，包括：

基于所述联合网络层的输入通道与所述任一待量化网络层的输出通道之间的对应关系，确定所述任一待量化网络层的权重矩阵中各输出通道的权重数据与所述联合网络层的权重矩阵中各输入通道的权重数据的对应关系；

确定所述各输出通道的权重数据的分布特征值，以及所述各输入通道的权重数据的分布特征值，并将所述各输出通道的权重数据的分布特征值与对应的各输入通道的权重数据的分布特征值分别相乘，得到所述各输出通道的权重数据的目标特征值；

按照所述各输出通道的权重数据的目标特征值由小到大的顺序，对所述任一待量化网络层的权重矩阵中各输出通道的权重数据进行位置变换，得到所述任一待量化网络层的待量化权重矩阵，并将所述任一待量化网络层的待量化权重矩阵相对于所述任一待量化网络层的权重矩阵的变换矩阵作为所述第一变换矩阵。

8.一种神经网络模型的量化装置，其特征在于，包括：

获取模块，用于分别针对待量化神经网络模型中的任一待量化网络层，获取所述任一待量化网络层的权重矩阵；

矩阵变换模块，用于对所述任一待量化网络层的权重矩阵进行矩阵变换，得到所述任

一待量化网络层的待量化权重矩阵；

量化模块，用于对所述任一待量化网络层的待量化权重矩阵进行量化，得到所述任一待量化网络层的量化后权重矩阵；

第一得到模块，用于基于所述待量化神经网络模型中的任一待量化网络层的量化后权重矩阵，得到量化后神经网络模型。

9.根据权利要求8所述的装置，其特征在于，所述量化模块，包括：

划分单元，用于根据存算一体加速器中的存算单元阵列的大小，将所述任一待量化网络层的待量化权重矩阵进行划分，得到多个待量化子权重矩阵；

量化单元，用于分别对所述多个待量化子权重矩阵中的任一待量化子权重矩阵进行量化，得到多个量化后的子权重矩阵；

映射单元，用于将所述多个量化后的子权重矩阵分别映射到所述存算一体加速器中的多个存算单元阵列，由所述多个存算单元阵列存储所述多个量化后的子权重矩阵。

10.根据权利要求8或9所述的装置，其特征在于，还包括：

第二得到模块，用于从校准样本集合中获取多个校准样本，将所述多个校准样本分别作为输入提供给所述量化后神经网络模型，经由所述量化后神经网络模型对输入的多个校准样本进行处理，得到所述量化后神经网络模型的输出；

调整模块，用于根据所述量化后神经网络模型的输出与所述待量化神经网络模型的输出之间的余弦距离，调整所述任一待量化网络层的待量化权重矩阵；其中所述待量化神经网络模型的输出为将所述多个校准样本分别作为输入，提供给所述待量化神经网络模型，经由所述待量化神经网络模型对输入的多个校准样本进行处理得到的输出。

11.根据权利要求8-10任一所述的装置，其特征在于，所述矩阵变换模块，还用于：

将矩阵变换次数初始化为预设值；

随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换，得到所述任一待量化网络层的新的待量化权重矩阵；对所述任一待量化网络层的新的待量化权重矩阵进行量化，得到所述任一待量化网络层的新的量化后权重矩阵，并基于所述任一待量化网络层的新的量化后权重矩阵，得到新的量化后神经网络模型；

确定所述新的量化后神经网络模型的输出与所述待量化神经网络模型的输出之间的余弦距离，并确定所述余弦距离是否小于预设余弦距离，以及确定所述矩阵变换次数是否小于预设阈值；

若所述余弦距离小于所述预设余弦距离且所述矩阵变换次数小于所述预设阈值，基于所述余弦距离和所述预设余弦距离确定目标概率，并确定所述目标概率是否大于预设概率；

若所述目标概率大于所述预设概率，将所述余弦距离作为预设余弦距离，所述矩阵变换次数加1，并执行所述随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的操作；

若所述目标概率不大于所述预设概率将所述任一待量化网络层的新的待量化权重矩阵作为所述任一待量化网络层的待量化权重矩阵。

12.根据权利要求11所述的装置，其特征在于，所述矩阵变换模块，还用于：

若所述余弦距离小于所述预设余弦距离且所述矩阵变换次数不小于所述预设阈值，将

所述任一待量化网络层的新的待量化权重矩阵作为所述任一待量化网络层的待量化权重矩阵；

若所述余弦距离不小于所述预设余弦距离且所述矩阵变换次数小于所述预设阈值，将所述余弦距离作为预设余弦距离，所述矩阵交换次数加1，并执行所述随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的操作。

13.根据权利要求11或12所述的装置，其特征在于，所述矩阵变换模块，包括：

确定单元，用于确定所述任一待量化网络层的联合网络层，所述联合网络层为所述待量化神经网络模型中的除所述任一待量化网络层以外的其它任一待量化网络层，所述联合网络层以所述任一待量化网络层的输出作为输入；

第一变换单元，用于确定所述任一待量化网络层的权重矩阵的第一变换矩阵，并利用所述第一变换矩阵对所述任一待量化网络层的权重矩阵进行矩阵变换得到所述任一待量化网络层的待量化权重矩阵；

第二变换单元，用于将所述第一变换矩阵的转置矩阵作为所述联合网络层的权重矩阵的第二变换矩阵，并利用所述第二变换矩阵对所述联合网络层的权重矩阵进行矩阵变换得到所述联合网络层的待量化权重矩阵。

14.根据权利要求13所述的装置，其特征在于，所述确定所述任一待量化网络层的权重矩阵的第一变换矩阵，包括：

基于所述联合网络层的输入通道与所述任一待量化网络层的输出通道之间的对应关系，确定所述任一待量化网络层的权重矩阵中各输出通道的权重数据与所述联合网络层的权重矩阵中各输入通道的权重数据的对应关系；

确定所述各输出通道的权重数据的分布特征值，以及所述各输入通道的权重数据的分布特征值，并将所述各输出通道的权重数据的分布特征值与对应的各输入通道的权重数据的分布特征值分别相乘，得到所述各输出通道的权重数据的目标特征值；

按照所述各输出通道的权重数据的目标特征值由小到大的顺序，对所述任一待量化网络层的权重矩阵中各输出通道的权重数据进行位置变换，得到所述任一待量化网络层的待量化权重矩阵，并将所述任一待量化网络层的待量化权重矩阵相对于所述任一待量化网络层的权重矩阵的变换矩阵作为所述第一变换矩阵。

15.一种神经网络模型的量化系统，其特征在于，包括存算一体加速器和权利要求8-14任一所述的神经网络模型的量化装置。

16.一种电子设备，其特征在于，包括：

存储器，用于存储计算机程序；

处理器，用于执行所述存储器中存储的计算机程序，且所述计算机程序被执行时，实现上述权利要求1-7任一所述的方法。

17.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该计算机程序被处理器执行时，实现上述权利要求1-7任一所述的方法。

神经网络模型的量化方法、装置和系统、电子设备和存储介质

技术领域

[0001] 本公开涉及人工智能技术，尤其是一种神经网络模型的量化方法、装置和系统、电子设备和存储介质。

背景技术

[0002] 随着人工智能的快速发展，人工神经网络的应用越来越广泛。在人工神经网络中主要操作为矩阵向量乘法操作，如卷积层和全连接层等都是矩阵向量乘法操作。存算一体神经网络加速器将计算单元整合到存储单元中，可以高效运行矩阵向量乘法，从而在很大程度上减少了计算单元与存储单元之间频繁的数据交互，同时也可以大幅减少中间数据与片外主存的数据交互。因此使用存算一体技术是未来神经网络加速器很有潜力的发展方向。

[0003] 存算一体结构使用的存算单元阵列(crossbar)有固定大小。例如128×128的大小的存算单元阵列可以支持最多128个输入与一个128×128的矩阵的乘法。一个比存算单元阵列的大小更大的矩阵需要映射到多个存算单元阵列上，由每个存算单元阵列执行矩阵向量乘法运算的一部分运算。由于几乎所有神经网络加速器都采用低比特的权重和激活值进行运算，因此神经网络需要被量化。通过量化算法将神经网络的权重和激活值从浮点数转化为较低比特的定点数，量化后的比特数越低，运行神经网络的能量开销以及存储神经网络的存储开销越低，但也会降低神经网络的预测性能。

[0004] 在实现本公开的过程中，发明人发现，现有以神经网络的层为粒度进行量化的量化算法，每层神经网络的权重共享一套量化参数，由于神经网络的同一层中不同通道的权重的分布一般会有很大差异，使用一套量化参数会让分布范围较小的通道有较大量化误差，如果使用较低量化比特数，神经网络的预测性能会有较大损失；现有以神经网络的通道为粒度进行量化的量化算法，每层神经网络的一个输出通道权重共享一套量化参数，由于每个通道的量化参数不同，细粒度地操作不同通道会增加硬件的设计开销、降低硬件的运行效率；现有以存算单元阵列为粒度进行量化的量化算法，神经网络的每层映射到多个存算单元阵列，每个存算单元阵列的权重共享一套量化参数，由于每个存算单元阵列中存储不同通道的权重，它们的分布也会有较大差异，使用一套量化参数也会有量化误差。

发明内容

[0005] 为了解决上述技术问题，提出了本公开。本公开的实施例提供一种神经网络模型的量化方法、装置和系统、电子设备和存储介质。

[0006] 本公开实施例的一个方面，提供一种神经网络模型的量化方法，包括：分别针对待量化神经网络模型中的任一待量化网络层，获取所述任一待量化网络层的权重矩阵；对所述任一待量化网络层的权重矩阵进行矩阵变换，得到所述任一待量化网络层的待量化权重矩阵；对所述任一待量化网络层的待量化权重矩阵进行量化，得到所述任一待量化网络层的量化后权重矩阵；基于所述待量化神经网络模型中的任一待量化网络层的量化后权重矩

阵，得到量化后神经网络模型。

[0007] 可选地，在本公开上述任一方法实施例中，所述对所述任一待量化网络层的待量化权重矩阵进行量化，得到所述任一待量化网络层的量化后权重矩阵，包括：根据存算一体加速器中的存算单元阵列的大小，将所述任一待量化网络层的待量化权重矩阵进行划分，得到多个待量化子权重矩阵；分别对所述多个待量化子权重矩阵中的任一待量化子权重矩阵进行量化，得到多个量化后的子权重矩阵；将所述多个量化后的子权重矩阵分别映射到所述存算一体加速器中的多个存算单元阵列，由所述多个存算单元阵列存储所述多个量化后的子权重矩阵。

[0008] 可选地，在本公开上述任一方法实施例中，所述基于所述待量化神经网络模型中的任一待量化网络层的量化后权重矩阵，得到量化后神经网络模型之后，还包括：从校准样本集合中获取多个校准样本，将所述多个校准样本分别作为输入提供给所述量化后神经网络模型，经由所述量化后神经网络模型对输入的多个校准样本进行处理，得到所述量化后神经网络模型的输出；根据所述量化后神经网络模型的输出与所述待量化神经网络模型的输出之间的余弦距离，调整所述任一待量化网络层的待量化权重矩阵；其中所述待量化神经网络模型的输出为将所述多个校准样本分别作为输入，提供给所述待量化神经网络模型，经由所述待量化神经网络模型对输入的多个校准样本进行处理得到的输出。

[0009] 可选地，在本公开上述任一方法实施例中，所述对所述任一待量化网络层的权重矩阵进行矩阵变换，得到所述任一待量化网络层的待量化权重矩阵之后，还包括：将矩阵变换次数初始化为预设值；随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换，得到所述任一待量化网络层的新的待量化权重矩阵；对所述任一待量化网络层的新的待量化权重矩阵进行量化，得到所述任一待量化网络层的新的量化后权重矩阵，并基于所述任一待量化网络层的新的量化后权重矩阵，得到新的量化后神经网络模型；确定所述新的量化后神经网络模型的输出与所述待量化神经网络模型的输出之间的余弦距离，并确定所述余弦距离是否小于预设余弦距离，以及确定所述矩阵变换次数是否小于预设阈值；若所述余弦距离小于所述预设余弦距离且所述矩阵变换次数小于所述预设阈值，基于所述余弦距离和所述预设余弦距离确定目标概率，并确定所述目标概率是否大于预设概率；若所述目标概率大于所述预设概率，将所述余弦距离作为预设余弦距离，所述矩阵变换次数加1，并执行所述随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的操作；若所述目标概率不大于所述预设概率将所述任一待量化网络层的新的待量化权重矩阵作为所述任一待量化网络层的待量化权重矩阵。

[0010] 可选地，在本公开上述任一方法实施例中，还包括：若所述余弦距离小于所述预设余弦距离且所述矩阵变换次数不小于所述预设阈值，将所述任一待量化网络层的新的待量化权重矩阵作为所述任一待量化网络层的待量化权重矩阵；若所述余弦距离不小于所述预设余弦距离且所述矩阵变换次数小于所述预设阈值，将所述余弦距离作为预设余弦距离，所述矩阵交换次数加1，并执行所述随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的操作。

[0011] 可选地，在本公开上述任一方法实施例中，所述对所述任一待量化网络层的权重矩阵进行矩阵变换，得到所述任一待量化网络层的待量化权重矩阵，包括：确定所述任一待量化网络层的联合网络层，所述联合网络层为所述待量化神经网络模型中的除所述任一待

量化网络层以外的其它任一待量化网络层，所述联合网络层以所述任一待量化网络层的输出作为输入；确定所述任一待量化网络层的权重矩阵的第一变换矩阵，并利用所述第一变换矩阵对所述任一待量化网络层的权重矩阵进行矩阵变换得到所述任一待量化网络层的待量化权重矩阵；将所述第一变换矩阵的转置矩阵作为所述联合网络层的权重矩阵的第二变换矩阵，并利用所述第二变换矩阵对所述联合网络层的权重矩阵进行矩阵变换得到所述联合网络层的待量化权重矩阵。

[0012] 可选地，在本公开上述任一方法实施例中，所述确定所述任一待量化网络层的权重矩阵的第一变换矩阵，包括：基于所述联合网络层的输入通道与所述任一待量化网络层的输出通道之间的对应关系，确定所述任一待量化网络层的权重矩阵中各输出通道的权重数据与所述联合网络层的权重矩阵中各输入通道的权重数据的对应关系；确定所述各输出通道的权重数据的分布特征值，以及所述各输入通道的权重数据的分布特征值，并将所述各输出通道的权重数据的分布特征值与对应的各输入通道的权重数据的分布特征值分别相乘，得到所述各输出通道的权重数据的目标特征值；按照所述各输出通道的权重数据的目标特征值由小到大的顺序，对所述任一待量化网络层的权重矩阵中各输出通道的权重数据进行位置变换，得到所述任一待量化网络层的待量化权重矩阵，并将所述任一待量化网络层的待量化权重矩阵相对于所述任一待量化网络层的权重矩阵的变换矩阵作为所述第一变换矩阵。

[0013] 本公开实施例的另一方面，提供一种神经网络模型的量化装置，包括：获取模块，用于分别针对待量化神经网络模型中的任一待量化网络层，获取所述任一待量化网络层的权重矩阵；矩阵变换模块，用于对所述任一待量化网络层的权重矩阵进行矩阵变换，得到所述任一待量化网络层的待量化权重矩阵；量化模块，用于对所述任一待量化网络层的待量化权重矩阵进行量化，得到所述任一待量化网络层的量化后权重矩阵；第一得到模块，用于基于所述待量化神经网络模型中的任一待量化网络层的量化后权重矩阵，得到量化后神经网络模型。

[0014] 可选地，在本公开上述任一装置实施例中，所述量化模块，包括：划分单元，用于根据存算一体加速器中的存算单元阵列的大小，将所述任一待量化网络层的待量化权重矩阵进行划分，得到多个待量化子权重矩阵；量化单元，用于分别对所述多个待量化子权重矩阵中的任一待量化子权重矩阵进行量化，得到多个量化后的子权重矩阵；映射单元，用于将所述多个量化后的子权重矩阵分别映射到所述存算一体加速器中的多个存算单元阵列，由所述多个存算单元阵列存储所述多个量化后的子权重矩阵。

[0015] 可选地，在本公开上述任一装置实施例中，还包括：第二得到模块，用于从校准样本集合中获取多个校准样本，将所述多个校准样本分别作为输入提供给所述量化后神经网络模型，经由所述量化后神经网络模型对输入的多个校准样本进行处理，得到所述量化后神经网络模型的输出；调整模块，用于根据所述量化后神经网络模型的输出与所述待量化神经网络模型的输出之间的余弦距离，调整所述任一待量化网络层的待量化权重矩阵；其中所述待量化神经网络模型的输出为将所述多个校准样本分别作为输入，提供给所述待量化神经网络模型，经由所述待量化神经网络模型对输入的多个校准样本进行处理得到的输出。

[0016] 可选地，在本公开上述任一装置实施例中，所述矩阵变换模块，还用于：将矩阵变

换次数初始化为预设值;随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换,得到所述任一待量化网络层的新的待量化权重矩阵;对所述任一待量化网络层的新的待量化权重矩阵进行量化,得到所述任一待量化网络层的新的量化后权重矩阵,并基于所述任一待量化网络层的新的量化后权重矩阵,得到新的量化后神经网络模型;确定所述新的量化后神经网络模型的输出与所述待量化神经网络模型的输出之间的余弦距离,并确定所述余弦距离是否小于预设余弦距离,以及确定所述矩阵变换次数是否小于预设阈值;若所述余弦距离小于所述预设余弦距离且所述矩阵变换次数小于所述预设阈值,基于所述余弦距离和所述预设余弦距离确定目标概率,并确定所述目标概率是否大于预设概率;若所述目标概率大于所述预设概率,将所述余弦距离作为预设余弦距离,所述矩阵变换次数加1,并执行所述随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的操作;若所述目标概率不大于所述预设概率将所述任一待量化网络层的新的待量化权重矩阵作为所述任一待量化网络层的待量化权重矩阵。

[0017] 可选地,在本公开上述任一装置实施例中,所述矩阵变换模块,还用于:若所述余弦距离小于所述预设余弦距离且所述矩阵变换次数不小于所述预设阈值,将所述任一待量化网络层的新的待量化权重矩阵作为所述任一待量化网络层的待量化权重矩阵;若所述余弦距离不小于所述预设余弦距离且所述矩阵变换次数小于所述预设阈值,将所述余弦距离作为预设余弦距离,所述矩阵交换次数加1,并执行所述随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的操作。

[0018] 可选地,在本公开上述任一装置实施例中,所述矩阵变换模块,包括:确定单元,用于确定所述任一待量化网络层的联合网络层,所述联合网络层为所述待量化神经网络模型中的除所述任一待量化网络层以外的其它任一待量化网络层,所述联合网络层以所述任一待量化网络层的输出作为输入;第一变换单元,用于确定所述任一待量化网络层的权重矩阵的第一变换矩阵,并利用所述第一变换矩阵对所述任一待量化网络层的权重矩阵进行矩阵变换得到所述任一待量化网络层的待量化权重矩阵;第二变换单元,用于将所述第一变换矩阵的转置矩阵作为所述联合网络层的权重矩阵的第二变换矩阵,并利用所述第二变换矩阵对所述联合网络层的权重矩阵进行矩阵变换得到所述联合网络层的待量化权重矩阵。

[0019] 可选地,在本公开上述任一装置实施例中,所述确定所述任一待量化网络层的权重矩阵的第一变换矩阵,包括:基于所述联合网络层的输入通道与所述任一待量化网络层的输出通道之间的对应关系,确定所述任一待量化网络层的权重矩阵中各输出通道的权重数据与所述联合网络层的权重矩阵中各输入通道的权重数据的对应关系;确定所述各输出通道的权重数据的分布特征值,以及所述各输入通道的权重数据的分布特征值,并将所述各输出通道的权重数据的分布特征值与对应的各输入通道的权重数据的分布特征值分别相乘,得到所述各输出通道的权重数据的目标特征值;按照所述各输出通道的权重数据的目标特征值由小到大的顺序,对所述任一待量化网络层的权重矩阵中各输出通道的权重数据进行位置变换,得到所述任一待量化网络层的待量化权重矩阵,并将所述任一待量化网络层的待量化权重矩阵相对于所述任一待量化网络层的权重矩阵的变换矩阵作为所述第一变换矩阵。

[0020] 本公开实施例的再一方面,提供一种神经网络模型的量化系统,包括存算一体加速器和本公开上述任一实施例所述的神经网络模型的量化装置。

[0021] 根据本公开实施例的又一方面，提供一种电子设备，所述电子设备包括：

[0022] 处理器；

[0023] 用于存储所述处理器可执行指令的存储器；

[0024] 所述处理器，用于从所述存储器中读取所述可执行指令，并执行所述指令以实现本公开上述任一实施例所述的神经网络的量化方法。

[0025] 根据本公开实施例的再一方面，提供一种计算机可读存储介质，所述存储介质存储有计算机程序，所述计算机程序用于执行本公开上述任一实施例所述的神经网络的量化方法。

[0026] 基于本公开上述实施例提供的神经网络模型的量化方法、装置和系统、电子设备和介质，通过对待量化神经网络模型中的任一待量化网络层的行重排和/或列重排后的权重矩阵进行量化，能够降低权重矩阵中各通道的权重数据的分布差异，从而可以减少量化误差，有助于提升量化后神经网络的精度。

[0027] 下面通过附图和实施例，对本公开的技术方案做进一步的详细描述。

附图说明

[0028] 构成说明书的一部分的附图描述了本公开的实施例，并且连同描述一起用于解释本公开的原理。

[0029] 参照附图，根据下面的详细描述，可以更加清楚地理解本公开，其中：

[0030] 图1为本公开神经网络模型的量化方法一个实施例的流程图。

[0031] 图2为本公开将权重矩阵映射到存算一体加速器的示意图。

[0032] 图3为本公开多层联合矩阵变换的示意图。

[0033] 图4为本公开神经网络模型的量化方法另一个实施例的流程图。

[0034] 图5为本公开神经网络模型的量化装置一个实施例的结构示意图。

[0035] 图6为本公开神经网络模型的量化装置另一个实施例的结构示意图。

[0036] 图7为本公开神经网络模型的量化系统一个实施例的结构示意图。

[0037] 图8为本公开电子设备一个应用实施例的结构示意图。

具体实施方式

[0038] 现在将参照附图来详细描述本公开的各种示例性实施例。应注意到：除非另外具体说明，否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本公开的范围。

[0039] 本领域技术人员可以理解，本公开实施例中的“第一”、“第二”等术语仅用于区别不同步骤、设备或模块等，既不代表任何特定技术含义，也不表示它们之间的必然逻辑顺序。

[0040] 还应理解，在本公开实施例中，“多个”可以指两个或两个以上，“至少一个”可以指一个、两个或两个以上。

[0041] 还应理解，对于本公开实施例中提及的任一部件、数据或结构，在没有明确限定或者在前后文给出相反启示的情况下，一般可以理解为一个或多个。

[0042] 另外，本公开中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存

在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。另外，本公开中字符“/”，一般表示前后关联对象是一种“或”的关系。

[0043] 还应理解，本公开对各个实施例的描述着重强调各个实施例之间的不同之处，其相同或相似之处可以相互参考，为了简洁，不再一一赘述。

[0044] 同时，应当明白，为了便于描述，附图中所示出的各个部分的尺寸并不是按照实际的比例关系绘制的。

[0045] 以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本公开及其应用或使用的任何限制。

[0046] 对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论，但在适当情况下，所述技术、方法和设备应当被视为说明书的一部分。

[0047] 应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步讨论。

[0048] 本公开实施例可以应用于终端设备、计算机系统、服务器等电子设备，其可与众多其它通用或专用计算系统环境或配置一起操作。适于与终端设备、计算机系统、服务器等电子设备一起使用的众所周知的终端设备、计算系统、环境和/或配置的例子包括但不限于：个人计算机系统、服务器计算机系统、瘦客户机、厚客户机、手持或膝上设备、基于微处理器的系统、机顶盒、可编程消费电子产品、网络个人电脑、小型计算机系统、大型计算机系统和包括上述任何系统的分布式云计算技术环境，等等。

[0049] 终端设备、计算机系统、服务器等电子设备可以在由计算机系统执行的计算机系统可执行指令(诸如程序模块)的一般语境下描述。通常，程序模块可以包括例程、程序、目标程序、组件、逻辑、数据结构等等，它们执行特定的任务或者实现特定的抽象数据类型。计算机系统/服务器可以在分布式云计算环境中实施，分布式云计算环境中，任务是由通过通信网络链接的远程处理设备执行的。在分布式云计算环境中，程序模块可以位于包括存储设备的本地或远程计算系统存储介质上。

[0050] 图1是本公开神经网络模型的量化方法一个实施例的流程图。如图1所示，本实施例包括如下步骤：

[0051] 步骤102，分别针对待量化神经网络模型中的任一待量化网络层，获取任一待量化网络层的权重矩阵。

[0052] 本公开实施例中，待量化神经网络模型可以为未预先训练的神经网络模型，也可以为预选训练的神经网络模型，本公开实施例不做限定。在待量化神经网络模型为预先训练得到的神经网络模型的情况下，本公开实施例还可以预先通过训练数据对建立的神经网络模型进行训练，获得预先训练的神经网络模型作为待量化神经网络模型，具体训练过程不再赘述。

[0053] 本公开实施例中，待量化神经网络模型可以为LeNet、AlexNet、VggNet和ResNet等卷积神经网络中的任一项，待量化神经网络模型的网络结构可以包括输入层、隐藏从和输出层，其中的隐藏层可以包括但不限于卷积层、激励层、池化层和全连接层，本公开实施例对待量化神经网络模型的类型、网络结构不做限定，对待量化神经网络模型中卷积层的数量和全连接层的数量也不做限定。

[0054] 本公开实施例中的待量化网络层可以为待量化神经网络模型中输入层和输出层

外的网络层中的任意一个或多个网络层，例如任一卷积层或任一全连接层，本公开实施例对待量化神经网络模型中的待量化网络层的类型和层数不做限定。

[0055] 步骤104，对任一待量化网络层的权重矩阵进行矩阵变换，得到任一待量化网络层的待量化权重矩阵。

[0056] 本公开实施例中，任一待量化网络层的权重矩阵中的一行权重数据，可以对应该待量化网络层中的某个神经元针对各输入的权重，也可以对应该量化神经网络层中的各神经元针对某个输入的权重。相应地，任一待量化网络层的权重矩阵中的一列权重数据，可以对应该量化神经网络层中的各神经元针对某个输入的权重，也可以对应该待量化网络层中的某个神经元针对各输入的权重。

[0057] 本公开实施例中的矩阵变换(也可以称为矩阵重排)可以为对权重矩阵的行变换(也可以称为行重排)，和/或列变换(也可以成为列重排)，其中的行变换可以为对权重矩阵中的行进行位置变换，列变换可以为对权重矩阵中的列进行位置变换。

[0058] 在一个可选示例中，可以随机选取任一待量化网络层的权重矩阵中的行，和/或列进行变换。

[0059] 在一个可选示例中，可以根据任一待量化网络层的权重矩阵中的各行权重数据的分布差异确定对权重矩阵的行变换方式，可以根据任一待量化网络层的权重矩阵中的各列权重数据的分别差异确定对权重矩阵的列变换方式。通过对权重矩阵的行变换，和/或列变换，可以减少权重矩阵中各行权重数据和各列权重数据的分布差异，从而可以提高网络模型量化精度。

[0060] 步骤106，对任一待量化网络层的待量化权重矩阵进行量化，得到任一待量化网络层的量化后权重矩阵。

[0061] 通过量化算法，可以将神经网络的权重数据从浮点数转化为较低比特的定点数。本公开实施例中，通过对待量化权重矩阵进行量化得到量化后权重矩阵，量化后权重矩阵中的权重数据可以被量化成较低比特的定点数据。在对待量化权重矩阵进行量化时，可以采用对称均匀量化的形式，其量化函数的可以为：

[0062] $$x_{int} = \text{clamp}(\text{round}\left(\frac{x}{\Delta}\right), -2^{k-1}, 2^{k-1}-1) \quad (1)$$

[0063] 其中，x_{int}是量化后的定点数据，x是量化前的浮点数据，Δ是量化的缩放系数，k是量化的比特数，k比特量化后的数值范围为$[-2^{k-1}, 2^{k-1}-1]$的整数。

[0064] 其中，round()函数用于返回浮点数$\frac{x}{\Delta}$的四舍五入值，clamp()函数用于将数值限制在一个给定的区间[min,max]内，小于最小值min的数值可以映射到min，大于最大值max的数值可以映射到max。

[0065] 本公开实施例中，待量化神经网络模型的量化粒度可以决定哪些数据共享一个量化缩放系数。若以层为量化粒度进行量化，每个层的权重数据共享一个缩放系数；若以输出通道为量化粒度进行量化，每个输出通道的权重数据共享一个缩放系数。

[0066] 步骤108，基于待量化神经网络模型中的任一待量化网络层的量化后权重矩阵，得到量化后神经网络模型。

[0067] 本公开实施例中，可以分别对待量化神经网络模型中的多个待量化网络层中的任

一待量化网络层进行量化，并可以在完成多个待量化网络层的量化后，得到量化后神经网络模型。

[0068] 本公开实施例的发明人，可以使用ResNet-20网络作为待量化神经网络模型，使用CIFAR-10数据集作为测试集合，随机对各待量化网络层的行列进行重排，对各待量化网络层使用4-bit量化(量化工具可以为EasyQuant)后，统计量化后神经网络模型的准确率，量化后的神经网络的准确率根据重排不同有较大变化。最优重排的准确率与最差重排的准确率差异高达2%。因此，进行权重重排可较大程度提升量化模型后的神经网络的准确率。

[0069] 基于本公开上述实施例提供的神经网络模型的量化方法，通过对待量化神经网络模型中的任一待量化网络层的行重排和/或列重排后的权重矩阵进行量化，能够降低权重矩阵中各行权重数据和各列权重数据的分布差异，从而可以减少量化误差，有助于提升量化后神经网络的精度。

[0070] 可选地，在本公开其中一些可能的实现方式中，上述步骤106中，在对任一待量化网络层的待量化权重矩阵进行量化，得到任一待量化网络层的量化后权重矩阵时，可以根据存算一体加速器中的存算单元阵列的大小，将任一待量化网络层的待量化权重矩阵进行划分，得到多个待量化子权重矩阵，然后分别对多个待量化子权重矩阵中的任一待量化子权重矩阵进行量化，得到多个量化后的子权重矩阵，并将多个量化后的子权重矩阵分别映射到存算一体加速器中的多个存算单元阵列，由多个存算单元阵列存储多个量化后的子权重矩阵。

[0071] 图2为本公开将权重矩阵映射到存算一体加速器的示意图。如图2所示，W为n×m的权重矩阵，箭头右侧的每个矩形为一个大小为R×C的存算单元阵列，在存算一体加速器中，权重矩阵W可以被映射到#R×#C个R×C存算单元阵列中，其中$\#R=\frac{n}{R},\ \#C=\frac{m}{C}$。

[0072] 基于此，本公开实施例可以根据存算一体加速器中的存算单元阵列的大小，将任一待量化网络层的待量化权重矩阵进行划分，得到多个待量化子权重矩阵，然后分别对多个待量化子权重矩阵中的任一待量化子权重矩阵进行量化，每个待量化子权重矩阵共享一个量化缩放系数，得到多个量化后的子权重矩阵，进而可以将多个量化后的子权重矩阵分别映射到存算一体加速器中的多个存算单元阵列，由多个存算单元阵列存储多个量化后的子权重矩阵。

[0073] 本公开实施例通过将待量化网络层的待量化权重矩阵按照存算单元阵列的大小划分为多个待量化子权重矩阵，然后分别对每个待量化权重矩阵进行量化，有助于减小每个存算单元阵列存储的权重数据的分布差异，有助于减少量化误差，提高网络模型量化精度。

[0074] 可选地，在本公开其中一些可能的实现方式中，上述步骤104之后还可以通过以下步骤优化待量化权重矩阵：

[0075] 步骤a，将矩阵变换次数初始化为预设值。

[0076] 本公开实施例中，矩阵变换次数的初始值为预设值，可以为0、1等数值，矩阵变换次数的最大值为预设阈值，可以根据实际需要进行设定，本公开实施例不做限定。

[0077] 步骤b，随机选取任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换，得到任一待量化网络层的新的待量化权重矩阵，进而可以对任一待量化网络层的新

的待量化权重矩阵进行量化，得到任一待量化网络层的新的量化后权重矩阵，并基于任一待量化网络层的新的量化后权重矩阵，得到新的量化后神经网络模型。

[0078] 本公开实施例中，对任一待量化网络层的新的待量化权重矩阵进行量化时，可以将任一待量化网络层的新的待量化权重矩阵映射到存算一体加速器中的多个存算单元阵列，由存算一体加速器针对多个存算单元阵列中的任一存算单元阵列存储的权重数据进行量化，以得到任一待量化网络层的新的量化后权重矩阵。

[0079] 步骤c，确定新的量化后神经网络模型的输出与待量化神经网络模型的输出之间的余弦距离，并确定余弦距离是否小于预设余弦距离，以及确定矩阵变换次数是否小于预设阈值。

[0080] 本公开实施例中，新的量化后神经网络模型的输出为将校准样本集合中的多个校准样本分别作为输入，提供给新的量化后神经网络模型，经由新的量化后神经网络模型对多个校准样本进行处理得到的输出，待量化神经网络模型的输出为将多个校准样本分别作为输入，提供给待量化神经网络模型，经由待量化神经网络模型对多个校准样本进行处理得到的输出。

[0081] 本公开实施例中，用于确定新的量化后神经网络模型的输出与待量化神经网络模型的输出之间的余弦距离的公式可以为：

[0082] $$cossim(Y_2,Y_1)=\frac{1}{N}\sum_{y_2\in Y_2,y_1\in Y_1}\frac{y_2\cdot y_1}{|y_2||y_1|} \quad (2)$$

[0083] 其中，Y_2新的量化后神经网络模型的输出，Y_1为待量化神经网络模型的输出。

[0084] 在利用该公式(2)确定了新的量化后神经网络模型的输出与待量化神经网络模型的输出之间的余弦距离之后，可以确定余弦距离是否小于预设余弦距离，以及确定矩阵变换次数是否小于预设阈值，其中的预设余弦距离可以根据实际需要进行确定，本公开实施例不做限定。

[0085] 步骤d，若余弦距离小于预设余弦距离且矩阵变换次数小于预设阈值，可以基于余弦距离和预设余弦距离确定目标概率，并确定目标概率是否大于预设概率。

[0086] 在本公开实施例中，用于基于余弦距离和预设余弦距离确定目标概率的公式可以为：

[0087] $$P=exp(\frac{cossim_i-cossim_{i-1}}{T}) \quad (3)$$

[0088] 其中exp()为以自然常数e为底的指数函数，P为目标概率，$cossim_i$为余弦距离，$cossim_{i-1}$为预设余弦距离，T为模拟退火算法的超参数。需要说明的是，T的取值可以根据实际需要进行设定，本公开实施例不做限定。

[0089] 步骤e，若目标概率大于预设概率，可以将余弦距离作为预设余弦距离，矩阵变换次数加1，并执行随机选取任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的操作。

[0090] 步骤f，若目标概率不大于预设概率，可以将任一待量化网络层的新的待量化权重矩阵作为任一待量化网络层的待量化权重矩阵。

[0091] 或者，在本公开另一些可能的实现方式中，上述优化待量化权重矩阵的步骤还可以包括：

[0092] 步骤g,若余弦距离小于预设余弦距离且矩阵变换次数不小于预设阈值,可以将任一待量化网络层的新的待量化权重矩阵作为任一待量化网络层的待量化权重矩阵。

[0093] 步骤h,若余弦距离不小于预设余弦距离且矩阵变换次数小于预设阈值,可以将余弦距离作为预设余弦距离,矩阵交换次数加1,并执行随机选取所述任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的操作。

[0094] 本公开实施例的发明人,可以使用ResNet-20网络作为待量化神经网络模型,使用CIFAR-10数据集作为测试集合,按照步骤a至步骤h对各待量化网络层的待量化权重矩阵进行优化,对各待量化网络层使用4-bit量化(量化工具可以为EasyQuant)后,统计量化后神经网络模型的准确率,发现按照步骤a至步骤h对各待量化网络层的待量化权重矩阵进行优化,可以有效提升量化后神经网络模型与待量化神经网络模型的cosine相似度,并能最终达到一个较高的神经网络模型量化精度。

[0095] 本公开实施例通过对随机选取任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的方式,进一步调整待量化权重矩阵中的行权重数据或列权重数据的分布差异,并通过新的量化后神经网络模型的输出与待量化神经网络模型的输出之间的余弦距离以及矩阵变换次数,确定是否可以将新的待量化权重矩阵作为任一待量化网络层的待量化权重矩阵,能够进一步减小权重矩阵中行权重数据和列权重数据的分布差异,提高网络模型量化精度。

[0096] 可选地,在本公开其中一些可能的实现方式中,上述步骤104中,在对任一待量化网络层的权重矩阵进行矩阵变换,得到任一待量化网络层的待量化权重矩阵时,可以确定任一待量化网络层的联合网络层,该联合网络层为待量化神经网络模型中的除任一待量化网络层以外的其它任一待量化网络层,联合网络层以任一待量化网络层的输出作为输入,然后可以确定任一待量化网络层的权重矩阵的第一变换矩阵,并利用第一变换矩阵对任一待量化网络层的权重矩阵进行矩阵变换得到任一待量化网络层的待量化权重矩阵,进而可以将第一变换矩阵的转置矩阵作为联合网络层的权重矩阵的第二变换矩阵,并利用第二变换矩阵对联合网络层的权重矩阵进行矩阵变换得到联合网络层的待量化权重矩阵。

[0097] 本公开实施例中,任一待量化网络层的权重矩阵中的一列权重数据,可以对应该待量化网络层中的某个神经元针对各输入的权重,任一待量化网络层的权重矩阵中的一行权重数据,可以对应该量化神经网络层中的各神经元针对某个输入的权重。可以将任一待量化网络层的权重矩阵中的一列权重数据作为一个输出通道的权重数据,将任意待量化网络层的权重矩阵中的一行权重数据作为一个输入通道的权重数据。权重矩阵的列交换可以让各输出通道的位置发生交换,权重矩阵的行交换可以让各输入通道的位置发生交换。

[0098] 本公开实施例中,若单独对任一待量化网络层的权重矩阵通过行变换进行行重排,或单独对任一待量化网络层的权重矩阵通过列交换进行列重排,在待量化网络层计算时,需要根据行重排的索引index获取对应位置的输入数据;待量化网络层的输出也需要根据列重排的索引index将计算结果存储在对应位置。这样的索引方式需要计算数据在内存中的位置,也会出现对存储单元的随机访存,产生较大开销。

[0099] 作为一个示例,对于待量化神经网络模型中的任一待量化网络层C1,C1的输入数据可以包括(d1,d2,d3),C1可以包括3个神经元(s1,s2,s3),那么C1的权重矩阵如下述表1所示:

[0100]

神经元 \ 输入	d1	d2	d3
s1	w11	w12	w13
s2	w21	w22	w23
s3	w23	w32	w33

[0101] 表1待量化网络层C1的权重矩阵

[0102] 对于该任一待量化网络层C1的联合网络层C2,C2以C1各神经元的输出作为输入,若C2包括3个神经元(t1,t2,t3),那么C2的权重矩阵如下述表2所示:

[0103]

神经元 \ 输入	s1-out	s2-out	s3-out
d1	u11	u12	u13
d2	u21	u22	u23
d3	u23	u32	u33

[0104] 表2联合网络层C2的权重矩阵

[0105] 对比C1和C2的权重矩阵,发现若将C1权重矩阵中的第一行和第二行进行了交换,就需要将C2权重矩阵中的第一列和第二列进行交换,以实现C2的输入通道与C1的输出通道的位置对应。

[0106] 基于此,本公开实施例中,在对任一待量化网络层进行量化时,可以将任一待量化网络层与,以该任一待量化网络层的输出作为输入的联合网络层进行联合矩阵变换(也可以称为联合重排)。由于联合网络层,以任一待量化网络层的输出作为输入,任一待量化网络层的输出通道与联合网络层的输入通道之间是对应的,因此可以同时调整任一待量化网络层的输出通道位置与联合网络层的输入通道位置。

[0107] 图3为本公开多层联合矩阵变换的示意图。如图3所示,W^l是第l层的权重矩阵,W^{l+1}是第l+1层的权重矩阵,f为激活函数,该激活函数f可以为常用的ReLU、PReLU、ReLU6等分段线性函数,本公开实施例不做限定。

[0108] 将网络中的批处理层(batch-normalization)计算融合到权重中,第l层的计算与第l+1层的计算分别为:

[0109] $X^{l+1}=f(W^lX^l)$ (4)

[0110] $X^{l+2}=f(W^{l+1}X^{l+1})$ (5)

[0111] 合并两层的计算公式为:

[0112] $X^{l+2}=f(W^{l+1}f(W^lX^l))$ (6)

[0113] 为了消除数据访存索引的开销,对第l层和第l+1层进行联合重排。使用矩阵S对第l层的列进行交换,得到W^lS,使用其转置矩阵S^T对l+1层的行进行交换,得到S^TW^{l+1}。如图3所示:

[0114] 图3中W^l的第i列与第j列进行交换,同时对W^{l+1}的第i行与第j行进行交换。

[0115] 使用重排后的权重直接进行计算,有:

[0116] $$\widehat{X^{l+2}}=f\left(S^TW^{l+1}f\left(W^lSX^l\right)\right) \quad (7)$$

[0117] 当激活函数f为常用的ReLU、PReLU、ReLU6等分段线性函数时,有以下等式:

[0118] $$\widehat{X^{l+2}} = f\left(S^T W^{l+1} f\left(W^l S X^l\right)\right) = f\left(W^{l+1} f\left(W^l X^l\right)\right) = X^{l+2} \quad (8)$$

[0119] 可见，联合重排后的计算结果$\widehat{X^{l+2}}$与联合重排前的计算结果X^{l+2}相同，可以避免数据访存索引的开销。

[0120] 在一个可选的示例中，在对第l层和第l+1层进行联合重排时，也可以使用矩阵S对第l层的行进行交换，得到SW^l，使用其转置矩阵S^T对l+1层的列进行交换，得到$W^{l+1}S^T$，具体联合重排过程不再赘述。

[0121] 本公开实施例的发明人，可以使用ResNet-20网络作为待量化神经网络模型，使用CIFAR-10数据集作为测试数据集，对待量化神经网络模型的各层使用联合随机重排，4-bit量化(量化工具可以为EasyQuant)后，统计模型预测准确率，发现联合重排的量化后神经网络的准确率与每层单独重排的量化后神经网络的准确率接近，多重联合重排相对于每层单独重排可以消除数据访存索引的开销并可以较大程度提升网络模模型的量化准确率。

[0122] 可选地，在本公开其中一些可能的实现方式中，上述步骤104中，在确定任一待量化网络层的权重矩阵的第一变换矩阵时，可以基于联合网络层的输入通道与任一待量化网络层的输出通道之间的对应关系，确定任一待量化网络层的权重矩阵中各输出通道的权重数据与联合网络层的权重矩阵中各输入通道的权重数据的对应关系，然后可以确定各输出通道的权重数据的分布特征值，以及各输入通道的权重数据的分布特征值，并将各输出通道的权重数据的分布特征值与对应的各输入通道的权重数据的分布特征值分别相乘，得到各输出通道的权重数据的目标特征值；进而可以按照各输出通道的权重数据的目标特征值由小到大的顺序，对任一待量化网络层的权重矩阵中各输出通道的权重数据进行位置变换，得到任一待量化网络层的待量化权重矩阵，并将任一待量化网络层的待量化权重矩阵相对于任一待量化网络层的权重矩阵的变换矩阵作为第一变换矩阵。

[0123] 本公开实施例中，权重数据的分布特征值可以为权重数据99.9%与0.1%分位数的范围。将权重数据按照由小对大的顺序进行排序，可以确定99.9%的权重数据对应的权重数值为99.9%分位数，确定0.1%的权重数据对应的权重数值位0.1%分位数。

[0124] 作为一个示例，设R_i^l为第l层网络的第i个输出通道的权重范围大小，R_i^{l+1}为第l+1层网络的第i个输入通道的权重范围大小。接下来可以根据$R_i^l \times R_i^{l+1}$对第l层网络的输出通道和第l+1层网络的输入通道进行排序，$R_i^l \times R_i^{l+1}$数值较小的输出通道或输入通道被重排到前面，$R_i^l \times R_i^{l+1}$数值较大的输出通道或输入通道被重排到后面。这样的排序考虑了两层网络的权重矩阵的权重数据分布，可以使重排之后第l层的权重矩阵和l+1层的权重矩阵在映射到存算一体加速器时，同一存算单元阵列存储的权重数据的分布更为接近，有助于减少量化误差，提高网络模型量化精度。

[0125] 图4为本公开神经网络模型的量化方法另一个实施例的流程图。如图4所示，在图1所示实施例的基础上，上述步骤108之后，上述方法还可以包括：

[0126] 步骤110，从校准样本集合中获取多个校准样本，将多个校准样本分别作为输入提供给量化后神经网络模型，经由量化后神经网络模型对输入的多个校准样本进行处理，得到量化后神经网络模型的输出。

[0127] 步骤112，根据量化后神经网络模型的输出与待量化神经网络模型的输出之间的余弦距离，调整任一待量化网络层的待量化权重矩阵。其中的待量化神经网络模型的输出为将多个校准样本分别作为输入，提供给待量化神经网络模型，经由待量化神经网络模型对输入的多个校准样本进行处理得到的输出。

[0128] 本公开实施例中，校准样本集合中的校准样本可以为不带标签的样本，因此无法基于校准样本的标签测量量化后神经网络模型的准确率。为了实现对量化后神经网络模型的准确率的测量，可以将量化后神经网络模型的输出与待量化神经网络模型的输出的余弦距离cosine距离作为测量量化后神经网络模型准确率的目标。

[0129] 神经网络模型量化的目的是让量化后神经网络模型的输出与待量化神经网络模型的输出尽可能接近，因此对权重矩阵的矩阵重排目标是最小化量化后神经网络模型输出与待量化神经网络模型输出的差异。可以定义矩阵重排目标：

[0130] $$\mathrm{argmax}_{S_{row},S_{col}}\ \mathrm{cossim}\left(Y_{quant},Y_{FP32}\right) \qquad (9)$$

[0131] 其中Y_{quant}是量化后神经网络模型的输出数值；Y_{FP32}是待量化神经网络模型的输出数值，FP32是采用4字节(32位)进行编码存储的一种数据类型；S_{row}，S_{col}是对权重矩阵进行矩阵变换的变换矩阵；cossim是计算不同校准输入平均cosine距离的函数，定义如下：

[0132] $$\mathrm{cossim}\left(Y_{quant},Y_{FP32}\right)=\frac{1}{N}\sum_{y_{quant}\in Y_{quant},y_{FP32}\in Y_{FP32}}\frac{y_{quant}\cdot y_{FP32}}{|y_{quant}||y_{FP32}|} \qquad (10)$$

[0133] 本公开实施例的发明人，可以使用ResNet-20网络作为待量化神经网络模型，使用CIFAR-10数据集作为校准样本集合，对待量化神经网络模型的各层使用4-bit量化(量化工具可以为EasyQuant)后，统计量化后神经网络模型的准确率，发现通过根据量化后神经网络模型的输出与待量化神经网络模型的输出之间的余弦距离，调整任一待量化网络层的待量化权重矩阵，使得量化后神经网络模型的输出与待量化网络模型的输出更接近，在保证量化后权重数据比特数较低的同时，可以提高量化后神经网络模型的预测准确度。

[0134] 图5为本公开神经网络模型的量化装置一个实施例的结构示意图。该实施例的神经网络模块的量化装置可用于实现本公开上述各神经网络模型的量化方法实施例。如图5所示，该实施例的装置可以包括：获取模块502，矩阵变换模块504，量化模块506和第一得到模块508。其中，

[0135] 获取模块502，用于分别针对待量化神经网络模型中的任一待量化网络层，获取任一待量化网络层的权重矩阵。

[0136] 矩阵变换模块504，用于对任一待量化网络层的权重矩阵进行矩阵变换，得到任一待量化网络层的待量化权重矩阵。

[0137] 量化模块506，用于对任一待量化网络层的待量化权重矩阵进行量化，得到任一待量化网络层的量化后权重矩阵。

[0138] 第一得到模块508，用于基于待量化神经网络模型中的任一待量化网络层的量化后权重矩阵，得到量化后神经网络模型。

[0139] 可选地，在本公开其中一些可能的实现方式中，上述量化模块506可以包括：划分单元、量化单元和传输单元。其中，划分单元用于根据存算一体加速器中的存算单元阵列的大小，将任一待量化网络层的待量化权重矩阵进行划分，得到多个待量化子权重矩阵；量化单元，用于分别对多个待量化子权重矩阵中的任一待量化子权重矩阵进行量化，得到多个

量化后的子权重矩阵;映射单元,用于将多个量化后的子权重矩阵分别映射到存算一体加速器中的多个存算单元阵列,由多个存算单元阵列存储多个量化后的子权重矩阵。

[0140] 可选地,在本公开其中一些可能的实现方式中,上述矩阵变换模块504还可以用于:将矩阵变换次数初始化为预设值;随机选取任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换,得到任一待量化网络层的新的待量化权重矩阵;对任一待量化网络层的新的待量化权重矩阵进行量化,得到任一待量化网络层的新的量化后权重矩阵,并基于任一待量化网络层的新的量化后权重矩阵,得到新的量化后神经网络模型;确定新的量化后神经网络模型的输出与待量化神经网络模型的输出之间的余弦距离,并确定余弦距离是否小于预设余弦距离,以及确定矩阵变换次数是否小于预设阈值;若余弦距离小于预设余弦距离且矩阵变换次数小于预设阈值,基于余弦距离和预设余弦距离确定目标概率,并确定目标概率是否大于预设概率;若目标概率大于预设概率,将余弦距离作为预设余弦距离,矩阵变换次数加1,并执行随机选取任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的操作;若目标概率不大于预设概率将任一待量化网络层的新的待量化权重矩阵作为任一待量化网络层的待量化权重矩阵。

[0141] 可选地,在本公开其中一些可能的实现方式中,上述矩阵变换模块504还可以用于:若余弦距离小于预设余弦距离且矩阵变换次数不小于预设阈值,将任一待量化网络层的新的待量化权重矩阵作为任一待量化网络层的待量化权重矩阵;若余弦距离不小于预设余弦距离且矩阵变换次数小于预设阈值,将余弦距离作为预设余弦距离,矩阵交换次数加1,并执行随机选取任一待量化网络层的待量化权重矩阵中的任意两行或两列进行交换的操作。

[0142] 可选地,在本公开其中一些可能的实现方式中,上述矩阵变换模块504可以包括:确定单元,第一变换单元和第二变换单元。其中,

[0143] 确定单元,用于确定任一待量化网络层的联合网络层,联合网络层为待量化神经网络模型中的除任一待量化网络层以外的其它任一待量化网络层,联合网络层以任一待量化网络层的输出作为输入;第一变换单元,用于确定任一待量化网络层的权重矩阵的第一变换矩阵,并利用第一变换矩阵对任一待量化网络层的权重矩阵进行矩阵变换得到任一待量化网络层的待量化权重矩阵;第二变换单元,用于将第一变换矩阵的转置矩阵作为联合网络层的权重矩阵的第二变换矩阵,并利用第二变换矩阵对联合网络层的权重矩阵进行矩阵变换得到联合网络层的待量化权重矩阵。

[0144] 可选地,在本公开其中一些可能的实现方式中,上述矩阵变换模块504确定任一待量化网络层的权重矩阵的第一变换矩阵时,可以基于联合网络层的输入通道与任一待量化网络层的输出通道之间的对应关系,确定任一待量化网络层的权重矩阵中各输出通道的权重数据与联合网络层的权重矩阵中各输入通道的权重数据的对应关系;确定各输出通道的权重数据的分布特征值,以及各输入通道的权重数据的分布特征值,并将各输出通道的权重数据的分布特征值与对应的各输入通道的权重数据的分布特征值分别相乘,得到各输出通道的权重数据的目标特征值;按照各输出通道的权重数据的目标特征值由小到大的顺序,对任一待量化网络层的权重矩阵中各输出通道的权重数据进行位置变换,得到任一待量化网络层的待量化权重矩阵,并将任一待量化网络层的待量化权重矩阵相对于任一待量化网络层的权重矩阵的变换矩阵作为第一变换矩阵。

[0145] 图6为本公开神经网络模型的量化装置另一个实施例的结构示意图。该另一个实施例的神经网络模型的量化装置，在图6所示神经网络模型的量化装置的基础上，还可以包括第二得到模块510和调整模块512。其中，

[0146] 第二得到模块510，用于从校准样本集合中获取多个校准样本，将多个校准样本分别作为输入提供给量化后神经网络模型，经由量化后神经网络模型对输入的多个校准样本进行处理，得到量化后神经网络模型的输出。

[0147] 调整模块512，用于根据量化后神经网络模型的输出与待量化神经网络模型的输出之间的余弦距离，调整任一待量化网络层的待量化权重矩阵。其中的待量化神经网络模型的输出为将多个校准样本分别作为输入，提供给待量化神经网络模型，经由待量化神经网络模型对输入的多个校准样本进行处理得到的输出。

[0148] 图7为本公开神经网络模型的量化系统一个实施例的结构示意图。该神经网络模型的量化系统可以包括存算一体加速器和本公开上述任一实施例所述的神经网络模型的量化装置。

[0149] 另外，本公开实施例还提供了一种电子设备，包括：

[0150] 存储器，用于存储计算机程序；

[0151] 处理器，用于执行所述存储器中存储的计算机程序，且所述计算机程序被执行时，实现本公开上述任一实施例所述的神经网络模型的量化方法。

[0152] 图8为本公开电子设备一个应用实施例的结构示意图。下面，参考图8来描述根据本公开实施例的电子设备。该电子设备可以是第一设备和第二设备中的任一个或两者、或与它们独立的单机设备，该单机设备可以与第一设备和第二设备进行通信，以从它们接收所采集到的输入信号。

[0153] 如图8所示，电子设备包括一个或多个处理器和存储器。

[0154] 处理器可以是中央处理单元(CPU)或者具有数据处理能力和/或指令执行能力的其他形式的处理单元，并且可以控制电子设备中的其他组件以执行期望的功能。

[0155] 存储器可以包括一个或多个计算机程序产品，所述计算机程序产品可以包括各种形式的计算机可读存储介质，例如易失性存储器和/或非易失性存储器。所述易失性存储器例如可以包括随机存取存储器(RAM)和/或高速缓冲存储器(cache)等。所述非易失性存储器例如可以包括只读存储器(ROM)、硬盘、闪存等。在所述计算机可读存储介质上可以存储一个或多个计算机程序指令，处理器可以运行所述程序指令，以实现上文所述的本公开的各个实施例的神经网络模型的量化方法以及/或者其他期望的功能。

[0156] 在一个示例中，电子设备还可以包括：输入装置和输出装置，这些组件通过总线系统和/或其他形式的连接机构(未示出)互连。

[0157] 此外，该输入设备还可以包括例如键盘、鼠标等等。

[0158] 该输出装置可以向外部输出各种信息，包括确定出的距离信息、方向信息等。该输出设备可以包括例如显示器、扬声器、打印机、以及通信网络及其所连接的远程输出设备等等。

[0159] 当然，为了简化，图8中仅示出了该电子设备中与本公开有关的组件中的一些，省略了诸如总线、输入/输出接口等等的组件。除此之外，根据具体应用情况，电子设备还可以包括任何其他适当的组件。

[0160] 除了上述方法和设备以外，本公开的实施例还可以是计算机程序产品，其包括计算机程序指令，所述计算机程序指令在被处理器运行时使得所述处理器执行本说明书上述部分中描述的根据本公开各种实施例的神经网络模型的量化方法中的步骤。

[0161] 所述计算机程序产品可以以一种或多种程序设计语言的任意组合来编写用于执行本公开实施例操作的程序代码，所述程序设计语言包括面向对象的程序设计语言，诸如Java、C++等，还包括常规的过程式程序设计语言，诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。

[0162] 此外，本公开的实施例还可以是计算机可读存储介质，其上存储有计算机程序指令，所述计算机程序指令在被处理器运行时使得所述处理器执行本说明书上述部分中描述的根据本公开各种实施例的神经网络模型的量化方法中的步骤。

[0163] 所述计算机可读存储介质可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以包括但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。

[0164] 本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储于一计算机可读取存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0165] 以上结合具体实施例描述了本公开的基本原理，但是，需要指出的是，在本公开中提及的优点、优势、效果等仅是示例而非限制，不能认为这些优点、优势、效果等是本公开的各个实施例必须具备的。另外，上述公开的具体细节仅是为了示例的作用和便于理解的作用，而非限制，上述细节并不限制本公开为必须采用上述具体的细节来实现。

[0166] 本说明书中各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其它实施例的不同之处，各个实施例之间相同或相似的部分相互参见即可。对于系统实施例而言，由于其与方法实施例基本对应，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0167] 本公开中涉及的器件、装置、设备、系统的方框图仅作为例示性的例子并且不意图要求或暗示必须按照方框图示出的方式进行连接、布置、配置。如本领域技术人员将认识到的，可以按任意方式连接、布置、配置这些器件、装置、设备、系统。诸如“包括”、“包含”、“具有”等等的词语是开放性词汇，指“包括但不限于”，且可与其互换使用。这里所使用的词汇“或”和“和”指词汇“和/或”，且可与其互换使用，除非上下文明确指示不是如此。这里所使用的词汇“诸如”指词组“诸如但不限于”，且可与其互换使用。

[0168] 可能以许多方式来实现本公开的方法和装置。例如，可通过软件、硬件、固件或者软件、硬件、固件的任何组合来实现本公开的方法和装置。用于所述方法的步骤的上述顺序仅是为了进行说明，本公开的方法的步骤不限于以上具体描述的顺序，除非以其它方式特

别说明。此外，在一些实施例中，还可将本公开实施为记录在记录介质中的程序，这些程序包括用于实现根据本公开的方法的机器可读指令。因而，本公开还覆盖存储用于执行根据本公开的方法的程序的记录介质。

[0169] 还需要指出的是，在本公开的装置、设备和方法中，各部件或各步骤是可以分解和/或重新组合的。这些分解和/或重新组合应视为本公开的等效方案。

[0170] 提供所公开的方面的以上描述以使本领域的任何技术人员能够做出或者使用本公开。对这些方面的各种修改对于本领域技术人员而言是非常显而易见的，并且在此定义的一般原理可以应用于其他方面而不脱离本公开的范围。因此，本公开不意图被限制到在此示出的方面，而是按照与在此公开的原理和新颖的特征一致的最宽范围。

[0171] 为了例示和描述的目的已经给出了以上描述。此外，此描述不意图将本公开的实施例限制到在此公开的形式。尽管以上已经讨论了多个示例方面和实施例，但是本领域技术人员将认识到其某些变型、修改、改变、添加和子组合。

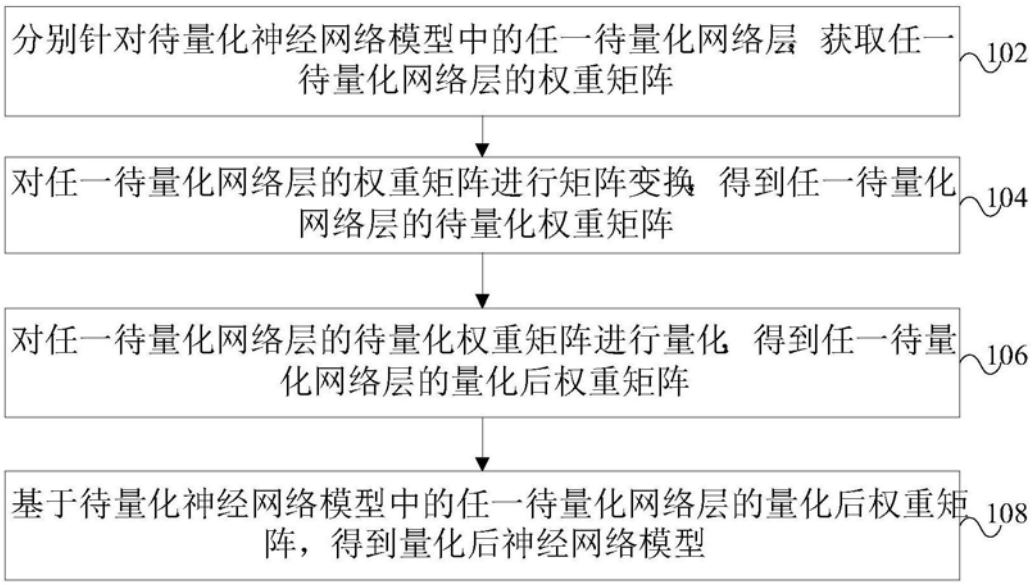
分别针对待量化神经网络模型中的任一待量化网络层，获取任一待量化网络层的权重矩阵
102
对任一待量化网络层的权重矩阵进行矩阵变换，得到任一待量化网络层的待量化权重矩阵
104
对任一待量化网络层的待量化权重矩阵进行量化，得到任一待量化网络层的量化后权重矩阵
106
基于待量化神经网络模型中的任一待量化网络层的量化后权重矩阵，得到量化后神经网络模型
108

图1

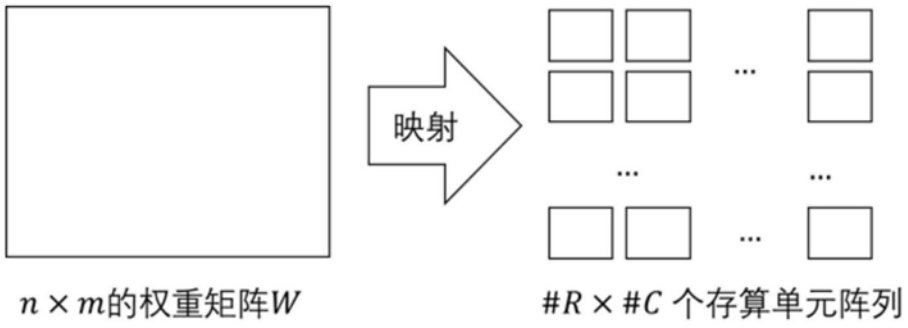
映射
$n \times m$的权重矩阵W
$\#R \times \#C$ 个存算单元阵列

图2

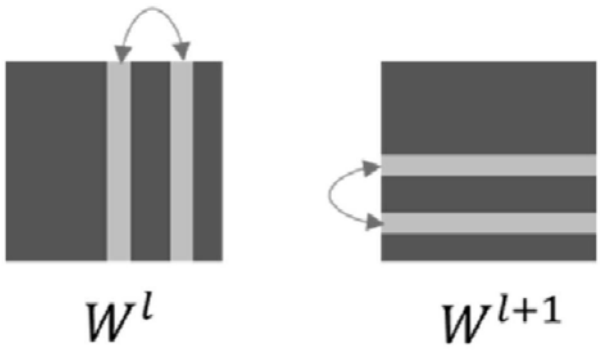
W^l
W^{l+1}

图3

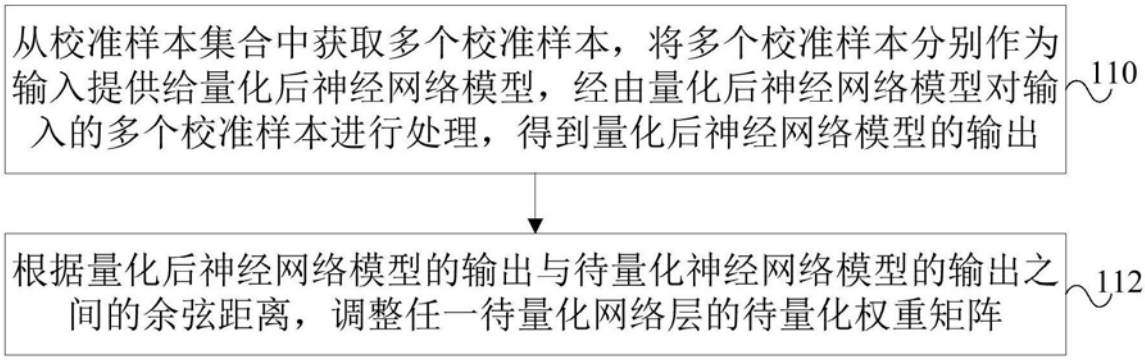
从校准样本集合中获取多个校准样本，将多个校准样本分别作为输入提供给量化后神经网络模型，经由量化后神经网络模型对输入的多个校准样本进行处理，得到量化后神经网络模型的输出
110
根据量化后神经网络模型的输出与待量化神经网络模型的输出之间的余弦距离，调整任一待量化网络层的待量化权重矩阵
112

图4

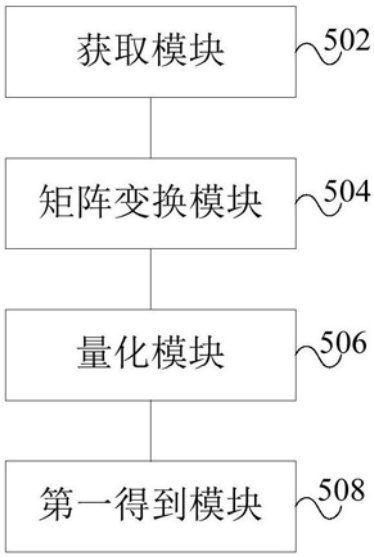
获取模块
502
矩阵变换模块
504
量化模块
506
第一得到模块
508

图5

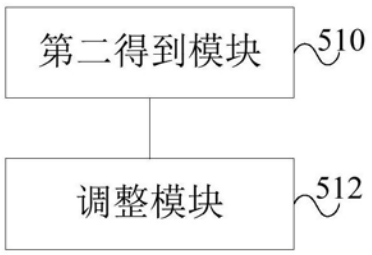
第二得到模块
510
调整模块
512

图6

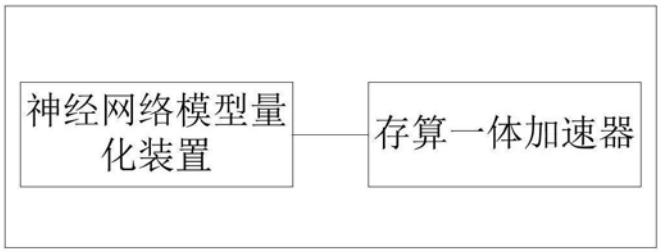

图7

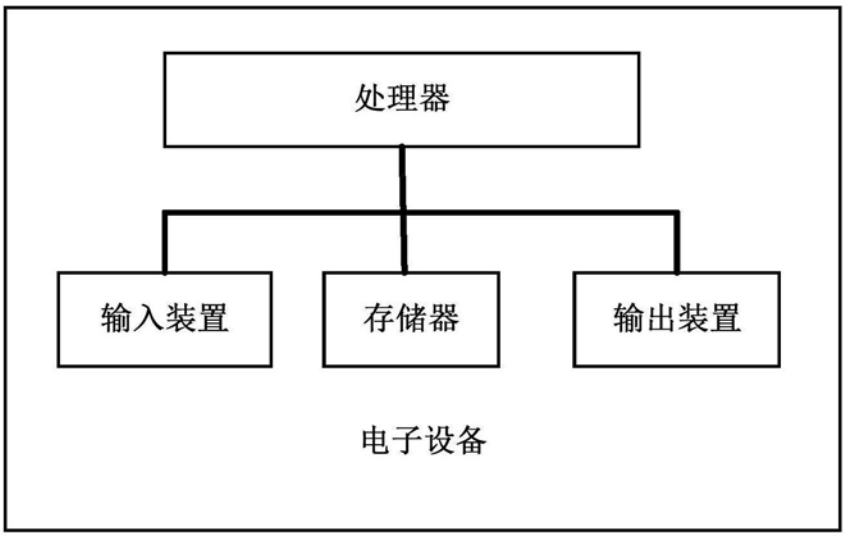

图8