(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 112532251 A
(43)申请公布日 2021.03.19

(21)申请号 201910891200.3

(22)申请日 2019.09.17

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 郑尚策 董永汉 于璠

(74)专利代理机构 深圳市深佳知识产权代理事务所(普通合伙) 44285
代理人 吴磊

(51)Int.Cl.
H03M 7/30(2006.01)
G06T 9/00(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

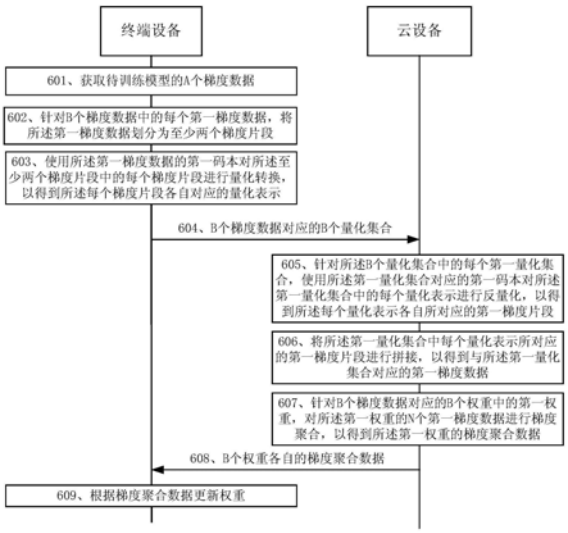
权利要求书5页 说明书24页 附图9页

(54)发明名称

一种数据处理的方法及设备

(57)摘要

本申请公开了一种数据处理的方法,可用于人工智能(artificial intelligence,AI)领域。该方法包括:终端设备将待训练模型的梯度数据划分为梯度片段,然后依据码本对梯度片段做量化转换,量化转换后得到的量化表示中元素的个数少于梯度片段中元素的个数,然后向云设备发送量化表示。本申请技术方案由于量化表示中元素的个数少于梯度片段中元素的个数,也就是相比于按元素量化技术提高了梯度数据的压缩比。因为量化表示的压缩比被提高了,终端设备向云设备传输量化表示时的通信开销也随之减少。



CN 112532251 A

1.一种数据处理的方法,其特征在于,包括:

获取待训练模型的A个梯度数据,A为正整数;

针对B个梯度数据中的每个第一梯度数据,将所述第一梯度数据划分为至少两个梯度片段,所述第一梯度数据包括d个元素,所述B个梯度数据包含于所述A个梯度数据中,B为正整数且B≤A,每个梯度片段中包括d′个元素,d′为大于2的正整数,且d能被d′整除;

使用所述第一梯度数据的第一码本对所述至少两个梯度片段中的每个梯度片段进行量化转换,以得到所述每个梯度片段各自对应的量化表示,所述量化表示中元素的个数小于d′,所述第一码本为d′行m列的第一矩阵或者d′列m行的第二矩阵,m为大于2的正整数,且m≥d’;

向云设备发送所述B个梯度数据对应的B个量化集合,其中,所述第一梯度数据对应的第一量化集合包括所述每个梯度片段各自对应的量化表示。

2.根据权利要求1所述的方法,其特征在于,所述方法还包括:

根据所述每个梯度数据各自的数据量,确定所述A个梯度数据中有B个梯度数据需要量化;

针对所述B个梯度数据中的每个第一梯度数据,根据所述第一梯度数据的元素数量d确定d′和m,d′为2的p次方,m为2的q次方,p和q都为正整数,且q≥p;

根据随机种子为所述第一梯度数据随机生成d′行m列的所述第一矩阵或者d′列m行的所述第二矩阵,且所述第一矩阵或所述第二矩阵中的每个高斯向量被缩放到模为1,以得到所述第一梯度数据的第一码本,所述第一矩阵中每一列为一个高斯向量,所述第二矩阵中每一行为一个高斯向量。

3.根据权利要求1所述的方法,其特征在于,所述方法还包括:

接收所述云设备发送的所述B个梯度数据各自对应的码本,所述B个梯度数据各自对应的码本中包括所述第一码本。

4.根据权利要求1-3任一项所述的方法,其特征在于,所述使用所述第一梯度数据的第一码本对所述至少两个梯度片段中的每个梯度片段进行量化转换,以得到所述每个梯度片段各自对应的量化表示,包括:

从所述第一码本中为第一梯度片段确定目标码,所述目标码为所述第一矩阵中的一列或者为所述第二矩阵中的一行,所述第一梯度片段为所述至少两个梯度片段中的任意一个;

确定所述目标码的伪模长和码索引;

将所述伪模长和所述码索引确定为所述第一梯度片段的量化表示。

5.根据权利要求4所述的方法,其特征在于,所述从所述第一码本中为第一梯度片段确定目标码,包括:

确定所述第一梯度片段的向量模长;

若所述向量模长等于0,则确定所述第一矩阵中的任意一列或者所述第二矩阵中的任意一行为目标码。

6.根据权利要求5所述的方法,其特征在于,所述方法还包括:

若所述向量模长不等于0,则根据所述第一码本和所述第一梯度片段确定所述第一梯度片段的第一系数向量,所述第一系数向量中包括m个元素;

对所述第一系数向量中的每个元素进行归一化，以得到第二系数向量，所述第二系数向量中包括m个归一化后的元素；

针对所述归一化后的m个元素采用轮盘赌的选择策略，确定其中第i个元素对应的所述第一矩阵中的第i列或所述第二矩阵中的第i行为目标码，所述第i个元素为采用轮盘赌的选择策略被选中的元素。

7.根据权利要求4所述的方法，其特征在于，所述从所述第一码本中为第一梯度片段确定目标码，包括：

根据所述第一码本和所述第一梯度片段确定所述第一梯度片段的投影向量，所述投影向量中包括m个元素；

确定第i个元素对应的所述第一码本中的第i列或所述第二矩阵中的第i行为目标码，所述第i元素的绝对值在所述m个元素的绝对值中最大。

8.根据权利要求1-7任一项所述的方法，其特征在于，还包括：

接收所述云设备发送的所述B个梯度数据对应的B个梯度聚合数据各自的量化集合，其中，每个梯度聚合数据的量化集合包括该梯度聚合数据的每个梯度片段对应的量化表示。

9.一种数据处理的方法，其特征在于，包括：

接收终端设备发送的B个梯度数据对应的B个量化集合，其中，每个量化集合包括对应梯度数据的每个梯度片段各自对应的量化表示；

针对所述B个量化集合中的每个第一量化集合，使用所述第一量化集合对应的第一码本对所述第一量化集合中的每个量化表示进行反量化，以得到所述每个量化表示各自所对应的第一梯度片段，所述第一梯度片段包括d′个元素，所述量化表示中元素的个数小于d′，所述第一码本为d′行m列的第一矩阵或者d′列m行的第二矩阵，d′和m为大于2的正整数，且m≥d’；

将所述第一量化集合中每个量化表示所对应的第一梯度片段进行拼接，以得到与所述第一量化集合对应的第一梯度数据，所述第一梯度数据包括d个元素，d为正整数，且d能被d′整除；

针对B个梯度数据对应的B个权重中的第一权重，对所述第一权重的N个第一梯度数据进行梯度聚合，以得到所述第一权重的梯度聚合数据，所述第一权重的多个第一梯度数据对应于N个终端设备，所述N为大于1的整数；

向所述终端设备发送所述B个权重各自的梯度聚合数据。

10.根据权利要求9所述的方法，其特征在于，所述方法还包括：

针对B个梯度数据中的每个第一梯度数据，根据随机种子为所述第一梯度数据随机生成d′行m列的所述第一矩阵或者d′列m行的所述第二矩阵，且所述第一矩阵或所述第二矩阵中的每个高斯向量被缩放到模为1，以得到所述第一梯度数据的第一码本，所述第一矩阵中每一列为一个高斯向量，所述第二矩阵中每一行为一个高斯向量。

11.根据权利要求10所述的方法，其特征在于，所述方法还包括：

向所述终端设备发送所述B个梯度数据各自对应的码本。

12.根据权利要求9-11任一项所述的方法，其特征在于，所述使用所述第一量化集合对应的第一码本对所述第一量化集合中的每个量化表示进行反量化，以得到所述每个量化表示各自所对应的第一梯度片段，包括：

针对所述第一量化集合中的每个第一量化表示，根据所述第一量化表示的码索引从所述第一码本中确定所述第一量化表示的目标码；

根据所述第一量化表示中的伪模长和所述目标码恢复出所述第一量化表示对应的第一梯度片段。

13.根据权利要求9-12任一项所述的方法，其特征在于，所述方法还包括：

对所述B个权重的梯度聚合数据分别进行量化转换，以得到所述B个梯度聚合数据各自的量化集合；

所述向所述终端设备发送所述B个权重各自的梯度聚合数据，包括：

向所述终端设备发送所述B个梯度聚合数据各自的量化集合，其中，每个梯度聚合数据的量化集合包括该梯度聚合数据的每个梯度片段对应的量化表示。

14.一种终端设备，包括收发器、处理器和存储器，所述处理器与所述存储器耦合，其特征在于，所述存储器，用于存储程序；

所述处理器用于：

获取待训练模型的A个梯度数据，A为正整数；

针对B个梯度数据中的每个第一梯度数据，将所述第一梯度数据划分为至少两个梯度片段，所述第一梯度数据包括d个元素，所述B个梯度数据包含于所述A个梯度数据中，B为正整数且B≤A，每个梯度片段中包括d′个元素，d′为大于2的正整数，且d能被d′整除；

使用所述第一梯度数据的第一码本对所述至少两个梯度片段中的每个梯度片段进行量化转换，以得到所述每个梯度片段各自对应的量化表示，所述量化表示中元素的个数小于d′，所述第一码本为d′行m列的第一矩阵或者d′列m行的第二矩阵，m为大于2的正整数，且m≥d’；

所述收发器用于向云设备发送所述B个梯度数据对应的B个量化集合，其中，所述第一梯度数据对应的第一量化集合包括所述每个梯度片段各自对应的量化表示。

15.根据权利要求14所述的终端设备，其特征在于，

所述处理器还用于：

根据所述每个梯度数据各自的数据量，确定所述A个梯度数据中有B个梯度数据需要量化；

针对所述B个梯度数据中的每个第一梯度数据，根据所述第一梯度数据的元素数量d确定d′和m，d′为2的p次方，m为2的q次方，p和q都为正整数，且q≥p；

根据随机种子为所述第一梯度数据随机生成d′行m列的所述第一矩阵或者d′列m行的所述第二矩阵，且所述第一矩阵或所述第二矩阵中的每个高斯向量被缩放到模为1，以得到所述第一梯度数据的第一码本，所述第一矩阵中每一列为一个高斯向量，所述第二矩阵中每一行为一个高斯向量。

16.根据权利要求14所述的终端设备，其特征在于，

所述收发器，用于接收所述云设备发送的所述B个梯度数据各自对应的码本，所述B个梯度数据各自对应的码本中包括所述第一码本。

17.根据权利要求14-16任一项所述的终端设备，其特征在于，

所述处理器用于：

从所述第一码本中为第一梯度片段确定目标码，所述目标码为所述第一矩阵中的一列

或者为所述第二矩阵中的一行，所述第一梯度片段为所述至少两个梯度片段中的任意一个；

确定所述目标码的伪模长和码索引；

将所述伪模长和所述码索引确定为所述第一梯度片段的量化表示。

18.根据权利要求17所述的终端设备，其特征在于，

所述处理器用于：

确定所述第一梯度片段的向量模长；

若所述向量模长等于0，则确定所述第一矩阵中的任意一列或者所述第二矩阵中的任意一行为目标码。

19.根据权利要求18所述的终端设备，其特征在于，

所述处理器还用于：

若所述向量模长不等于0，则根据所述第一码本和所述第一梯度片段确定所述第一梯度片段的第一系数向量，所述第一系数向量中包括m个元素；

对所述第一系数向量中的每个元素进行归一化，以得到第二系数向量，所述第二系数向量中包括m个归一化后的元素；

针对所述归一化后的m个元素采用轮盘赌的选择策略，确定其中第i个元素对应的所述第一矩阵中的第i列或所述第二矩阵中的第i行为目标码，所述第i个元素为采用轮盘赌的选择策略被选中的元素。

20.根据权利要求17所述的终端设备，其特征在于，

所述处理器用于：

根据所述第一码本和所述第一梯度片段确定所述第一梯度片段的投影向量，所述投影向量中包括m个元素；

确定第i个元素对应的所述第一码本中的第i列或所述第二矩阵中的第i行为目标码，所述第i元素的绝对值在所述m个元素的绝对值中最大。

21.根据权利要求14-21任一项所述的终端设备，其特征在于，

所述收发器，用于接收所述云设备发送的所述B个梯度数据对应的B个梯度聚合数据各自的量化集合，其中，每个梯度聚合数据的量化集合包括该梯度聚合数据的每个梯度片段对应的量化表示。

22.一种云设备，包括通信端口、处理器和存储器，所述处理器与所述存储器耦合，其特征在于，所述存储器，用于存储程序；

所述通信端口，用于接收终端设备发送的B个梯度数据对应的B个量化集合，其中，每个量化集合包括对应梯度数据的每个梯度片段各自对应的量化表示；

所述处理器用于：

针对所述B个量化集合中的每个第一量化集合，使用所述第一量化集合对应的第一码本对所述第一量化集合中的每个量化表示进行反量化，以得到所述每个量化表示各自所对应的第一梯度片段，所述第一梯度片段包括d′个元素，所述量化表示中元素的个数小于d′，所述第一码本为d′行m列的第一矩阵或者d′列m行的第二矩阵，d′和m为大于2的正整数，且m≥d’；

将所述第一量化集合中每个量化表示所对应的第一梯度片段进行拼接，以得到与所述

第一量化集合对应的第一梯度数据，所述第一梯度数据包括d个元素，d为正整数，且d能被d′整除；

针对B个梯度数据对应的B个权重中的第一权重，对所述第一权重的N个第一梯度数据进行梯度聚合，以得到所述第一权重的梯度聚合数据，所述第一权重的多个第一梯度数据对应于N个终端设备，所述N为大于1的整数；

所述通信端口，用于向所述终端设备发送所述B个权重各自的梯度聚合数据。

23.根据权利要求22所述的云设备，其特征在于，

所述处理器还用于：针对B个梯度数据中的每个第一梯度数据，根据随机种子为所述第一梯度数据随机生成d′行m列的所述第一矩阵或者d′列m行的所述第二矩阵，且所述第一矩阵或所述第二矩阵中的每个高斯向量被缩放到模为1，以得到所述第一梯度数据的第一码本，所述第一矩阵中每一列为一个高斯向量，所述第二矩阵中每一行为一个高斯向量。

24.根据权利要求23所述的云设备，其特征在于，

所述通信端口，还用于向所述终端设备发送所述B个梯度数据各自对应的码本。

25.根据权利要求22-24任一项所述的云设备，其特征在于，

所述处理器用于：

针对所述第一量化集合中的每个第一量化表示，根据所述第一量化表示的码索引从所述第一码本中确定所述第一量化表示的目标码；

根据所述第一量化表示中的伪模长和所述目标码恢复出所述第一量化表示对应的第一梯度片段。

26.根据权利要求22-25任一项所述的云设备，其特征在于，

所述处理器，还用于对所述B个权重的梯度聚合数据分别进行量化转换，以得到所述B个梯度聚合数据各自的量化集合；

所述通信端口，用于向所述终端设备发送所述B个梯度聚合数据各自的量化集合，其中，每个梯度聚合数据的量化集合包括该梯度聚合数据的每个梯度片段对应的量化表示。

27.一种计算机可读存储介质，包括程序，当其在计算机上运行时，使得计算机执行如权利要求1至8中任一项所述的方法，或者，使得计算机执行如权利要求9至13中任一项所述的方法。

一种数据处理的方法及设备

技术领域

[0001] 本申请涉及计算机技术领域,具体涉及一种数据处理的方法及设备。

背景技术

[0002] 近年来随着硬件算力的提升和大数据的出现,深度学习模型正以前所未有的速度和热度被发展和应用。深度学习模型在被应用前需要经过模型训练。联邦学习是一种模型训练方案,云设备可以联合多个终端设备共同训练模型,同时解决隐私问题。联邦学习有多轮训练过程,每轮按照一定的规则筛选多个终端设备共同参与训练,一定程度上解决了端侧样本数据量较少的问题。联邦学习过程中用户隐私数据在端侧本地使用,无需上传到云侧,可以解决隐私泄露问题。每个端侧训练得到的梯度要上传到云侧,云侧对多个端侧上传的梯度数据进行聚合,然后再将聚合后的梯度数据下发给端侧。

[0003] 大量的梯度数据在端侧和云侧之间传输,导致联邦学习面临着巨大的通信瓶颈问题,亟待解决。

发明内容

[0004] 本申请实施例提供一种数据处理的方法,可以提高梯度数据的压缩比,有效减少了梯度数据传输的通信开销。本申请实施例还提供了相应的设备。

[0005] 本申请第一方面提供一种数据处理的方法,该方法可以应用于端侧设备,例如终端设备。该方法可以包括:获取待训练模型的A个梯度数据,A为正整数。针对B个梯度数据中的每个梯度数据(或称为第一梯度数据)执行如下操作:将所述梯度数据划分为至少两个梯度片段,所述梯度数据包括d个元素,所述B个梯度数据包含于所述A个梯度数据中,B为正整数且B≤A,每个梯度片段中包括d′个元素,d′为大于2的正整数,且d能被d′整除;使用所述梯度数据的码本(code book)(或称第一码本)对所述至少两个梯度片段(gradient segment)中的每个梯度片段进行量化转换,以得到所述每个梯度片段各自对应的量化表示,所述量化表示中元素的个数小于d′,所述码本为d′行m列的第一矩阵或者d′列m行的第二矩阵,m为大于2的正整数,且m≥d′。向云设备发送所述B个梯度数据对应的B个量化集合,其中,所述第一梯度数据对应的第一量化集合包括所述每个梯度片段各自对应的量化表示。

[0006] 需要说明的是,为方便引用,本申请中可以用“第一梯度数据”代指B个梯度数据中任意的一个梯度数据。待训练模型以及A个梯度数据的来源本申请不做限定。

[0007] 这里“发送所述B个梯度数据对应的B个量化集合”可以是B个梯度数据全部完成量化转换后统一发送,也可以分多次发送或不等待全部完成,部分(包括一个)完成后就发送。

[0008] 该第一方面中,待训练模型中可以有A个与权重相关的算子,每个算子都会有一个对应的梯度数据,每个梯度数据中都会包括至少一个元素,每个元素可以表示一个梯度。B个梯度数据可以是A个梯度数据中的部分,也可以是全部,例如:B=A时,表示A个梯度数据中的每一个都要做梯度片段的划分和量化转换,B<A时,表示其中部分需要做梯度片段的

划分和量化转换，剩余的(A-B)个梯度数据不需要做梯度片段的划分和量化转换。梯度片段在划分时，可以根据d的数量，确定d′，d能被d′整除，例如：d＝1200，d′＝16，d/d′＝75，则表示该第一梯度数据可以划分为75个梯度片段。每个第一梯度数据都会对应有一个码本，不同梯度数据的d′可以不相同，不同码本中的d′和m也可以不相同。m≥d′可以确保在第一矩阵或第二矩阵都处于满秩状态，这样可以为各梯度片段找到匹配精度更高的列或行。在模型训练过程中，通常会规定使用第一矩阵或者第二矩阵。量化表示中元素的个数小于d′，在量化压缩的基础上还进一步缩减了量化数据的数量，从而提高了压缩比。另外，因为量化表示的压缩比被提高了，终端设备向云设备传输量化表示时的通信开销也随之减少。

[0009] 在第一方面的一种可能的实现方式中，当A＞B时，该方法还可以包括：

[0010] 向所述云设备发送所述A个梯度数据中除所述B个梯度数据之外的(A-B)个梯度数据。

[0011] 该种可能的实现方式中，(A-B)个梯度数据可能本身的元素个数就很少，量化转换的过程还会消耗计算资源，从综合收益的角度考虑，可以不经压缩处理直接发送(A-B)个梯度数据，也可以采用其他的压缩处理方式，对(A-B)个梯度数据进行压缩处理，然后再发送。

[0012] 在第一方面的一种可能的实现方式中，该方法还可以包括：

[0013] 根据所述每个梯度数据各自的数据量，确定所述A个梯度数据中有B个梯度数据需要量化；

[0014] 针对所述B个梯度数据中的每个第一梯度数据，根据所述第一梯度数据的元素数量d确定d′和m，d′为2的p次方，m为2的q次方，p和q都为正整数，且q≥p；

[0015] 根据随机种子为所述第一梯度数据随机生成d′行m列的所述第一矩阵或者d′列m行的所述第二矩阵，且所述第一矩阵或所述第二矩阵中的每个高斯向量被缩放到模为1，以得到所述第一梯度数据的第一码本，所述第一矩阵中每一列为一个高斯向量，所述第二矩阵中每一行为一个高斯向量。

[0016] 该种可能的实现方式中，可以从综合收益的角度确定一个需要进行量化转换的数据量阈值，例如：设定该数据量阈值为512个元素，若梯度数据中的元素个数大于或等于512，则表示需要进行量化转换，若梯度数据中的元素个数小于512，则表示不需要进行量化转换。终端设备只需要为需要进行量化转换的梯度数据生成码本。这样可以减少因生成码本而带来的计算开销，提高计算开销和通信开销的综合收益。为了确保云设备可以正常恢复梯度数据，终端设备生成码本所使用的随机种子与云设备相同。随机种子可以是一个数值，向随机函数中输入随机种子就可以随机生成梯度数据的码本。

[0017] 在第一方面的一种可能的实现方式中，该方法还可以包括：

[0018] 接收所述云设备发送的所述B个梯度数据各自对应的码本，所述B个梯度数据各自对应的码本中包括所述第一码本。

[0019] 该种可能的实现方式中，也可以不需要终端设备自行生成码本，可以由云设备生成码本，然后下发给终端设备，这样可以减少终端设备的计算开销。

[0020] 在第一方面的一种可能的实现方式中，上述步骤：使用所述第一梯度数据的第一码本对所述至少两个梯度片段中的每个梯度片段进行量化转换，以得到所述每个梯度片段各自对应的量化表示，可以包括：

[0021] 从所述第一码本中为第一梯度片段确定目标码(target code)，所述目标码为所

述第一矩阵中的一列或者为所述第二矩阵中的一行,所述第一梯度片段为所述至少两个梯度片段中的任意一个;

[0022] 确定所述目标码的伪模长(pseudo norm)和码索引(code index);

[0023] 将所述伪模长和所述码索引确定为所述第一梯度片段的量化表示。

[0024] 该种可能的实现方式中,伪模长与模长是相对的,伪模长表示不是按照数学上的向量模长的计算方法得到的,而是通过码本和梯度片段计算得到的。码索引表示第一矩阵的一列的索引,例如:index(5)表示第五列。或者,表示第二矩阵的一行的索引,例如:index(6)表示第六行。若为模长用u表示,码索引index(c)表示,则量化表示可以为(u,index(c))。由该种可能的实现方式可知,一个梯度片段转换成了只包括伪模长和码索引两个元素,极大的缩减了数据量,提高了梯度数据的压缩比,也减少了传输梯度数据的通信开销。

[0025] 在第一方面的一种可能的实现方式中,上述步骤:从所述第一码本中为第一梯度片段确定目标码,可以包括:

[0026] 确定所述第一梯度片段的向量模长(vector norm);

[0027] 若所述向量模长等于0,则确定所述第一矩阵中的任意一列或者所述第二矩阵中的任意一行为目标码。

[0028] 该种可能的实现方式中,向量模长为第一梯度片段上的各元素的平方和然后再求平方根。若向量模长等于0则表示各元素都为0,所以只需要随机选择一列或一行凑成符合量化表示的格式即可,例如:可以选择第一列或第一行。

[0029] 在第一方面的一种可能的实现方式中,该方法还可以包括:

[0030] 若所述向量模长不等于0,则根据所述第一码本和所述第一梯度片段确定所述第一梯度片段的第一系数向量,所述第一系数向量中包括m个元素;

[0031] 对所述第一系数向量中的每个元素进行归一化,以得到第二系数向量,所述第二系数向量中包括m个归一化后的元素;

[0032] 针对所述归一化后的m个元素采用轮盘赌的选择策略,确定其中第i个元素对应的所述第一矩阵中的第i列或所述第二矩阵中的第i行为目标码,所述第i个元素为采用轮盘赌的选择策略被选中的元素。

[0033] 该种可能的实现方式中,若第一码本用C表示,第一梯度片段用g表示,第一系数向量用p表示,$p=C^{T}(CC^{T})^{-1}g$,其中,CT表示C的转置矩阵,第二系数向量用$\tilde{p}$表示,$\tilde{p}$中包括m个归一化后的元素,元素归一化可以理解为:$\widetilde{p_b}=|p_b|/\|p\|_1$,其中$\widetilde{p_b}$表示m个归一化后的元素中的第b个,$|p_b|$表示第一系数向量p中的第b个元素的绝对值,$||p||_1$表示第一系数向量p的m元素中各元素的绝对值之和。轮盘赌的选择策略在该方案中可以认为每个归一化的元素都有各自的概率,各元素的概率有大有小,所有概率之和等于1,概率大的被选中的概率也大,概率小的被选中的概率也小,但最终的选择结果还是以轮盘赌转动后选择到的元素为准。若选中的是$\widetilde{p_b}$,则就选第b个元素所对应列或行的码作为目标码。当然,若选中的是第i个归一化后的元素,则选第i个元素所对应列或行的码作为目标码。

[0034] 在第一方面的一种可能的实现方式中,上述步骤:从所述第一码本中为第一梯度片段确定目标码,包括:

[0035] 根据所述第一码本和所述第一梯度片段确定所述第一梯度片段的投影向量(projection vector),所述投影向量中包括m个元素;

[0036] 确定第i个元素对应的所述第一码本中的第i列或所述第二矩阵中的第i行为目标码，所述第i元素的绝对值在所述m个元素的绝对值中最大。

[0037] 该种可能的实现方式中，若第一码本用C表示，第一梯度片段用g表示，投影向量用p表示，则$p=C^{T}g$，其中，C^{T}表示C的转置矩阵。投影向量p中包括m个元素；$|p_i|$表示m个元素中第i个元素的绝对值。若i＝5，则表示选择第一码本中的第5列或所述第二矩阵中的第5行作为目标码。

[0038] 在第一方面的一种可能的实现方式中，该方法还可以包括：

[0039] 接收所述云设备发送的所述B个梯度数据对应的B个梯度聚合数据各自的量化集合，其中，每个梯度聚合数据的量化集合包括该梯度聚合数据的每个梯度片段对应的量化表示。

[0040] 该种可能的实现方式中，云设备对执行梯度聚合后，针对聚合后得到的梯度聚合数据，也可以采用上述终端设备的量化转换方式进行量化转换，这样可以进一步减少云设备与终端设备的通信开销。

[0041] 本申请第二方面提供一种数据处理的方法，该方法可以应用于云侧设备，例如云设备，云设备可以是服务器，也可以是其他设备或一块虚拟资源。该方法可以包括：接收终端设备发送的B个梯度数据对应的B个量化集合，其中，每个量化集合包括对应梯度数据的每个梯度片段各自对应的量化表示；针对所述B个量化集合中的每个第一量化集合，使用所述第一量化集合对应的第一码本对所述第一量化集合中的每个量化表示进行反量化，以得到所述每个量化表示各自所对应的第一梯度片段，所述第一梯度片段包括d′个元素，所述量化表示中元素的个数小于d′，所述第一码本为d′行m列的第一矩阵或者d′列m行的第二矩阵，d和m为大于2的正整数，且m≥d′；将所述第一量化集合中每个量化表示所对应的第一梯度片段进行拼接，以得到与所述第一量化集合对应的第一梯度数据，所述第一梯度数据包括个元素，d为正整数，且d能被d′整除；针对B个梯度数据对应的B个权重中的第一权重，对所述第一权重的N个第一梯度数据进行梯度聚合，以得到所述第一权重的梯度聚合数据，所述第一权重的多个第一梯度数据对应于N个终端设备，所述N为大于1的整数；向所述终端设备发送所述B个权重各自的梯度聚合数据。

[0042] 需要说明的是，为方便引用，本申请中可以用“第一量化集合”代指B个量化集合中任意的一个量化集合。“第一梯度数据”代指B个梯度数据中任意的一个梯度数据。

[0043] 这里“接收B个梯度数据对应的B个量化集合”可以是统一接收，也可以分多次接收的。

[0044] 这里“发送所述B个权重各自的梯度聚合数据”可以是梯度聚合完统一发送，也可以分多次发送或不等待全部完成，部分(包括一个)完成后就发送。

[0045] 该第二方面中，该云设备接收到量化表示后，会使用终端设备在量化转换时相同的码本做反量化，也就是量化恢复，进而进行梯度聚合。梯度聚合的过程指的是云设备对从多个终端设备接收到的同一个权重的各个梯度进行聚合，例如：针对权重A接收到了500个梯度，则可以将这500个梯度相加再求平均，该平均值作为该权重A的聚合后的梯度。云设备将该聚合后的梯度聚合数据发送给终端设备后，终端设备可以更新各权重，从而使待训练模型向收敛更进一步。权重更新的过程可以是用本轮当前的权重减去本轮计算得到的梯度聚合数据，得到更新后的权重，该更新后的权重用于下一轮训练。该第二方面中，因为云设

备从终端设备接收的是终端设备量化转换后的量化表示，因为该量化表示中元素的个数小于每个梯度片段中的元素个数d′，所以只需要用很小的通信开销就能实现梯度数据的接收。

[0046] 在第二方面的一种可能的实现方式中，该方法还可以包括：

[0047] 针对B个梯度数据中的每个第一梯度数据，根据随机种子为所述第一梯度数据随机生成d′行m列的所述第一矩阵或者d′列m行的所述第二矩阵，且所述第一矩阵或所述第二矩阵中的每个高斯向量被缩放到模为1，以得到所述第一梯度数据的第一码本，所述第一矩阵中每一列为一个高斯向量，所述第二矩阵中每一行为一个高斯向量。

[0048] 该种可能的实现方式中，随机种子可以是一个数值，向随机函数中输入随机种子就可以随机生成梯度数据的码本。

[0049] 在第二方面的一种可能的实现方式中，该方法还可以包括：

[0050] 向所述终端设备发送所述B个梯度数据各自对应的码本。

[0051] 该种可能的实现方式中，云设备向终端设备发送生成的码本，可以减少终端设备的计算开销。

[0052] 在第二方面的一种可能的实现方式中，上述步骤：使用所述第一量化集合对应的第一码本对所述第一量化集合中的每个量化表示进行反量化，以得到所述每个量化表示各自所对应的第一梯度片段，可以包括：

[0053] 针对所述第一量化集合中的每个第一量化表示，根据所述第一量化表示的码索引从所述第一码本中确定所述第一量化表示的目标码；

[0054] 根据所述第一量化表示中的伪模长和所述目标码恢复出所述第一量化表示对应的第一梯度片段。

[0055] 该种可能的实现方式中，云设备根据码索引可以从第一码本中找到目标码，然后使用伪模长与目标码中的元素相乘就可以恢复出对应的梯度片段。

[0056] 在第二方面的一种可能的实现方式中，该方法还可以包括：

[0057] 对所述B个权重的梯度聚合数据分别进行量化转换，以得到所述B个梯度聚合数据各自的量化集合；

[0058] 所述向所述终端设备发送所述B个权重各自的梯度聚合数据，包括：

[0059] 向所述终端设备发送所述B个梯度聚合数据各自的量化集合，其中，每个梯度聚合数据的量化集合包括该梯度聚合数据的每个梯度片段对应的量化表示。

[0060] 该种可能的实现方式中，云设备可以对要发送给终端设备的梯度聚合数据也进行类似于终端设备对梯度数据所进行的量化转换，这样可以进一步减少云设备与终端设备的通信开销。

[0061] 本申请第三方面提供一种终端设备，该终端设备具有实现上述第一方面或第一方面任意一种可能实现方式的方法的功能。该功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。该硬件或软件包括一个或多个与上述功能相对应的模块，例如：接收单元、处理单元和发送单元。

[0062] 本申请第四方面提供一种云设备，该云设备具有实现上述第二方面或第二方面任意一种可能实现方式的方法的功能。该功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。该硬件或软件包括一个或多个与上述功能相对应的模块，例如：接收单元、处

理单元和发送单元。本申请中的云设备可以是任意一种部署在网络侧或者云侧的计算机设备。

[0063] 本申请第五方面提供一种终端设备，该终端设备包括至少一个处理器、存储器、收发器以及存储在存储器中并可在处理器上运行的计算机执行指令，当所述计算机执行指令被所述处理器执行时，所述处理器执行如上述第一方面或第一方面任意一种可能的实现方式所述的方法。

[0064] 本申请第六方面提供一种云设备，该云设备包括至少一个处理器、存储器、通信端口以及存储在存储器中并可在处理器上运行的计算机执行指令，当所述计算机执行指令被所述处理器执行时，所述处理器执行如上述第二方面或第二方面任意一种可能的实现方式所述的方法。

[0065] 本申请第七方面提供一种存储一个或多个计算机执行指令的计算机可读存储介质，当所述计算机执行指令被处理器执行时，所述处理器执行如上述第一方面或第一方面任意一种可能的实现方式所述的方法。

[0066] 本申请第八方面提供一种存储一个或多个计算机执行指令的计算机可读存储介质，当所述计算机执行指令被处理器执行时，所述处理器执行如上述第二方面或第二方面任意一种可能的实现方式所述的方法。

[0067] 本申请第九方面提供一种存储一个或多个计算机执行指令的计算机程序产品(或称计算机程序)，当所述计算机执行指令被所述处理器执行时，所述处理器执行上述第一方面或第一方面任意一种可能实现方式的方法。

[0068] 本申请第十方面提供一种存储一个或多个计算机执行指令的计算机程序产品，当所述计算机执行指令被所述处理器执行时，所述处理器执行上述第二方面或第二方面任意一种可能实现方式的方法。

[0069] 本申请第十一方面提供了一种芯片系统，该芯片系统包括处理器，用于支持终端设备实现上述第一方面或第一方面任意一种可能的实现方式中所涉及的功能。在一种可能的设计中，芯片系统还可以包括存储器，存储器，用于保存终端设备必要的程序指令和数据。该芯片系统，可以由芯片构成，也可以包含芯片和其他分立器件。

[0070] 本申请第十二方面提供了一种芯片系统，该芯片系统包括处理器，用于支持云设备实现上述第二方面或第二方面任意一种可能的实现方式中所涉及的功能。在一种可能的设计中，芯片系统还可以包括存储器，存储器，用于保存云设备必要的程序指令和数据。该芯片系统，可以由芯片构成，也可以包含芯片和其他分立器件。

[0071] 其中，第三、第五、第七、第九和第十一方面或者其中任一种可能实现方式所带来的技术效果可参见第一方面或第一方面不同可能实现方式所带来的技术效果，此处不再赘述。

[0072] 其中，第四、第六、第八、第十和第十二方面或者其中任一种可能实现方式所带来的技术效果可参见第二方面或第二方面不同可能实现方式所带来的技术效果，此处不再赘述。

[0073] 本申请实施例采用将要传输的梯度数据进行梯度片段划分，然后再依据码本进行量化转换，得到量化表示，量化表示中元素的个数小于每个梯度片段中元素的个数。这样在量化压缩的基础上还进一步缩减了量化数据的数量，从而提高了压缩比。另外，因为量化表

示的压缩比被提高了，终端设备向云设备传输量化表示时的通信开销也随之减少。

附图说明

[0074] 图1为本申请实施例提供的一种人工智能主体框架示意图；

[0075] 图2为本申请实施例提供的一种应用环境示意图；

[0076] 图3为本申请实施例提供的一种卷积神经网络结构示意图；

[0077] 图4为本申请实施例提供的一种卷积神经网络结构示意图；

[0078] 图5为本申请实施例提供的一种神经网络处理器的结构示意图；

[0079] 图6为本申请实施例提供的端云结合的模型训练系统的一示意图；

[0080] 图7为本申请实施例提供的数据处理的方法的一实施例示意图；

[0081] 图8为本申请实施例提供的数据处理的方法的另一实施例示意图；

[0082] 图9为本申请实施例提供的终端设备的一实施例示意图；

[0083] 图10为本申请实施例提供的云设备的一实施例示意图；

[0084] 图11为本申请实施例提供的终端设备的一实施例示意图；

[0085] 图12为本申请实施例提供的云设备的一实施例示意图。

具体实施方式

[0086] 下面结合附图，对本申请的实施例进行描述，显然，所描述的实施例仅仅是本申请一部分的实施例，而不是全部的实施例。本领域普通技术人员可知，随着技术的发展和新场景的出现，本申请实施例提供的技术方案对于类似的技术问题，同样适用。

[0087] 本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的实施例能够以除了在这里图示或描述的内容以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0088] 本申请实施例提供一种数据处理的方法，可以提高端侧和云侧在梯度传输过程中梯度数据的压缩比，有效减少了梯度数据传输的通信开销。本申请实施例还提供了相应的设备。以下分别进行详细说明。

[0089] 图1示出一种人工智能主体框架示意图，该主体框架描述了人工智能系统总体工作流程，适用于通用的人工智能领域需求。

[0090] 下面从“智能信息链”(水平轴)和“IT价值链”(垂直轴)两个维度对上述人工智能主题框架进行阐述。

[0091] “智能信息链”反映从数据的获取到处理的一列过程。举例来说，可以是智能信息感知、智能信息表示与形成、智能推理、智能决策、智能执行与输出的一般过程。在这个过程中，数据经历了“数据—信息—知识—智慧”的凝练过程。

[0092] “IT价值链”从人智能的底层基础设施、信息(提供和处理技术实现)到系统的产业生态过程，反映人工智能为信息技术产业带来的价值。

[0093] (1)基础设施：

[0094] 基础设施为人工智能系统提供计算能力支持，实现与外部世界的沟通，并通过基础平台实现支撑。通过传感器与外部沟通；计算能力由智能芯片(中央处理器(central processing unit,CPU)、神经网络处理器(network processing unit,NPU)、图形处理器(graphic processing unit GPU)、专用集成电路(application specific integrated circuit,ASIC)、现场可编程门阵列(field-programmable gate array,FPGA)等硬件加速芯片)提供；基础平台包括分布式计算框架及网络等相关的平台保障和支持，可以包括云存储和计算、互联互通网络等。举例来说，传感器和外部沟通获取数据，这些数据提供给基础平台提供的分布式计算系统中的智能芯片进行计算。

[0095] (2)数据

[0096] 基础设施的上一层的数据用于表示人工智能领域的数据来源。数据涉及到图形、图像、语音、文本，还涉及到传统设备的物联网数据，包括已有系统的业务数据以及力、位移、液位、温度、湿度等感知数据。

[0097] (3)数据处理

[0098] 数据处理通常包括数据训练，机器学习，深度学习，搜索，推理，决策等方式。

[0099] 其中，机器学习和深度学习可以对数据进行符号化和形式化的智能信息建模、抽取、预处理、训练等。

[0100] 推理是指在计算机或智能系统中，模拟人类的智能推理方式，依据推理控制策略，利用形式化的信息进行机器思维和求解问题的过程，典型的功能是搜索与匹配。

[0101] 决策是指智能信息经过推理后进行决策的过程，通常提供分类、排序、预测等功能。

[0102] (4)通用能力

[0103] 对数据经过上面提到的数据处理后，进一步基于数据处理的结果可以形成一些通用的能力，比如可以是算法或者一个通用系统，例如，翻译，文本的分析，计算机视觉的处理，语音识别，图像的识别等等。

[0104] (5)智能产品及行业应用

[0105] 智能产品及行业应用指人工智能系统在各领域的产品和应用，是对人工智能整体解决方案的封装，将智能信息决策产品化、实现落地应用，其应用领域主要包括：智能制造、智能交通、智能家居、智能医疗、智能安防、自动驾驶，平安城市，智能终端等。

[0106] 参见附图2，本申请实施例提供了一种系统架构200。数据采集设备260用于采集训练数据并存入数据库230，训练设备220基于数据库230中维护的训练数据生成目标模型/规则201。下面将更详细地描述训练设备220如何基于训练数据得到目标模型/规则201，目标模型/规则201能够用于图像识别、视频分类、语音识别和语言翻译等应用场景。

[0107] 该目标模型/规则201可以是基于深度神经网络或者卷积神经网络convolutional neuron nrtwork,CNN)得到的，下面对深度神经网络或者卷积神经网络分别进行介绍。

[0108] 深度神经网络中的每一层的工作可以用数学表达式$\vec{y}=a(W\bullet\vec{x}+b)$来描述：从物理层面深度神经网络中的每一层的工作可以理解为通过五种对输入空间(输入向量的集合)的操作，完成输入空间到输出空间的变换(即矩阵的行空间到列空间)，这五种操作包括：1、

升维/降维;2、放大/缩小;3、旋转;4、平移;5、“弯曲”。其中1、2、3的操作由$W \cdot \vec{x}$完成,4的操作由+b完成,5的操作则由a()来实现。这里之所以用“空间”二字来表述是因为被分类的对象并不是单个事物,而是一类事物,空间是指这类事物所有个体的集合。其中,W是权重向量,该向量中的每一个值表示该层神经网络中的一个神经元的权重值。该向量W决定着上文所述的输入空间到输出空间的空间变换,即每一层的权重W控制着如何变换空间。训练深度神经网络的目的,也就是最终得到训练好的神经网络的所有层的权重矩阵(由很多层的向量W形成的权重矩阵)。因此,神经网络的训练过程本质上就是学习控制空间变换的方式,更具体的就是学习权重矩阵。

[0109] 因为希望深度神经网络的输出尽可能的接近真正想要预测的值,所以可以通过比较当前网络的预测值和真正想要的目标值,再根据两者之间的差异情况来更新每一层神经网络的权重向量(当然,在第一次更新之前通常会有初始化的过程,即为深度神经网络中的各层预先配置参数),比如,如果网络的预测值高了,就调整权重向量让它预测低一些,不断的调整,直到神经网络能够预测出真正想要的目标值。因此,就需要预先定义“如何比较预测值和目标值之间的差异”,这便是损失函数(loss function)或目标函数(objective function),它们是用于衡量预测值和目标值的差异的重要方程。其中,以损失函数举例,损失函数的输出值(loss)越高表示差异越大,那么深度神经网络的训练就变成了尽可能缩小这个loss的过程。

[0110] 训练设备220得到的目标模型/规则可以应用不同的系统或设备中。在附图2中,执行设备210配置有I/O接口212,与外部设备进行数据交互,“用户”可以通过客户设备240向I/O接口212输入数据。

[0111] 执行设备210可以调用数据存储系统250中的数据、代码等,也可以将数据、指令等存入数据存储系统250中。

[0112] 计算模块211使用目标模型/规则201对输入的数据进行处理,以文本类型的语言翻译为例,计算模块211可以对第一语言的文本中的语句进行解析,得到每条语句中的主语、谓语和宾语等词语。

[0113] 关联功能模块213可以对计算模块211中第一条语句中主语、谓语和宾语等词语翻译成第二语言,再结合第二语言的语法逻辑组织该条语句。

[0114] 关联功能模块214可以对计算模块211中第二条语句中主语、谓语和宾语等词语翻译成第二语言,再结合第二语言的语法逻辑组织该条语句。

[0115] 最后,I/O接口212将处理结果返回给客户设备240,提供给用户。

[0116] 更深层地,训练设备220可以针对不同的目标,基于不同的数据生成相应的目标模型/规则201,以给用户提供更佳的结果。

[0117] 在附图2中所示情况下,用户可以手动指定输入执行设备210中的数据,例如,在I/O接口212提供的界面中操作。另一种情况下,客户设备240可以自动地向I/O接口212输入数据并获得结果,如果客户设备240自动输入数据需要获得用户的授权,用户可以在客户设备240中设置相应权限。用户可以在客户设备240查看执行设备210输出的结果,具体的呈现形式可以是显示、声音、动作等具体方式。客户设备240也可以作为数据采集端将采集到训练数据存入数据库230。

[0118] 值得注意的,附图2仅是本申请实施例提供的一种系统架构的示意图,图中所示设

备、器件、模块等之间的位置关系不构成任何限制，例如，在附图2中，数据存储系统250相对执行设备210是外部存储器，在其它情况下，也可以将数据存储系统250置于执行设备210中。

[0119] 卷积神经网络是一种带有卷积结构的深度神经网络，是一种深度学习(deep learning)架构，深度学习架构是指通过机器学习的算法，在不同的抽象层级上进行多个层次的学习。作为一种深度学习架构，CNN是一种前馈(feed-forward)人工神经网络，该前馈人工神经网络中的各个神经元对输入其中的图像中的重叠区域作出响应。

[0120] 如图3所示，卷积神经网络(CNN)100可以包括输入层110，卷积层/池化层120，其中池化层为可选的，以及神经网络层130。

[0121] 卷积层/池化层120：

[0122] 卷积层：

[0123] 如图3所示卷积层/池化层120可以包括如示例121-126层，在一种实现中，121层为卷积层，122层为池化层，123层为卷积层，124层为池化层，125为卷积层，126为池化层；在另一种实现方式中，121、122为卷积层，123为池化层，124、125为卷积层，126为池化层。即卷积层的输出可以作为随后的池化层的输入，也可以作为另一个卷积层的输入以继续进行卷积操作。

[0124] 以卷积层121为例，卷积层121可以包括很多个卷积算子，卷积算子也称为核，其在图像处理中的作用相当于一个从输入图像矩阵中提取特定信息的过滤器，卷积算子本质上可以是一个权重矩阵，这个权重矩阵通常被预先定义，在对图像进行卷积操作的过程中，权重矩阵通常在输入图像上沿着水平方向一个像素接着一个像素(或两个像素接着两个像素……这取决于步长stride的取值)的进行处理，从而完成从图像中提取特定特征的工作。该权重矩阵的大小应该与图像的大小相关，需要注意的是，权重矩阵的纵深维度(depth dimension)和输入图像的纵深维度是相同的，在进行卷积运算的过程中，权重矩阵会延伸到输入图像的整个深度。因此，和一个单一的权重矩阵进行卷积会产生一个单一纵深维度的卷积化输出，但是大多数情况下不使用单一权重矩阵，而是应用维度相同的多个权重矩阵。每个权重矩阵的输出被堆叠起来形成卷积图像的纵深维度。不同的权重矩阵可以用来提取图像中不同的特征，例如一个权重矩阵用来提取图像边缘信息，另一个权重矩阵用来提取图像的特定颜色，又一个权重矩阵用来对图像中不需要的噪点进行模糊化……该多个权重矩阵维度相同，经过该多个维度相同的权重矩阵提取后的特征图维度也相同，再将提取到的多个维度相同的特征图合并形成卷积运算的输出。

[0125] 这些权重矩阵中的权重值在实际应用中需要经过大量的训练得到，通过训练得到的权重值形成的各个权重矩阵可以从输入图像中提取信息，从而帮助卷积神经网络100进行正确的预测。

[0126] 当卷积神经网络100有多个卷积层的时候，初始的卷积层(例如121)往往提取较多的一般特征，该一般特征也可以称之为低级别的特征；随着卷积神经网络100深度的加深，越往后的卷积层(例如126)提取到的特征越来越复杂，比如高级别的语义之类的特征，语义越高的特征越适用于待解决的问题。

[0127] 池化层：

[0128] 由于常常需要减少训练参数的数量，因此卷积层之后常常需要周期性的引入池化

层,即如图3中120所示例的121-126各层,可以是一层卷积层后面跟一层池化层,也可以是多层卷积层后面接一层或多层池化层。在图像处理过程中,池化层的唯一目的就是减少图像的空间大小。池化层可以包括平均池化算子和/或最大池化算子,以用于对输入图像进行采样得到较小尺寸的图像。平均池化算子可以在特定范围内对图像中的像素值进行计算产生平均值。最大池化算子可以在特定范围内取该范围内值最大的像素作为最大池化的结果。另外,就像卷积层中用权重矩阵的大小应该与图像大小相关一样,池化层中的运算符也应该与图像的大小相关。通过池化层处理后输出的图像尺寸可以小于输入池化层的图像的尺寸,池化层输出的图像中每个像素点表示输入池化层的图像的对应子区域的平均值或最大值。

[0129] 神经网络层130:

[0130] 在经过卷积层/池化层120的处理后,卷积神经网络100还不足以输出所需要的输出信息。因为如前所述,卷积层/池化层120只会提取特征,并减少输入图像带来的参数。然而为了生成最终的输出信息(所需要的类信息或别的相关信息),卷积神经网络100需要利用神经网络层130来生成一个或者一组所需要的类的数量的输出。因此,在神经网络层130中可以包括多层隐含层(如图3所示的131、132至13n)以及输出层140,该多层隐含层中所包含的参数可以根据具体的任务类型的相关训练数据进行预先训练得到,例如该任务类型可以包括图像识别,图像分类,图像超分辨率重建等等……。

[0131] 在神经网络层130中的多层隐含层之后,也就是整个卷积神经网络100的最后层为输出层140,该输出层140具有类似分类交叉熵的损失函数,具体用于计算预测误差,一旦整个卷积神经网络100的前向传播(如图3由110至140的传播为前向传播)完成,反向传播(如图3由140至110的传播为反向传播)就会开始更新前面提到的各层的权重值以及偏差,以减少卷积神经网络100的损失及卷积神经网络100通过输出层输出的结果和理想结果之间的误差。

[0132] 需要说明的是,如图3所示的卷积神经网络100仅作为一种卷积神经网络的示例,在具体的应用中,卷积神经网络还可以以其他网络模型的形式存在,例如,如图4所示的多个卷积层/池化层并行,将分别提取的特征均输入给全神经网络层130进行处理。

[0133] 图3和图4所示的基于卷积神经网络的算法可以在图5所示的NPU芯片中实现。

[0134] 图5是本申请实施例提供的一种芯片硬件结构图。

[0135] 神经网络处理器NPU 50NPU作为协处理器挂载到主CPU(Host CPU)上,由Host CPU分配任务。NPU的核心部分为运算电路50,通过控制器504控制运算电路503提取存储器中的矩阵数据并进行乘法运算。

[0136] 在一些实现中,运算电路503内部包括多个处理单元(Process Engine,PE)。在一些实现中,运算电路503是二维脉动阵列。运算电路503还可以是一维脉动阵列或者能够执行例如乘法和加法这样的数学运算的其它电子线路。在一些实现中,运算电路503是通用的矩阵处理器。

[0137] 举例来说,假设有输入矩阵A,权重矩阵B,输出矩阵C。运算电路从权重存储器502中取矩阵B相应的数据,并缓存在运算电路中每一个PE上。运算电路从输入存储器501中取矩阵A数据与矩阵B进行矩阵运算,得到的矩阵的部分结果或最终结果,保存在累加器508accumulator中。

[0138] 统一存储器506用于存放输入数据以及输出数据。权重数据直接通过存储单元访问控制器505Direct Memory Access Controller,DMAC被搬运到权重存储器502中。输入数据也通过DMAC被搬运到统一存储器506中。

[0139] BIU为Bus Interface Unit即,总线接口单元510,用于AXI总线与DMAC和取指存储器509Instruction Fetch Buffer的交互。

[0140] 总线接口单元510(Bus Interface Unit,简称BIU),用于取指存储器509从外部存储器获取指令,还用于存储单元访问控制器505从外部存储器获取输入矩阵A或者权重矩阵B的原数据。

[0141] DMAC主要用于将外部存储器DDR中的输入数据搬运到统一存储器506或将权重数据搬运到权重存储器502中或将输入数据数据搬运到输入存储器501中。

[0142] 向量计算单元507多个运算处理单元,在需要的情况下,对运算电路的输出做进一步处理,如向量乘,向量加,指数运算,对数运算,大小比较等等。主要用于神经网络中非卷积/FC层网络计算,如Pooling(池化),Batch Normalization(批归一化),Local Response Normalization(局部响应归一化)等。

[0143] 在一些实现种,向量计算单元能507将经处理的输出的向量存储到统一缓存器506。例如,向量计算单元507可以将非线性函数应用到运算电路503的输出,例如累加值的向量,用以生成激活值。在一些实现中,向量计算单元507生成归一化的值、合并值,或二者均有。在一些实现中,处理过的输出的向量能够用作到运算电路503的激活输入,例如用于在神经网络中的后续层中的使用。

[0144] 控制器504连接的取指存储器(instruction fetch buffer)509,用于存储控制器504使用的指令;

[0145] 统一存储器506,输入存储器501,权重存储器502以及取指存储器509均为On-Chip存储器。外部存储器私有于该NPU硬件架构。

[0146] 其中,图3和图4所示的卷积神经网络中各层的运算可以由矩阵计算单元212或向量计算单元507执行。

[0147] 上述图1至图5描述了人工智能的相关内容,本申请实施例提供了一种数据处理的方法。本申请实施例所提供的数据处理的方法可以是基于例如图2中的训练设备220实现的,该训练设备220可以相当于本申请的模型训练系统。需要说明的是本申请实施例所提供的模型训练系统的表现形式可能与上述图2中的训练设备220不同,但本申请实施例所训练的模型可以应用于图1所描述的各种场景中,该模型可以采用图3至图4中的任一可能神经网络结构。

[0148] 本申请实施例的模型训练系统可以是端云结合的模型训练系统。参阅图6,介绍本申请实施例提供一种端云结合的模型训练系统。

[0149] 该模型训练系统包括云设备和多个终端设备,云设备和多个终端设备之间通过通信网络通信连接。

[0150] 云设备可以是具有计算和数据收发功能的资源集合,可以是一个独立的计算机设备,也可以是多个独立的计算机设备组成的集群。也可以是虚拟机((virtual machine,VM)。

[0151] 终端设备(也可以称为用户设备(user equipment,UE))是一种具有无线收发功能

的设备，可以部署在陆地上，包括室内或室外、手持或车载；也可以部署在水面上(如轮船等)；还可以部署在空中(例如飞机、气球和卫星上等)。所述终端可以是手机(mobile phone)、平板电脑(pad)、带无线收发功能的电脑、虚拟现实(virtual reality，VR)终端、增强现实(augmented reality，AR)终端、工业控制(industrial control)中的无线终端、无人驾驶(self driving)中的无线终端、远程医疗(remote medical)中的无线终端、智能电网(smart grid)中的无线终端、运输安全(transportation safety)中的无线终端、智慧城市(smart city)中的无线终端、智慧家庭(smart home)中的无线终端等。

[0152] 在基于端云结合的模型训练系统中，云设备和每个终端设备上都保存有一份待训练模型，当然，该待训练模型在云设备和每个终端设备上的表现形式可以是包含算子和边的计算图，也可以是其他形式的文件。

[0153] 如前文图2所对应内容中对深度神经网络模型所描述的，模型训练的过程就是通过不断训练更新权重的过程。因为待训练模型在权重初始化时设定的权重通常都较大，所以权重更新的过程通常是用本轮当前的权重减去本轮该权重的聚合梯度，得到更新后的权重，该更新后的权重用于下一轮训练。

[0154] 模型训练的过程通常是计算各权重的梯度的过程，梯度通常是权重的导数。因为每个终端设备针对同一个权重(例如权重A)都会计算出一个梯度，每个终端设备在模型训练时通常使用不同的训练数据，所以每个训练节点针对同一个权重计算出来的梯度通常是不相同的，所以需要云设备对各终端设备计算得到的梯度进行聚合，得到一个聚合梯度。梯度聚合的过程通常是将各终端设备计算得到的梯度相加然后再求平均值的过程。如：针对权重A，终端设备0计算出的梯度是a0、终端设备1计算出的梯度是a1、终端设备2计算出的梯度是a2、终端设备3计算出的梯度是a3，那么该权重A的聚合梯度就可以为(a0+a1+a2+a3)/4。当然，梯度聚合的方法也不限于这一种，其他可适用的梯度聚合方法也适用于本申请，此处不做更多介绍。

[0155] 本申请实施例中需要终端设备向云设备传输梯度数据，就会产生终端设备与云设备的通信开销。为了尽量减少通信开销，本申请实施例提供了一种数据处理的方法。

[0156] 参阅图7，本申请实施例提供的数据处理的方法的一实施例可以包括：

[0157] 601、终端设备获取待训练模型的A个梯度数据。

[0158] A为正整数。本申请实施例中A的数量与待训练模型的结构有关，以待训练模型是一种卷积神经网络为例，该待训练模型的结构为卷积(convolution，Conv)1-Conv2-全连接(full connection，FC)1-FC2)，其中，Conv1和Conv2都分别包括过滤器(filter)和偏离(bias)两种逻辑单元，所以该模型中的A＝6。也可以理解为，该模型包括6个与权重相关的算子。该处只是以这种结构的模型为例进行说明，关于A的取值，与具体模型结构有关，对此，本申请不做限定。每个梯度数据中都会包括至少一个元素，每个元素可以表示一个梯度。

[0159] 602、终端设备针对B个梯度数据中的每个第一梯度数据，将所述第一梯度数据划分为至少两个梯度片段。

[0160] 需要说明的是，为方便引用，本申请中可以用“第一梯度数据”代指B个梯度数据中任意的一个梯度数据。待训练模型以及A个梯度数据的来源本申请不做限定。

[0161] 所述第一梯度数据包括d个元素，所述B个梯度数据包含于所述A个梯度数据中，B

为正整数且B≤A,每个梯度片段中包括d′个元素,d′为大于2的正整数,且d能被d′整除。

[0162] B个梯度数据可以是A个梯度数据中的部分,也可以是全部,例如:B=A时,表示A个梯度数据中的每一个都要做梯度片段的划分,B<A时,表示其中部分需要做梯度片段的划分,剩余的(A-B)个梯度数据不需要做梯度片段的划分。梯度片段在划分时,可以根据d的数量,确定d′,d能被d′整除,例如:d=1200,d′=16,d/d′=75,则表示该第一梯度数据可以划分为75个梯度片段。

[0163] 603、终端设备使用所述第一梯度数据的第一码本对所述至少两个梯度片段中的每个梯度片段进行量化转换,以得到所述每个梯度片段各自对应的量化表示。

[0164] 所述量化表示中元素的个数小于d′,所述第一码本为d′行m列的第一矩阵或者d′列m行的第二矩阵,m为大于2的正整数,且m≥d′。

[0165] 每个第一梯度数据都会对应有一个码本,不同梯度数据的d′可以不相同,不同码本中的d′和m也可以不相同。m≥d′可以确保在第一矩阵或第二矩阵中为各梯度片段找到匹配的列或行。在模型训练过程中,通常会规定使用第一矩阵或者第二矩阵。

[0166] 关于终端设备对哪些梯度数据进行量化转换,以及所采用的第一矩阵或第二矩阵可以参阅下表1进行理解。

[0167] 表1:梯度数据与码本的关系表

[0168]

模型层	Con1	Con1	Con2	Con2	FC1	FC2
权重/梯度	filter	bias	filter	bias	weight	weight
元素个数	150	6	2400	96	48000	1200
是否量化	N	N	Y	N	Y	Y
码本维度 d′*m			32*32		64*64	16*32

[0169] 由表1可知,针对步骤601部分所描述的卷积神经网络结构的待训练模型,可以有6个梯度数据,每个梯度数据都有多个元素,如:Con1-filter的梯度数据有150个元素,Con1-bias的梯度数据有6个元素,Con2-filter的梯度数据有2400个元素,Con2-bias的梯度数据有96个元素,FC1-weight的梯度数据有48000个元素,FC2-weight的梯度数据有1200个元素。其中,是否量化一行,N表示不需要量化,Y表示需要量化。若设定的是否需要量化的数据量阈值为512,则其中Con1-filter、Con1-bias和Con2-bias的梯度数据都小于512,不需要量化,则也不需要码本。另外,Con2-filter、FC1-weight和FC2-weight的梯度数据都超过了512,表示都需要量化,则每个梯度数据都需要对应的码本,例如:Con2-filter的梯度数据有2400个元素,若d′=32,则可以划分成75个梯度片段,m=32,表示该码本是一个32*32的矩阵。同理,FC1-weight的梯度数据对应的码本为一个64*64的矩阵。FC2-weight的梯度数据对应的码本为一个16*32的矩阵。

[0170] 604、终端设备向云设备发送所述B个梯度数据对应的B个量化集合,其中,所述第一梯度数据对应的第一量化集合包括所述每个梯度片段各自对应的量化表示。

[0171] 这里“发送所述B个梯度数据对应的B个量化集合”可以是B个梯度数据全部完成量化转换后统一发送,也可以分多次发送或不等待全部完成,部分(包括一个)完成后就发送。

[0172] 每个量化集合对应一个梯度数据,若B=3,梯度数据1有75个梯度片段,则该梯度数据1对应的量化集合中就有75个量化表示。梯度数据2有120个梯度片段,则该梯度数据2

对应的量化集合中就有120个量化表示。梯度数据3有150个梯度片段，则该梯度数据3对应的量化集合中就有150个量化表示。

[0173] 当A＞B时，终端设备还向所述云设备发送所述A个梯度数据中除所述B个梯度数据之外的(A-B)个梯度数据。

[0174] (A-B)个梯度数据可能本身的元素个数就很少，量化转换的过程还会消耗计算资源，从综合收益的角度考虑，可以不经压缩处理直接发送(A-B)个梯度数据，也可以采用其他的压缩处理方式，对(A-B)个梯度数据进行压缩处理，然后再发送。

[0175] 605、云设备接收终端设备发送的B个梯度数据对应的B个量化集合后，针对所述B个量化集合中的每个第一量化集合，使用所述第一量化集合对应的第一码本对所述第一量化集合中的每个量化表示进行反量化，以得到所述每个量化表示各自所对应的第一梯度片段。

[0176] 所述第一梯度片段包括d′个元素，所述量化表示中元素的个数小于d′，所述第一码本为d′行m列的第一矩阵或者d′列m行的第二矩阵，d′和m为大于2的正整数，且m≥d′。

[0177] 该云设备接收到量化表示后，会使用终端设备在量化转换时相同的码本做反量化，也就是量化恢复。

[0178] 606、云设备将所述第一量化集合中每个量化表示所对应的第一梯度片段进行拼接，以得到与所述第一量化集合对应的第一梯度数据。

[0179] 所述第一梯度数据包括d个元素，d为正整数，且d能被d′整除。

[0180] 拼接的过程就是排序的过程，将每个梯度片段按照终端设备划分梯度片段时的顺序排列好，从而恢复出第一梯度数据。

[0181] 607、云设备针对B个梯度数据对应的B个权重中的第一权重，对所述第一权重的N个第一梯度数据进行梯度聚合，以得到所述第一权重的梯度聚合数据。

[0182] 所述第一权重的多个第一梯度数据对应于N个终端设备，所述N为大于1的整数。

[0183] 梯度聚合可以参阅前述对梯度聚合过程的相关描述进行理解，此处不再重复赘述。

[0184] 608、云设备向所述终端设备发送所述B个权重各自的梯度聚合数据。

[0185] 609、终端设备根据梯度聚合数据更新权重。

[0186] 更新权重可以参阅前述对更新权重过程的相关描述进行理解，此处不再重复赘述。

[0187] 本申请实施例采用将要传输的梯度数据进行梯度片段划分，然后再依据码本进行量化转换，得到量化表示，量化表示中元素的个数小于每个梯度片段中元素的个数。这样在量化压缩的基础上还进一步缩减了量化数据的数量，从而提高了压缩比。另外，因为量化表示的压缩比被提高了，终端设备向云设备传输量化表示时的通信开销也随之减少。

[0188] 本申请实施例中，无论是在量化转换，还是在反量化过程中都使用到了码本，码本可以是终端设备和云设备根据相同的随机种子各自生成的，也可以是由云设备生成后再下发给终端设备的，下面对这两种情况分别进行介绍。

[0189] 1、终端设备和云设备自行生成码本。

[0190] 终端设备根据所述每个梯度数据各自的数据量，确定所述A个梯度数据中有B个梯度数据需要量化；

[0191] 终端设备针对所述B个梯度数据中的每个第一梯度数据，根据所述第一梯度数据的元素数量d确定d′和m，d′为2的p次方，m为2的q次方，p和q都为正整数，且q≥p；

[0192] 终端设备根据随机种子为所述第一梯度数据随机生成d′行m列的所述第一矩阵或者d列m行的所述第二矩阵，且所述第一矩阵或所述第二矩阵中的每个高斯向量被缩放到模为1，以得到所述第一梯度数据的第一码本，所述第一矩阵中每一列为一个高斯向量，所述第二矩阵中每一行为一个高斯向量。

[0193] 云设备针对B个梯度数据中的每个第一梯度数据，根据随机种子为所述第一梯度数据随机生成d′行m列的所述第一矩阵或者d′列m行的所述第二矩阵，且所述第一矩阵或所述第二矩阵中的每个高斯向量被缩放到模为1，以得到所述第一梯度数据的第一码本，所述第一矩阵中每一列为一个高斯向量，所述第二矩阵中每一行为一个高斯向量。

[0194] 终端设备可以从综合收益的角度确定一个需要进行量化转换的数据量阈值，例如：设定该数据量阈值为512个元素，若梯度数据中的元素个数大于或等于512，则表示需要进行量化转换，若梯度数据中的元素个数小于512，则表示不需要进行量化转换。终端设备只需要为需要进行量化转换的梯度数据生成码本。这样可以减少因生成码本而带来的计算开销，提高计算开销和通信开销的综合收益。为了确保云设备可以正常恢复梯度数据，终端设备生成码本所使用的随机种子与云设备相同。随机种子可以是一个数值，例如：随机种子s＝798，该随机种子通常是云设备设定的，然后同步给各个终端数合并。向随机函数中输入随机种子就可以随机生成梯度数据的码本。因为随机种子相同，所以，终端设备自行生成的码本和云设备自行生成的码本是相同的。

[0195] 2、云设备生成码本，再向终端设备下发。

[0196] 云设备针对B个梯度数据中的每个第一梯度数据，根据随机种子为所述第一梯度数据随机生成d′行m列的所述第一矩阵或者d′列m行的所述第二矩阵，且所述第一矩阵或所述第二矩阵中的每个高斯向量被缩放到模为1，以得到所述第一梯度数据的第一码本，所述第一矩阵中每一列为一个高斯向量，所述第二矩阵中每一行为一个高斯向量。

[0197] 云设备向所述终端设备发送所述B个梯度数据各自对应的码本。对应地，终端设备接收所述云设备发送的所述B个梯度数据各自对应的码本，所述B个梯度数据各自对应的码本中包括所述第一码本。

[0198] 这种实现方案中，云设备向终端设备发送生成的码本，可以减少终端设备的计算开销。

[0199] 本申请实施例中，上述步骤603：终端设备使用所述第一梯度数据的第一码本对所述至少两个梯度片段中的每个梯度片段进行量化转换，以得到所述每个梯度片段各自对应的量化表示，可以包括：

[0200] 终端设备从所述第一码本中为第一梯度片段确定目标码，所述目标码为所述第一矩阵中的一列或者为所述第二矩阵中的一行，所述第一梯度片段为所述至少两个梯度片段中的任意一个；

[0201] 终端设备确定所述目标码的伪模长和码索引；

[0202] 终端设备将所述伪模长和所述码索引确定为所述第一梯度片段的量化表示。

[0203] 本申请实施例中伪模长与模长是相对的，伪模长表示不是按照数学上通常的向量模长的计算方法得到的，而是通过码本和梯度片段计算得到的。码索引表示第一矩阵的一

列的索引，例如：index(5)表示第五列。若有多个码本，还需要带上码本的标记，如index(3,5)，表示码本3的第五列。或者，码索引表示第二矩阵的一行的索引，例如：index(3,6)表示示码本3的第六行。若为模长用u表示，码索引index(c)表示，则量化表示可以为(u,index(c))。本申请实施例中，一个梯度片段转换成了只包括伪模长和码索引两个元素，极大的缩减了数据量，提高了梯度数据的压缩比，也减少了传输梯度数据的通信开销。

[0204] 其中，步骤：从所述第一码本中为第一梯度片段确定目标码可以有两种实现方式。

[0205] 参阅图8，方式一可以包括：

[0206] 701、确定所述第一梯度片段的向量模长(vector norm)。

[0207] 702、确定向量模长是否等于0，若等于0，则执行步骤703，若不等于0，则执行步骤704。

[0208] 向量模长为第一梯度片段上的各元素的平方和然后再求平方根。

[0209] 703、若所述向量模长等于0，则确定所述第一矩阵中的任意一列或者所述第二矩阵中的任意一行为目标码。

[0210] 若向量模长等于0则表示各元素都为0，所以只需要随机选择一列或一行凑成符合量化表示的格式即可，例如：可以选择第一列或第一行。

[0211] 若第一梯度片段用g表示，则会有If$||g||_2=0$，则return$(0,c_1)$。$||g||_2$表示该第一梯度片段中的各元素的平方和然后求平方根，也就是该第一梯度片段g的向量模长。针对第一梯度片段g，若向量模长等于0，则可以输出向量表示(0,index(x,1))，表示向量模长＝0，码索引是第x码本中的第一行或第一列，x表示码本的索引，可以是码本的标识或序号。当然，也可以不选择第一列或第一行，只需要随机选择一列或一行即可。

[0212] 704、若所述向量模长不等于0，则根据所述第一码本和所述第一梯度片段确定所述第一梯度片段的第一系数向量。

[0213] 所述第一系数向量中包括m个元素。

[0214] 若第一码本用C表示，第一梯度片段用g表示，第一系数向量用p表示，则有$p=C^T(CC^T)^{-1}g$，其中，C^T表示C的转置矩阵。P中包括m个元素。

[0215] 705、对所述第一系数向量中的每个元素进行归一化，以得到第二系数向量。

[0216] 所述第二系数向量中包括m个归一化后的元素。

[0217] 第二系数向量用$\tilde{p}$表示，$\tilde{p}$中包括m个归一化后的元素，元素归一化可以理解为：$\widetilde{p_b}=|p_b|/\|p\|_1$，其中$\widetilde{p_b}$表示m个归一化后的元素中的第b个，$|p_b|$表示第一系数向量p中的第b个元素的绝对值，$||p||_1$表示第一系数向量p的m元素中各元素的绝对值之和。

[0218] 706、针对所述归一化后的m个元素采用轮盘组的选择策略，确定其中第i个元素对应的所述第一矩阵中的第i列或所述第二矩阵中的第i行为目标码。

[0219] 所述第i个元素为采用轮盘组的选择策略被选中的元素。

[0220] 轮盘组的选择策略在该方案中可以认为每个归一化的元素都有各自的概率，各元素的概率有大有小，所有概率之和等于1，概率大的被选中的概率也大，概率小的被选中的概率也小，但最终的选择结果还是以轮盘组转动后选择到的元素为准。若选中的是$\widetilde{p_b}$，则就选第b个元素所对应列或行的码作为目标码。当然，若选中的是第i个归一化后的元素，则选第i个元素所对应列或行的码作为目标码。

[0221] 若该步骤选择的目标码是c_b:则伪模长$u = sign(\widetilde{p_b})\|p\|_1$,其中,||p||$_1$表示第一系数中各元素的绝对值之和。Sign函数符号符号。若目标码是c_b,则伪模长表示对归一化后的第b个元素与第一系数中各元素的绝对值之和的乘积。

[0222] 码索引为index(x,b),return(u,index(x,b)),表示输出量化表示(u,index(x,b))。

[0223] 方式二可以包括:

[0224] 根据所述第一码本和所述第一梯度片段确定所述第一梯度片段的投影向量(projection vector),所述投影向量中包括m个元素;

[0225] 确定第i个元素对应的所述第一码本中的第i列或所述第二矩阵中的第i行为目标码,所述第i元素的绝对值在所述m个元素的绝对值中最大。

[0226] 该种可能的实现方式中,若第一码本用C表示,第一梯度片段用g表示,投影向量用p表示,则$p=C^Tg$,其中,C^T表示C的转置矩阵。投影向量p中包括m个元素;$|p_i|$表示m个元素中第i个元素的绝对值。若i=5,则表示选择第一码本中的第5列或所述第二矩阵中的第5行作为目标码。

[0227] 若该步骤选择的目标码c_i,则伪模长$u=p_i$,即伪模长等于投影向量p中的第i个元素。

[0228] 码索引为index(x,i),return(u,index(x,i)),表示输出量化表示(u,index(x,i))。

[0229] 若在只有一个码本的情况下,码索引中可以不包含码本的索引x。

[0230] 云设备在反量化时可以:针对所述第一量化集合中的每个第一量化表示,根据所述第一量化表示的码索引从所述第一码本中确定所述第一量化表示的目标码;根据所述第一量化表示中的伪模长和所述目标码恢复出所述第一量化表示对应的第一梯度片段。

[0231] 也就是云设备可以通过index(i)确定目标码c_i,然后使用伪模长u与目标码中的每个元素相乘,恢复出对应的梯度片段。

[0232] 云设备接收到终端设备发送来的量化表示,对量化表示进行逆向操作,得到梯度数据的近似值。以FC2-weight的梯度为例,该梯度有75个量化的梯度片段,对于每个梯度片段设其量化表示为(0.523641,(3,25)),表示从码本3中选择第25个码c,则该梯度片段的近似值为0.523641*c。将75个梯度片段的近似值拼接后即可得到FC2-weight梯度数据的近似表示。

[0233] 需要说明的是,上述示例中的码本是以d′行m列的第一矩阵的结构为例进行说明的,若码本是d′列m行的所述第二矩阵,则需要对上述公式中的C先进转置,从第二矩阵的形式转换为第一矩阵的形式后再通过上述相应公式进行计算。

[0234] 为了进一步节省通信开销,云设备还可以:

[0235] 对所述B个权重的梯度聚合数据分别进行量化转换,以得到所述B个梯度聚合数据各自的量化集合;

[0236] 向所述终端设备发送所述B个梯度聚合数据各自的量化集合,其中,每个梯度聚合数据的量化集合包括该梯度聚合数据的每个梯度片段对应的量化表示。

[0237] 该种可能的实现方式中,云设备可以对要发送给终端设备的梯度聚合数据也进行类似于终端设备对梯度数据所进行的量化转换,这样可以进一步减少云设备与终端设备的

通信开销。

[0238] 终端设备接收所述云设备发送的所述B个梯度数据对应的B个梯度聚合数据各自的量化集合后，可以采用上述云设备反量化的过程恢复出B个梯度聚合数据，进而执行权重更新。

[0239] 目前在梯度数据传输过程中有几种主流的压缩方法，分别为随机梯度下降(Stochastic Gradient Descent，SGD)、量化的SGD(Quantized SGD，QSGD)、符号SGD(Sign SGD)和三元梯度(Ternary Gradients，TernGrad)。本申请实施例的量化转换也是一种梯度压缩的方法，该方法可以称为超球量化(Hyper-Sphere Quantizatio，HSQ)的梯度压缩方法。

[0240] 在开发HSQ的过程中，工程人员在模拟环境上基于相同的数据集针对上述几种压缩方法试验了3种流行的深度学习模型，VGG19、ResNet50和ResNet101。表2中分别列出了压缩率(纯SGD为基线)和算法的收敛精度，d′表示梯度片段大小。

[0241] 表2：几种压缩方法对应的压缩比和收敛精度表

[0242]

压缩方法	SGD	Sign SGD	TernGrad	QSGD (4bit)	QSGD (8bit)	HSQ $d'=8$	HSQ $d'=16$	HSQ $d'=64$
压缩比	1	32	20.2	~8	~4	18.3	36.6	**146.3**
收敛精度(VGG19)	92.65	90.79	91.1	92.6	92.71	**92.76**	92.38	91.13
收敛精度(ResNet50)	94.19	92.6	93.29	94.64	94.03	94.68	**94.77**	93.77
收敛精度(ResNet101)	94.63	92.01	93.15	94.35	94.67	94.48	**94.7**	93.87

[0243] 由上表2可以看出，当d′＝8，d′＝16时，HSQ可以获得比SGD更高的收敛精度，压缩比分别为18.3和36.6倍。当d′＝64时，HSQ的压缩比显着高于其他算法，并且收敛精度的降低很小。相比已有的几种梯度压缩方法，本申请方法在梯度压缩比和收敛精度上都更有优势。

[0244] 以上描述了基于端云结合的模型训练系统，在模型训练中的数据处理的方法，下面结合附图介绍本申请实施例提供的终端设备和云设备。

[0245] 如图9所示，本申请实施例提供的终端设备80的一实施例可以包括：

[0246] 处理单元801用于：

[0247] 获取待训练模型的A个梯度数据，A为正整数；

[0248] 针对B个梯度数据中的每个第一梯度数据，将所述第一梯度数据划分为至少两个梯度片段，所述第一梯度数据包括d个元素，所述B个梯度数据包含于所述A个梯度数据中，B为正整数且B≤A，每个梯度片段中包括d′个元素，d′为大于2的正整数，且d能被d′整除；

[0249] 使用所述第一梯度数据的第一码本对所述至少两个梯度片段中的每个梯度片段进行量化转换，以得到所述每个梯度片段各自对应的量化表示，所述量化表示中元素的个数小于d′，所述第一码本为d′行m列的第一矩阵或者d′列m行的第二矩阵，m为大于2的正整数，且m≥d′；

[0250] 发送单元802，用于向云设备发送所述B个梯度数据对应的B个量化集合，其中，所

述第一梯度数据对应的第一量化集合包括所述每个梯度片段各自对应的量化表示。

[0251] 本申请实施例采用将要传输的梯度数据进行梯度片段划分，然后再依据码本进行量化转换，得到量化表示，量化表示中元素的个数小于每个梯度片段中元素的个数。这样在量化压缩的基础上还进一步缩减了量化数据的数量，从而提高了压缩比。另外，因为量化表示的压缩比被提高了，终端设备向云设备传输量化表示时的通信开销也随之减少。

[0252] 一种可能的实施例中，发送单元802，还用于向所述云设备发送所述A个梯度数据中除所述B个梯度数据之外的(A-B)个梯度数据。

[0253] 该种可能的实现方式中，(A-B)个梯度数据可能本身的元素个数就很少，量化转换的过程还会消耗计算资源，从综合收益的角度考虑，可以不经压缩处理直接发送(A-B)个梯度数据，也可以采用其他的压缩处理方式，对(A-B)个梯度数据进行压缩处理，然后再发送。

[0254] 一种可能的实施例中，所述处理单元801还用于：

[0255] 根据所述每个梯度数据各自的数据量，确定所述A个梯度数据中有B个梯度数据需要量化；

[0256] 针对所述B个梯度数据中的每个第一梯度数据，根据所述第一梯度数据的元素数量d确定d′和m，d′为2的p次方，m为2的q次方，p和q都为正整数，且q≥p；

[0257] 根据随机种子为所述第一梯度数据随机生成d′行m列的所述第一矩阵或者d′列m行的所述第二矩阵，且所述第一矩阵或所述第二矩阵中的每个高斯向量被缩放到模为1，以得到所述第一梯度数据的第一码本，所述第一矩阵中每一列为一个高斯向量，所述第二矩阵中每一行为一个高斯向量。

[0258] 该种可能的实现方式中，可以从综合收益的角度确定一个需要进行量化转换的数据量阈值，例如：设定该数据量阈值为512个元素，若梯度数据中的元素个数大于或等于512，则表示需要进行量化转换，若梯度数据中的元素个数小于512，则表示不需要进行量化转换。终端设备只需要为需要进行量化转换的梯度数据生成码本。这样可以减少因生成码本而带来的计算开销，提高计算开销和通信开销的综合收益。为了确保云设备可以正常恢复梯度数据，终端设备生成码本所使用的随机种子与云设备相同。随机种子可以是一个数值，向随机函数中输入随机种子就可以随机生成梯度数据的码本。

[0259] 一种可能的实施例中，接收单元803，用于接收所述云设备发送的所述B个梯度数据各自对应的码本，所述B个梯度数据各自对应的码本中包括所述第一码本。

[0260] 该种可能的实现方式中，也可以不需要终端设备自行生成码本，可以由云设备生成码本，然后下发给终端设备，这样可以减少终端设备的计算开销。

[0261] 一种可能的实施例中，所述处理单元801用于：

[0262] 从所述第一码本中为第一梯度片段确定目标码，所述目标码为所述第一矩阵中的一列或者为所述第二矩阵中的一行，所述第一梯度片段为所述至少两个梯度片段中的任意一个；

[0263] 确定所述目标码的伪模长和码索引；

[0264] 将所述伪模长和所述码索引确定为所述第一梯度片段的量化表示。

[0265] 一种可能的实施例中，所述处理单元801用于：

[0266] 确定所述第一梯度片段的向量模长；

[0267] 若所述向量模长等于0，则确定所述第一矩阵中的任意一列或者所述第二矩阵中

的任意一行为目标码。

[0268] 一种可能的实施例中，所述处理单元801还用于：

[0269] 若所述向量模长不等于0，则根据所述第一码本和所述第一梯度片段确定所述第一梯度片段的第一系数向量，所述第一系数向量中包括m个元素；

[0270] 对所述第一系数向量中的每个元素进行归一化，以得到第二系数向量，所述第二系数向量中包括m个归一化后的元素；

[0271] 针对所述归一化后的m个元素采用轮盘组的选择策略，确定其中第i个元素对应的所述第一矩阵中的第i列或所述第二矩阵中的第i行为目标码，所述第i个元素为采用轮盘组的选择策略被选中的元素。

[0272] 一种可能的实施例中，所述处理单元801用于：

[0273] 根据所述第一码本和所述第一梯度片段确定所述第一梯度片段的投影向量，所述投影向量中包括m个元素；

[0274] 确定第i个元素对应的所述第一码本中的第i列或所述第二矩阵中的第i行为目标码，所述第i元素的绝对值在所述m个元素的绝对值中最大。

[0275] 一种可能的实施例中，接收单元803，用于接收所述云设备发送的所述B个梯度数据对应的B个梯度聚合数据各自的量化集合，其中，每个梯度聚合数据的量化集合包括该梯度聚合数据的每个梯度片段对应的量化表示。

[0276] 需要说明的是，上述终端设备80的各模块之间的信息交互、执行过程等内容，由于与本申请方法实施例基于同一构思，其带来的技术效果与本发明方法实施例相同，具体内容可参见本申请前述所示的方法实施例中的叙述，此处不再赘述。

[0277] 参阅图10，本申请实施例提供的云设备90的一实施例可以包括：

[0278] 接收单元901，用于接收终端设备发送的B个梯度数据对应的B个量化集合，其中，每个量化集合包括对应梯度数据的每个梯度片段各自对应的量化表示；

[0279] 处理单元902用于：

[0280] 针对所述B个量化集合中的每个第一量化集合，使用所述第一量化集合对应的第一码本对所述第一量化集合中的每个量化表示进行反量化，以得到所述每个量化表示各自所对应的第一梯度片段，所述第一梯度片段包括d′个元素，所述量化表示中元素的个数小于d′，所述第一码本为d′行m列的第一矩阵或者d′列m行的第二矩阵，d′和m为大于2的正整数，且m≥d′；

[0281] 将所述第一量化集合中每个量化表示所对应的第一梯度片段进行拼接，以得到与所述第一量化集合对应的第一梯度数据，所述第一梯度数据包括d个元素，d为正整数，且d能被d′整除；

[0282] 针对B个梯度数据对应的B个权重中的第一权重，对所述第一权重的N个第一梯度数据进行梯度聚合，以得到所述第一权重的梯度聚合数据，所述第一权重的多个第一梯度数据对应于N个终端设备，所述N为大于1的整数；

[0283] 发送单元903，用于向所述终端设备发送所述B个权重各自的梯度聚合数据。

[0284] 本申请实施例中，该云设备接收到量化表示后，会使用终端设备在量化转换时相同的码本做反量化，也就是量化恢复，进而进行梯度聚合。梯度聚合的过程指的是云设备对从多个终端设备接收到的同一个权重的各个梯度进行聚合，云设备将该聚合后的梯度聚合

数据发送给终端设备后，终端设备可以更新各权重，从而使待训练模型向收敛更进一步。权重更新的过程可以是用本轮当前的权重减去本轮计算得到的梯度聚合数据，得到更新后的权重，该更新后的权重用于下一轮训练。因为云设备从终端设备接收的是终端设备量化转换后的量化表示，因为该量化表示中元素的个数小于每个梯度片段中的元素个数d′，所以只需要用很小的通信开销就能实现梯度数据的接收。

[0285] 一种可能的实施例中，处理单元902还用于：针对B个梯度数据中的每个第一梯度数据，根据随机种子为所述第一梯度数据随机生成d′行m列的所述第一矩阵或者d′列m行的所述第二矩阵，且所述第一矩阵或所述第二矩阵中的每个高斯向量被缩放到模为1，以得到所述第一梯度数据的第一码本，所述第一矩阵中每一列为一个高斯向量，所述第二矩阵中每一行为一个高斯向量。

[0286] 一种可能的实施例中，发送单元903，还用于向所述终端设备发送所述B个梯度数据各自对应的码本。

[0287] 一种可能的实施例中，处理单元902用于：

[0288] 针对所述第一量化集合中的每个第一量化表示，根据所述第一量化表示的码索引从所述第一码本中确定所述第一量化表示的目标码；

[0289] 根据所述第一量化表示中的伪模长和所述目标码恢复出所述第一量化表示对应的第一梯度片段。

[0290] 一种可能的实施例中，处理单元902还用于：对所述B个权重的梯度聚合数据分别进行量化转换，以得到所述B个梯度聚合数据各自的量化集合；

[0291] 发送单元903，用于向所述终端设备发送所述B个梯度聚合数据各自的量化集合，其中，每个梯度聚合数据的量化集合包括该梯度聚合数据的每个梯度片段对应的量化表示。

[0292] 该种可能的实施例中，云设备可以对要发送给终端设备的梯度聚合数据也进行类似于终端设备对梯度数据所进行的量化转换，这样可以进一步减少云设备与终端设备的通信开销。

[0293] 需要说明的是，上述云设备90的各模块之间的信息交互、执行过程等内容，由于与本申请方法实施例基于同一构思，其带来的技术效果与本发明方法实施例相同，具体内容可参见本申请前述所示的方法实施例中的叙述，此处不再赘述。

[0294] 如图11所示，为本申请实施例的又一种设备的结构示意图，该设备为终端设备，该终端设备可以包括：处理器1001(例如CPU)、存储器1002、发送器1004和接收器1003；发送器1004和接收器1003耦合至处理器1001，处理器1001控制发送器1004的发送动作和接收器1003的接收动作。存储器1002可能包含高速RAM存储器，也可能还包括非易失性存储器NVM，例如至少一个磁盘存储器，存储器1002中可以存储各种指令，以用于完成各种处理功能以及实现本申请实施例的方法步骤。可选的，本申请实施例涉及的终端设备还可以包括：电源1005、以及通信端口1006中的一个或多个，图11中所描述的各器件可以是通过通信总线连接，也可以是通过其他连接方式连接，对此，本申请实施例中不做限定。接收器1003和发送器1004可以集成在终端设备的收发器中，也可以为终端设备上分别独立的收、发天线。通信总线用于实现元件之间的通信连接。上述通信端口1006用于实现终端设备与其他外设之间进行连接通信。

[0295] 在一些实施例中，终端设备中的处理器1001可以执行图9中处理单元801执行的动作，终端设备中的接收器1003可以执行图9中接收单元803执行的动作，终端设备中的发送器1004可以执行图9中发送单元802执行的动作，其实现原理和技术效果类似，在此不再赘述。

[0296] 本申请还提供了一种芯片系统，该芯片系统包括处理器，用于支持上述终端设备实现其所涉及的功能，例如，例如接收或处理上述方法实施例中所涉及的数据。在一种可能的设计中，所述芯片系统还包括存储器，所述存储器，用于保存终端设备必要的程序指令和数据。该芯片系统，可以由芯片构成，也可以包含芯片和其他分立器件。

[0297] 图12所示，为本申请的实施例提供的上述实施例中所涉及的云设备110的一种可能的逻辑结构示意图。云设备110包括：处理器1101、通信端口1102、存储器1103以及总线1104。处理器1101、通信端口1102以及存储器1103通过总线1104相互连接。在本申请的实施例中，处理器1101用于对云设备110的动作进行控制管理，例如，处理器1101用于执行图10中的处理单元902所执行的功能。通信端口1102用于支持云设备110进行通信。存储器1103，用于存储云设备110的程序代码和数据。

[0298] 其中，处理器1101可以是中央处理器单元，通用处理器，数字信号处理器，专用集成电路，现场可编程门阵列或者其他可编程逻辑器件、晶体管逻辑器件、硬件部件或者其任意组合。其可以实现或执行结合本申请公开内容所描述的各种示例性的逻辑方框，模块和电路。所述处理器也可以是实现计算功能的组合，例如包含一个或多个微处理器组合，数字信号处理器和微处理器的组合等等。总线1104可以是外设部件互连标准(Peripheral Component Interconnect，PCI)总线或扩展工业标准结构(Extended Industry Standard Architecture，EISA)总线等。所述总线可以分为地址总线、数据总线、控制总线等。为便于表示，图12中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

[0299] 本申请还提供了一种芯片系统，该芯片系统包括处理器，用于支持上述云设备实现其所涉及的功能，例如，例如接收或处理上述方法实施例中所涉及的数据。在一种可能的设计中，所述芯片系统还包括存储器，所述存储器，用于保存终端设备必要的程序指令和数据。该芯片系统，可以由芯片构成，也可以包含芯片和其他分立器件。

[0300] 在本申请的另一实施例中，还提供一种计算机可读存储介质，计算机可读存储介质中存储有计算机执行指令，当设备的至少一个处理器执行该计算机执行指令时，设备执行上述图6至图8部分实施例所描述的方法。

[0301] 在本申请的另一实施例中，还提供一种计算机程序产品，该计算机程序产品包括计算机执行指令，该计算机执行指令存储在计算机可读存储介质中；设备的至少一个处理器可以从计算机可读存储介质读取该计算机执行指令，至少一个处理器执行该计算机执行指令使得设备执行上述图6至图8部分实施例所描述的方法。

[0302] 本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请实施例的范围。

[0303] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统、

装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0304] 在本申请实施例所提供的几个实施例中,应该理解到,所揭露的系统、装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0305] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0306] 另外,在本申请实施例各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。

[0307] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本申请实施例的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本申请实施例各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(Read-Only Memory,ROM)、随机存取存储器(Random Access Memory,RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0308] 以上所述,仅为本申请实施例的具体实施方式,但本申请实施例的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请实施例揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本申请实施例的保护范围之内。因此,本申请实施例的保护范围应以所述权利要求的保护范围为准。

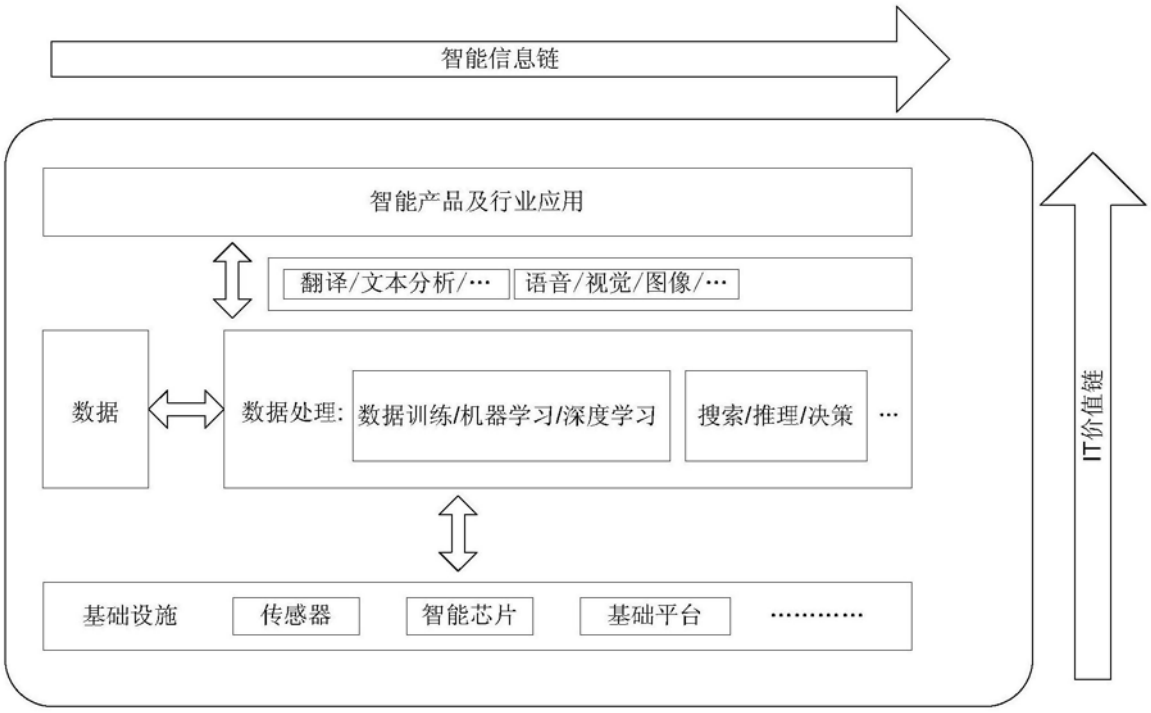

智能信息链
智能产品及行业应用
翻译/文本分析/…
语音/视觉/图像/…
数据
数据处理:
数据训练/机器学习/深度学习
搜索/推理/决策
…
基础设施
传感器
智能芯片
基础平台
……
IT价值链

图1

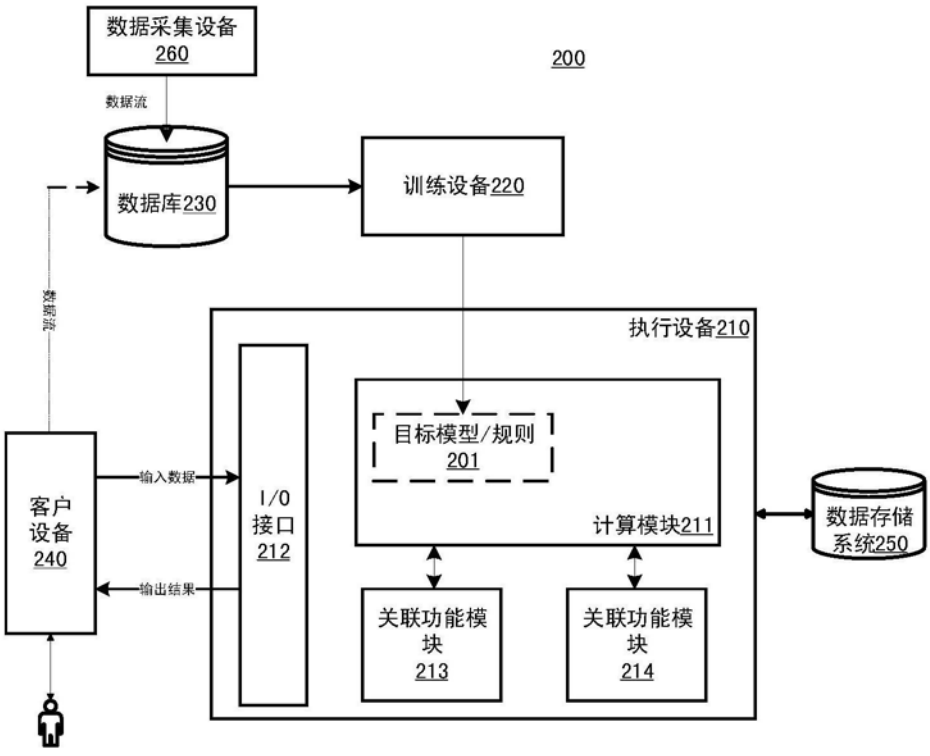

数据采集设备
260
200
数据流
数据库230
训练设备220
执行设备210
目标模型/规则
201
计算模块211
数据存储
系统250
客户
设备
240
输入数据
输出结果
I/O
接口
212
关联功能模
块
213
关联功能模
块
214

图2

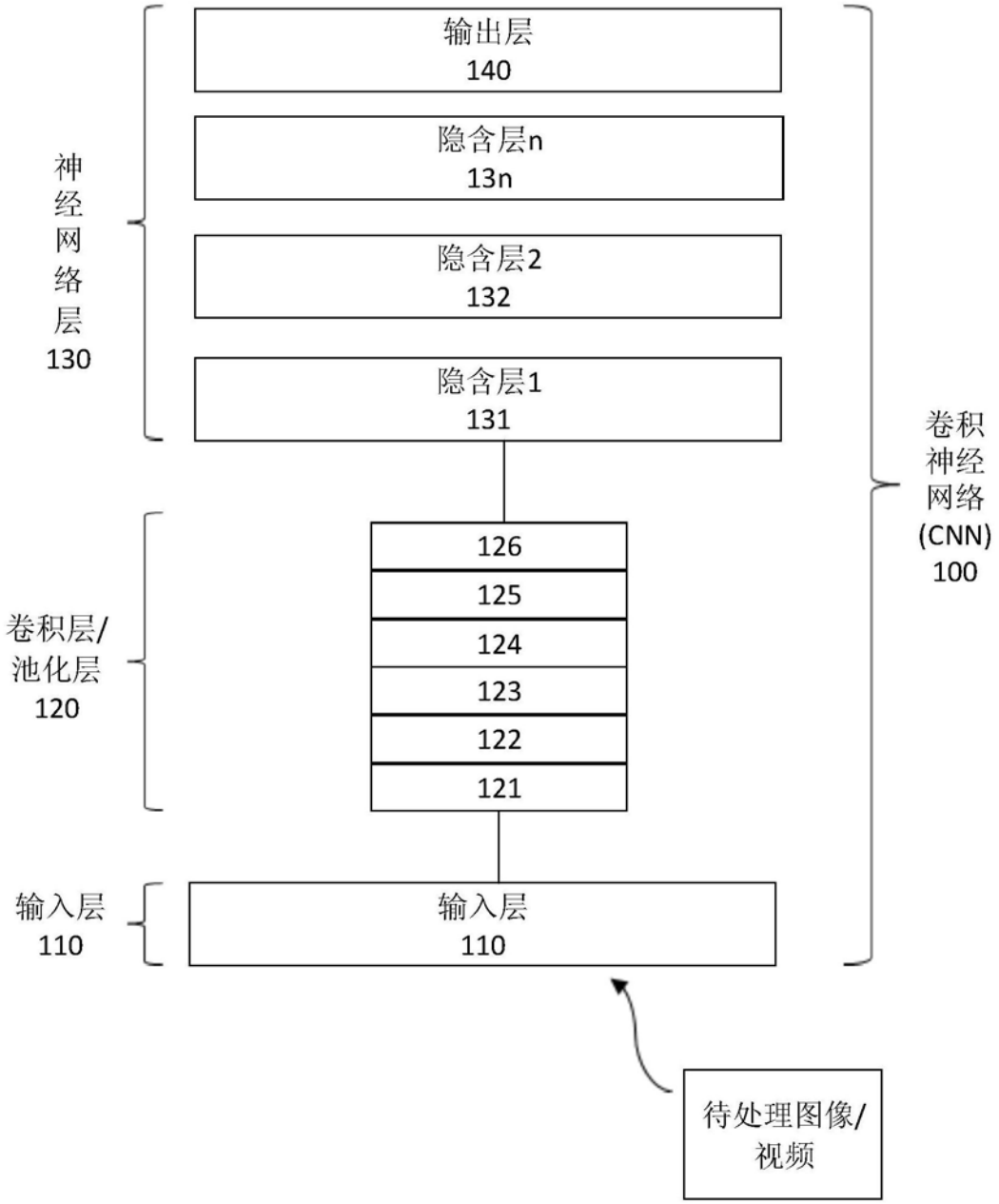

输出层
140
隐含层n
13n
隐含层2
132
隐含层1
131
神经网络层
130
126
125
124
123
122
121
卷积层/池化层
120
输入层
110
输入层
110
卷积神经网络(CNN)
100
待处理图像/视频

图3

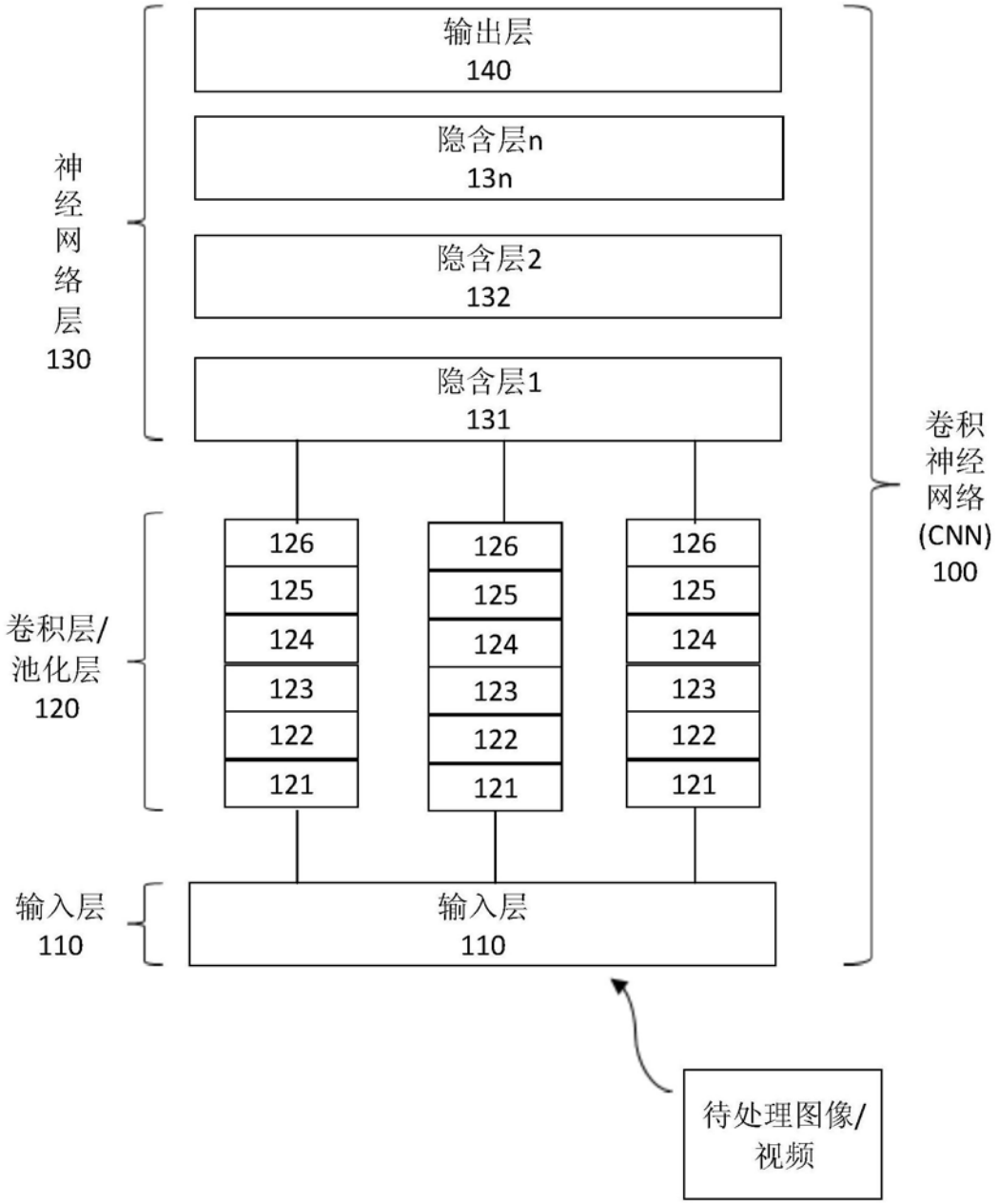

输出层
140
隐含层n
13n
隐含层2
132
隐含层1
131
神经网络层
130
卷积神经网络(CNN)
100
126
125
124
123
122
121
卷积层/池化层
120
输入层
110
输入层
110
待处理图像/视频

图4

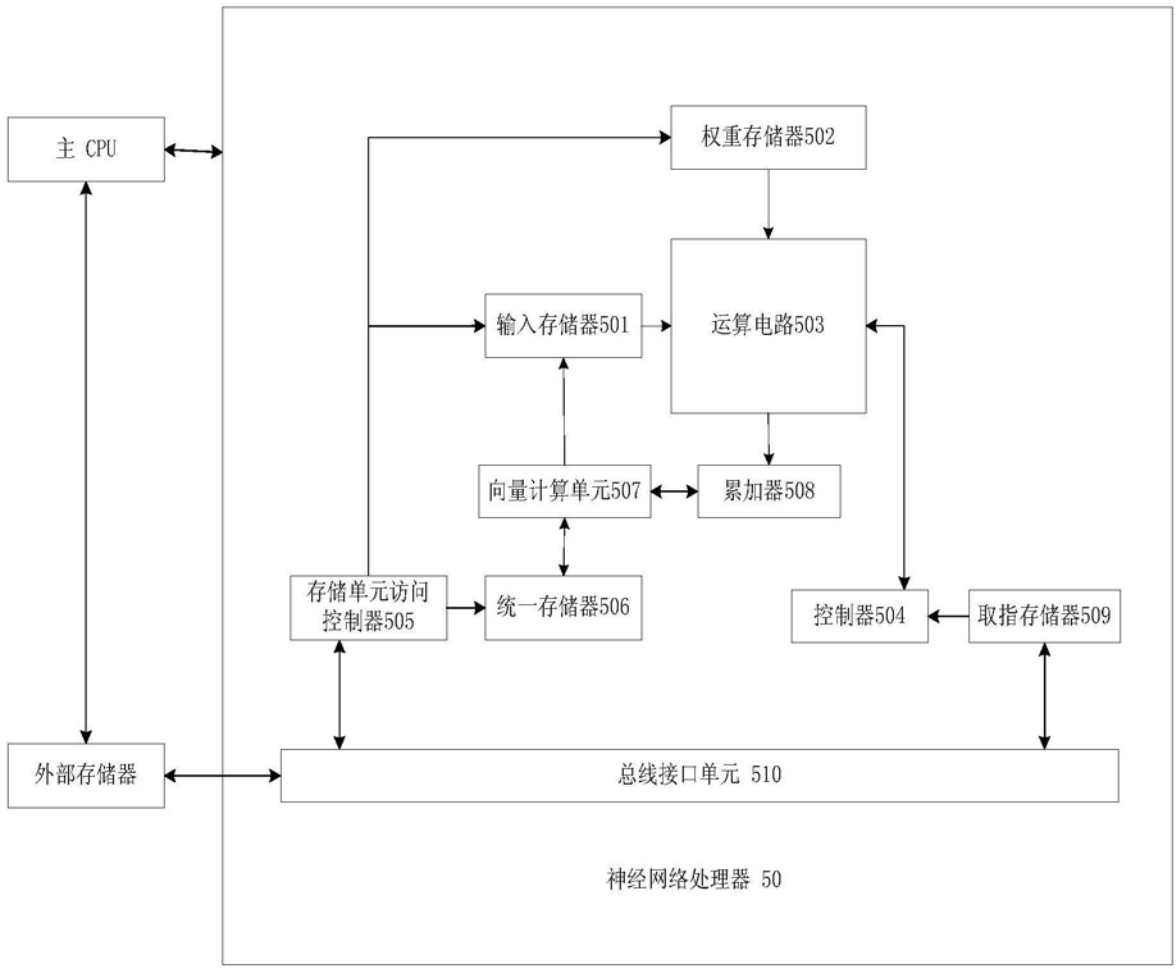

主 CPU
权重存储器502
输入存储器501
运算电路503
向量计算单元507
累加器508
存储单元访问
控制器505
统一存储器506
控制器504
取指存储器509
外部存储器
总线接口单元 510
神经网络处理器 50

图5

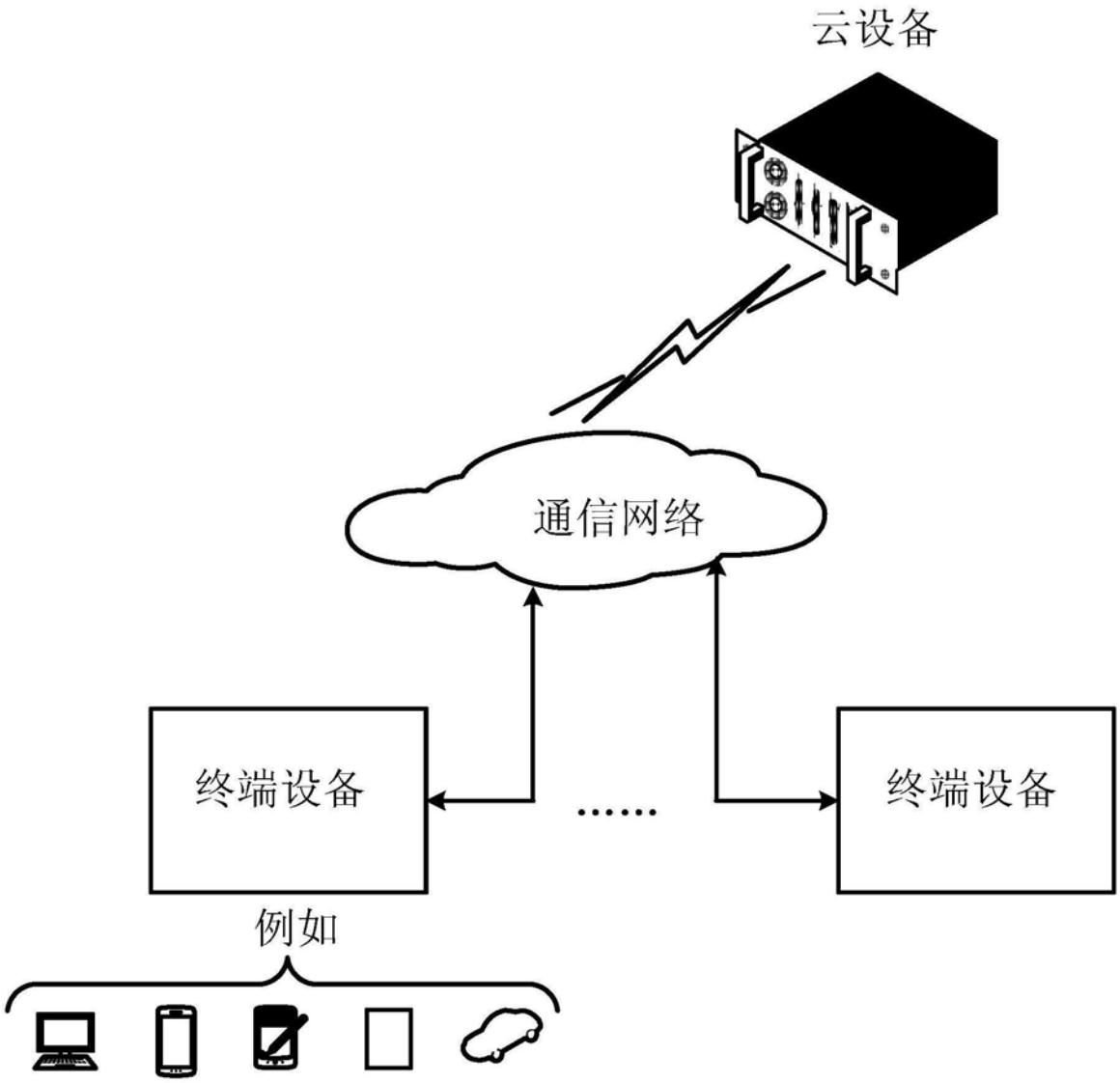


图6

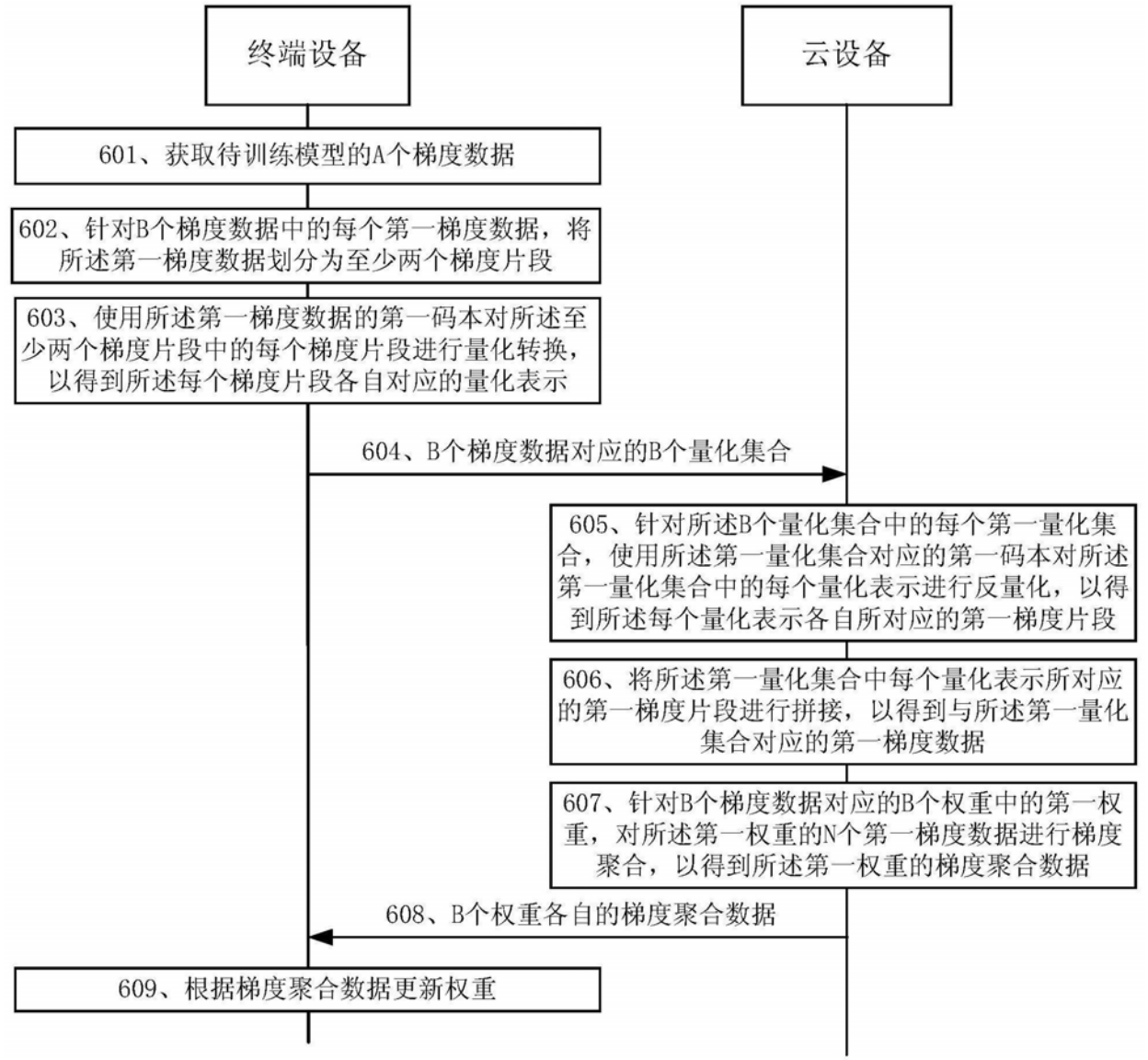

终端设备
云设备
601、获取待训练模型的A个梯度数据
602、针对B个梯度数据中的每个第一梯度数据，将所述第一梯度数据划分为至少两个梯度片段
603、使用所述第一梯度数据的第一码本对所述至少两个梯度片段中的每个梯度片段进行量化转换，以得到所述每个梯度片段各自对应的量化表示
604、B个梯度数据对应的B个量化集合
605、针对所述B个量化集合中的每个第一量化集合，使用所述第一量化集合对应的第一码本对所述第一量化集合中的每个量化表示进行反量化，以得到所述每个量化表示各自所对应的第一梯度片段
606、将所述第一量化集合中每个量化表示所对应的第一梯度片段进行拼接，以得到与所述第一量化集合对应的第一梯度数据
607、针对B个梯度数据对应的B个权重中的第一权重，对所述第一权重的N个第一梯度数据进行梯度聚合，以得到所述第一权重的梯度聚合数据
608、B个权重各自的梯度聚合数据
609、根据梯度聚合数据更新权重

图7

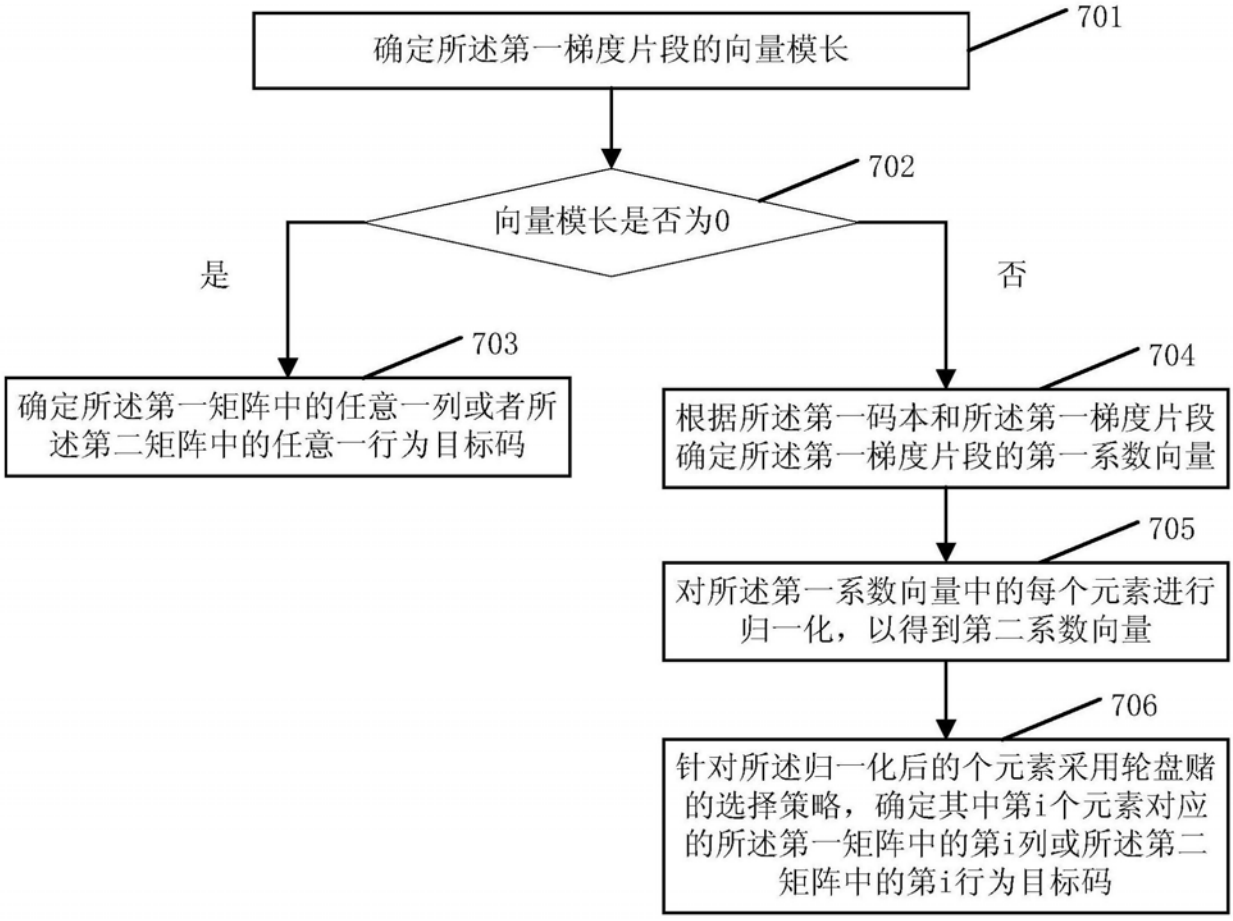

701
确定所述第一梯度片段的向量模长
702
向量模长是否为0
是
否
703
确定所述第一矩阵中的任意一列或者所述第二矩阵中的任意一行为目标码
704
根据所述第一码本和所述第一梯度片段确定所述第一梯度片段的第一系数向量
705
对所述第一系数向量中的每个元素进行归一化，以得到第二系数向量
706
针对所述归一化后的个元素采用轮盘赌的选择策略，确定其中第i个元素对应的所述第一矩阵中的第i列或所述第二矩阵中的第i行为目标码

图8

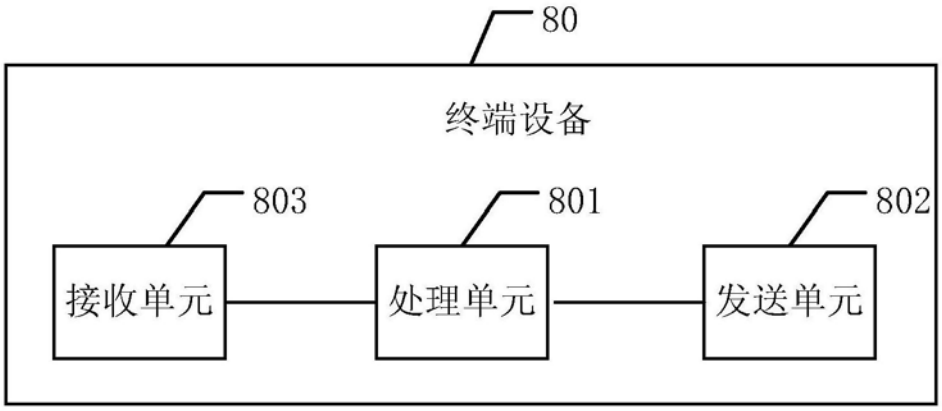

80
终端设备
803
801
802
接收单元
处理单元
发送单元

图9

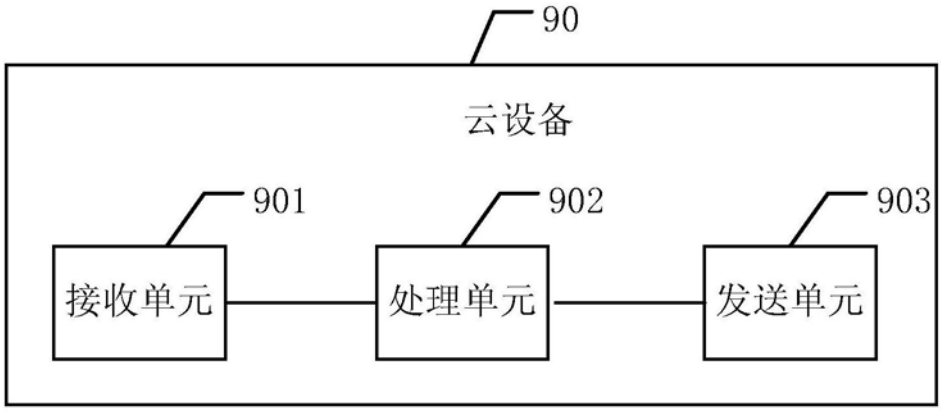

90
云设备
901
902
903
接收单元
处理单元
发送单元

图10

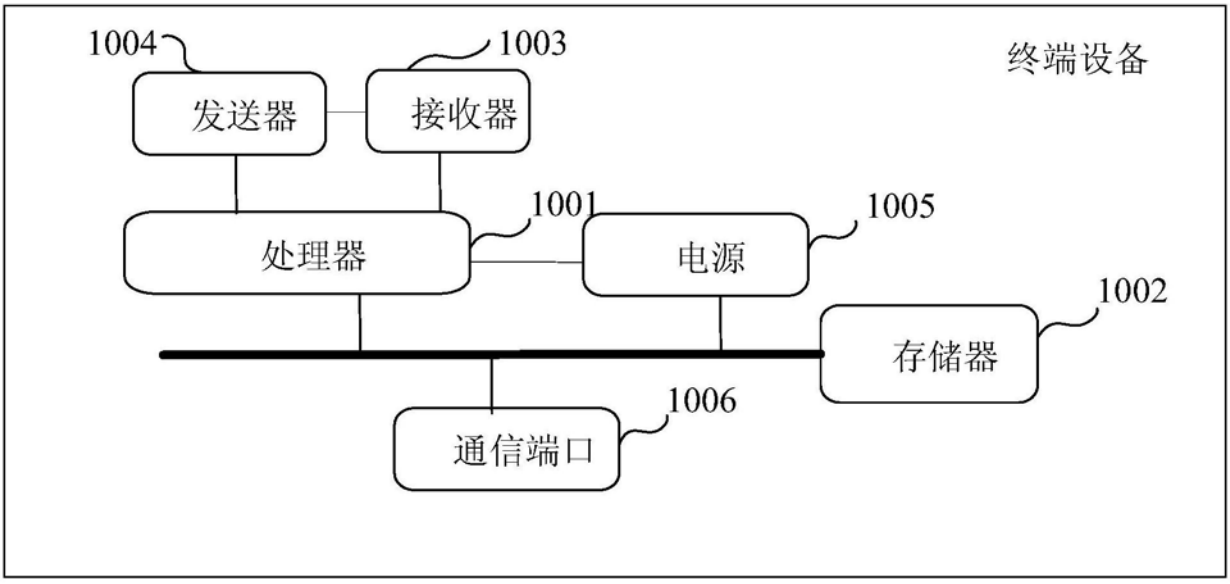

1004
1003
终端设备
发送器
接收器
1001
1005
处理器
电源
1002
存储器
1006
通信端口

图11

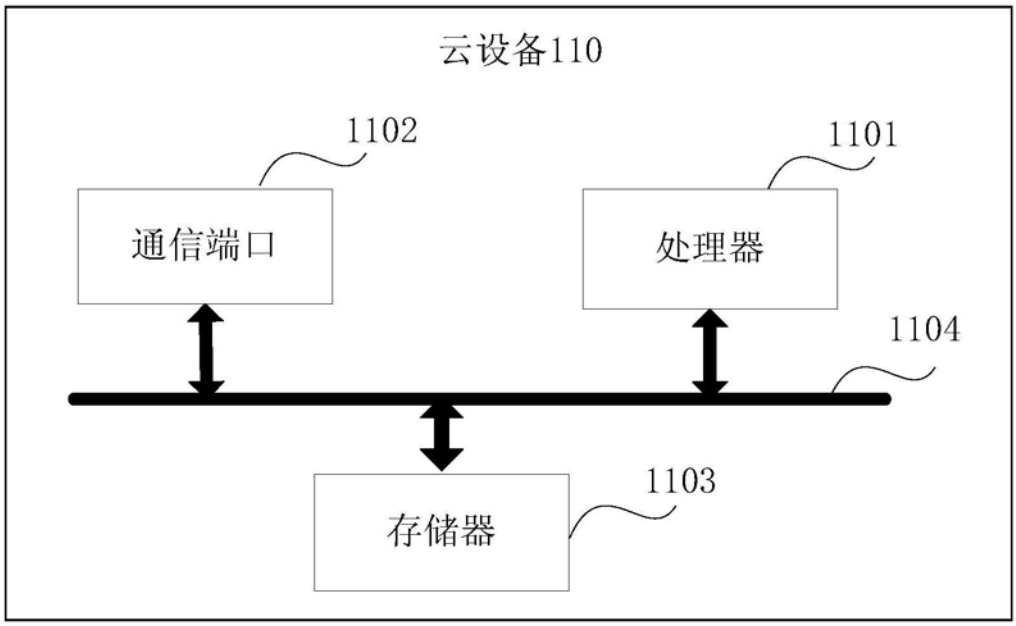


图12