



(12) 发明专利

(10) 授权公告号 CN 112948105 B

(45) 授权公告日 2023. 10. 17

(21) 申请号 201911268718.8

G06F 9/54 (2006.01)

(22) 申请日 2019.12.11

(56) 对比文件

(65) 同一申请的已公布的文献号

CN 110134636 A, 2019.08.16

申请公布号 CN 112948105 A

CN 110245743 A, 2019.09.17

(43) 申请公布日 2021.06.11

US 2016125316 A1, 2016.05.05

(73) 专利权人 香港理工大学深圳研究院

US 2019087723 A1, 2019.03.21

地址 518057 广东省深圳市南山区高新园

CN 109740755 A, 2019.05.10

南区粤兴一道18号香港理工大学产学

CN 110287031 A, 2019.09.27

研大楼205室

CN 109951438 A, 2019.06.28

(72) 发明人 郭嵩 王号召 詹玉峰

Haozhao Wang; Song Guo; Bin Tang;

(74) 专利代理机构 深圳中一专利商标事务所

Ruixuan Li; Chengjie Li. Heterogeneity-aware Gradient Coding for Straggler Tolerance. 2019, 全文.

44237

专利代理师 曹小翠

审查员 李一

(51) Int. Cl.

G06F 9/50 (2006.01)

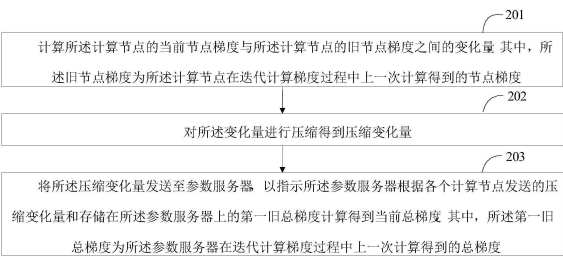
权利要求书2页 说明书13页 附图4页

(54) 发明名称

一种梯度传输方法、梯度传输装置及参数服务器

(57) 摘要

本申请适用于数据处理技术领域,提供了一种梯度传输方法、梯度传输装置及参数服务器,所述方法包括:计算所述计算节点的当前节点梯度与所述计算节点的旧节点梯度之间的变化量,其中,所述旧节点梯度为所述计算节点在迭代计算梯度过程中上一次计算得到的节点梯度;对所述变化量进行压缩得到压缩变化量;将所述压缩变化量发送至参数服务器,以指示所述参数服务器根据各个计算节点发送的压缩变化量和存储在所述参数服务器上的第一旧总梯度计算得到当前总梯度,其中,所述第一旧总梯度为所述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度。通过上述方法,可以在保证梯度精度损失较少的同时减小计算节点与参数服务器之间的通信量。



1. 一种梯度传输方法,其特征在于,应用于计算节点,包括:

计算所述计算节点的当前节点梯度与所述计算节点的旧节点梯度之间的变化量,其中,所述旧节点梯度为所述计算节点在迭代计算梯度过程中上一次计算得到的节点梯度;

对所述变化量进行压缩得到压缩变化量;

将所述压缩变化量发送至参数服务器,以指示所述参数服务器根据各个计算节点发送的压缩变化量和存储在所述参数服务器上的第一旧总梯度计算得到当前总梯度,所述参数服务器接收各个计算节点发送的压缩变化量,根据各个压缩变化量计算得到总梯度变化量并根据所述总梯度变化量和第一旧总梯度计算得到当前总梯度,其中,所述第一旧总梯度为所述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度;

所述计算所述计算节点的当前节点梯度与所述计算节点的旧节点梯度之间的变化量,包括:

分别计算当前节点梯度中各个维度的梯度分量与旧节点梯度中对应的各个维度的梯度分量之间的变化量;

相应地,所述对所述变化量进行压缩得到压缩变化量,包括:

根据存储在所述计算节点上的第二旧总梯度计算压缩阈值,其中,所述第二旧总梯度为所述计算节点在迭代计算梯度过程中上一次计算得到的总梯度;

将各个维度的梯度分量的变化量分别与所述压缩阈值进行比较;

将目标维度的梯度分量的变化量保持不变,将非目标维度的梯度分量的变化量设置为0,得到压缩变化量,其中,所述目标维度为梯度分量的变化量大于所述压缩阈值的维度。

2. 根据权利要求1所述的梯度传输方法,其特征在于,所述根据存储在所述计算节点上的第二旧总梯度计算压缩阈值,包括:

根据预设的阈值计算公式计算所述压缩阈值,所述阈值计算公式为 $H = \lambda \frac{\|g_0\|^2}{\sqrt{D}}$, 其中H为压缩阈值, λ 为预设的超参数, g_0 为第二旧总梯度,D为第二旧总梯度的维度数。

3. 根据权利要求1所述的梯度传输方法,其特征在于,所述对所述变化量进行压缩得到压缩变化量之后,还包括:

根据所述压缩变化量和所述旧节点梯度更新所述当前节点梯度,其中,更新后的当前节点梯度作为所述计算节点在迭代计算梯度过程中下一次计算得到的节点梯度对应的旧节点梯度。

4. 根据权利要求1所述的梯度传输方法,其特征在于,所述将所述压缩变化量发送至参数服务器之后,还包括:

接收所述参数服务器发送的总梯度变化量;

根据所述总梯度变化量更新所述计算节点的模型参数。

5. 根据权利要求4所述的梯度传输方法,其特征在于,所述根据所述总梯度变化量更新所述计算节点的模型参数,包括:

根据所述总梯度变化量和存储在所述计算节点上的第二旧总梯度计算得到所述当前总梯度;

根据所述当前总梯度更新所述计算节点的模型参数。

6. 一种梯度传输装置, 其特征在于, 应用于计算节点, 包括:

变化量计算单元, 用于计算所述计算节点的当前节点梯度与所述计算节点的旧节点梯度之间的变化量, 其中, 所述旧节点梯度为所述计算节点在迭代计算梯度过程中上一次计算得到的节点梯度;

变化量压缩单元, 用于对所述变化量进行压缩得到压缩变化量;

变化量发送单元, 用于将所述压缩变化量发送至参数服务器, 以指示所述参数服务器根据各个计算节点发送的压缩变化量和存储在所述参数服务器上的第一旧总梯度计算得到当前总梯度, 所述参数服务器接收各个计算节点发送的压缩变化量, 根据各个压缩变化量计算得到总梯度变化量并根据所述总梯度变化量和第一旧总梯度计算得到当前总梯度, 其中, 所述第一旧总梯度为所述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度;

所述变化量计算单元进一步地用于: 分别计算当前节点梯度中各个维度的梯度分量与旧节点梯度中对应的各个维度的梯度分量之间的变化量;

所述变化量压缩单元进一步地用于:

根据存储在所述计算节点上的第二旧总梯度计算压缩阈值, 其中, 所述第二旧总梯度为所述计算节点在迭代计算梯度过程中上一次计算得到的总梯度;

将各个维度的梯度分量的变化量分别与所述压缩阈值进行比较;

将目标维度的梯度分量的变化量保持不变, 将非目标维度的梯度分量的变化量设置为0, 得到压缩变化量, 其中, 所述目标维度为梯度分量的变化量大于所述压缩阈值的维度。

7. 一种终端设备, 包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序, 其特征在于, 所述处理器执行所述计算机程序时实现如权利要求1至5任一项所述的方法。

一种梯度传输方法、梯度传输装置及参数服务器

技术领域

[0001] 本申请属于数据处理技术领域,尤其涉及一种梯度传输方法、梯度传输装置、参数服务器及终端设备。

背景技术

[0002] 在分布式机器学习领域中,计算节点往往数量比较多,参数服务器与各个计算节点之间所要传输的参数的数据量也较大,这将导致参数服务器与各个计算节点之间的通信出现瓶颈。

[0003] 为了解决上述问题,现有的技术是计算节点在计算出梯度后,将梯度稀疏化,然后传输给参数服务器。但是,稀疏化的梯度会丢失较多数据,造成了梯度的精度损失。

发明内容

[0004] 有鉴于此,本申请提供了一种梯度传输方法、梯度传输装置、参数服务器及终端设备,可以在保证梯度精度损失较少的同时减小计算节点与参数服务器之间的通信量。

[0005] 第一方面,本申请提供了一种梯度传输方法,应用于计算节点,包括:

[0006] 计算上述计算节点的当前节点梯度与上述计算节点的旧节点梯度之间的变化量,其中,上述旧节点梯度为上述计算节点在迭代计算梯度过程中上一次计算得到的节点梯度;

[0007] 对上述变化量进行压缩得到压缩变化量;

[0008] 将上述压缩变化量发送至参数服务器,以指示上述参数服务器根据各个计算节点发送的压缩变化量和存储在上述参数服务器上的第一旧总梯度计算得到当前总梯度,其中,上述第一旧总梯度为上述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度。

[0009] 第二方面,本申请提供了一种梯度传输方法,应用于参数服务器,包括:

[0010] 接收各个计算节点发送的压缩变化量,其中,压缩变化量由计算节点的当前节点梯度与旧节点梯度之间的变化量经过压缩得到,旧节点梯度为计算节点在迭代计算梯度过程中上一次计算得到的节点梯度;

[0011] 根据各个压缩变化量计算得到总梯度变化量;

[0012] 根据上述总梯度变化量和第一旧总梯度计算得到当前总梯度,其中,上述第一旧总梯度为上述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度。

[0013] 第三方面,本申请提供了一种梯度传输装置,应用于计算节点,包括:

[0014] 变化量计算单元,用于计算上述计算节点的当前节点梯度与上述计算节点的旧节点梯度之间的变化量,其中,上述旧节点梯度为上述计算节点在迭代计算梯度过程中上一次计算得到的节点梯度;

[0015] 变化量压缩单元,用于对上述变化量进行压缩得到压缩变化量;

[0016] 变化量发送单元,用于将上述压缩变化量发送至参数服务器,以指示上述参数服

务器根据各个计算节点发送的压缩变化量和存储在上述参数服务器上的第一旧总梯度计算得到当前总梯度,其中,上述第一旧总梯度为上述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度。

[0017] 第四方面,本申请提供了一种参数服务器,包括:

[0018] 变化量接收单元,用于接收各个计算节点发送的压缩变化量,其中,压缩变化量由计算节点的当前节点梯度与旧节点梯度之间的变化量经过压缩得到,旧节点梯度为计算节点在迭代计算梯度过程中上一次计算得到的节点梯度;

[0019] 总变化量计算单元,用于根据各个压缩变化量计算得到总梯度变化量;

[0020] 总梯度计算单元,用于根据上述总梯度变化量和第一旧总梯度计算得到当前总梯度,其中,上述第一旧总梯度为上述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度。

[0021] 第五方面,本申请提供了一种终端设备,包括第一存储器、第一处理器以及存储在上述第一存储器中并可在上述第一处理器上运行的第一计算机程序,上述第一处理器执行上述第一计算机程序时实现如上述第一方面所提供的方法。

[0022] 第六方面,本申请提供了一种参数服务器,包括第二存储器、第二处理器以及存储在上述第二存储器中并可在上述第二处理器上运行的第二计算机程序,上述第二处理器执行上述第二计算机程序时实现如上述第二方面所提供的方法。

[0023] 第七方面,本申请提供了一种计算机可读存储介质,上述计算机可读存储介质存储有第一计算机程序,上述第一计算机程序被处理器执行时实现如第一方面所提供的方法。

[0024] 第八方面,本申请提供了一种计算机可读存储介质,上述计算机可读存储介质存储有第二计算机程序,上述第二计算机程序被处理器执行时实现如第二方面所提供的方法。

[0025] 第九方面,本申请提供了一种计算机程序产品,当计算机程序产品在终端设备上运行时,使得终端设备执行上述第一方面中所提供的方法。

[0026] 第十方面,本申请提供了一种计算机程序产品,当计算机程序产品在终端设备上运行时,使得终端设备执行上述第二方面中所提供的方法。

[0027] 由上可见,本申请方案中首先计算上述计算节点的当前节点梯度与上述计算节点的旧节点梯度之间的变化量,其中,上述旧节点梯度为上述计算节点在迭代计算梯度过程中上一次计算得到的节点梯度;然后对上述变化量进行压缩得到压缩变化量;最后将上述压缩变化量发送至参数服务器,以指示上述参数服务器根据各个计算节点发送的压缩变化量和存储在上述参数服务器上的第一旧总梯度计算得到当前总梯度,其中,上述第一旧总梯度为上述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度。通过本申请方案,在计算节点与参数服务器之间传输梯度变化量,可以在保证梯度精度损失较少的同时减小计算节点与参数服务器之间的通信量。

附图说明

[0028] 为了更清楚地说明本申请实施例中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本申请的一些

实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0029] 图1是本申请实施例提供的参数服务器与计算节点的连接关系图;

[0030] 图2是本申请实施例提供的一种梯度传输方法的流程示意图;

[0031] 图3是本申请实施例提供的另一种梯度传输方法的流程示意图;

[0032] 图4是本申请实施例提供的参数服务器与计算节点的交互流程示意图;

[0033] 图5是本申请实施例提供的梯度传输装置的结构示意图;

[0034] 图6是本申请实施例提供的参数服务器的结构示意图;

[0035] 图7是本申请实施例提供的终端设备的结构示意图;

[0036] 图8为本申请实施例提供的参数服务器的结构示意图。

具体实施方式

[0037] 以下描述中,为了说明而不是为了限定,提出了诸如特定系统结构、技术之类的具体细节,以便透彻理解本申请实施例。然而,本领域的技术人员应当清楚,在没有这些具体细节的其它实施例中也可以实现本申请。在其它情况中,省略对众所周知的系统、装置、电路以及方法的详细说明,以免不必要的细节妨碍本申请的描述。

[0038] 应当理解,当在本申请说明书和所附权利要求书中使用时,术语“包括”指示所描述特征、整体、步骤、操作、元素和/或组件的存在,但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。

[0039] 还应当理解,在本申请说明书和所附权利要求书中使用的术语“和/或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合,并且包括这些组合。

[0040] 如在本申请说明书和所附权利要求书中所使用的那样,术语“如果”可以依据上下文被解释为“当...时”或“一旦”或“响应于确定”或“响应于检测到”。类似地,短语“如果确定”或“如果检测到[所描述条件或事件]”可以依据上下文被解释为意指“一旦确定”或“响应于确定”或“一旦检测到[所描述条件或事件]”或“响应于检测到[所描述条件或事件]”。

[0041] 另外,在本申请说明书和所附权利要求书的描述中,术语“第一”、“第二”、“第三”等仅用于区分描述,而不能理解为指示或暗示相对重要性。

[0042] 在本申请说明书中描述的参考“一个实施例”或“一些实施例”等意味着在本申请的一个或多个实施例中包括结合该实施例描述的特定特征、结构或特点。由此,在本说明书中的不同之处出现的语句“在一个实施例中”、“在一些实施例中”、“在其他一些实施例中”、“在另外一些实施例中”等不是必然都参考相同的实施例,而是意味着“一个或多个但不是所有的实施例”,除非是以其他方式另外特别强调。术语“包括”、“包含”、“具有”及它们的变形都意味着“包括但不限于”,除非是以其他方式另外特别强调。

[0043] 如图1所示,本申请所提供的梯度传输方法分别应用于计算节点及参数服务器,其中,上述计算节点有多个,每个计算节点都可以执行本申请实施例提供的应用于计算节点的梯度传输方法的步骤,且每个计算节点与上述参数服务器基于互联网相互连接,以使二者之间能够进行数据交互。每个计算节点存储有用于训练的部分数据,各个计算节点根据各自用于训练的部分数据计算各自的节点梯度,上述参数服务器根据接收到的各个计算节点发送的数据计算一个总梯度。

[0044] 图2示出了本申请实施例提供的一种梯度传输方法的流程图,该梯度传输方法应用于计算节点,详述如下:

[0045] 步骤201,计算上述计算节点的当前节点梯度与上述计算节点的旧节点梯度之间的变化量;

[0046] 在本申请实施例中,上述当前节点梯度为上述计算节点在迭代计算梯度过程中本次计算得到的节点梯度,而上述旧节点梯度为上述计算节点在迭代计算梯度过程中上一次计算得到的节点梯度。在机器学习中,梯度用于更新机器学习模型的模型参数,梯度的计算为一个迭代的过程,即通过不断地计算新的梯度来求得机器学习模型的模型参数的最优解。需要说明的是,在整个机器学习迭代计算梯度的过程中,上述计算节点每一次计算得到的梯度都具有相同数量的维度,并且每个梯度都包括多个维度,每个维度对应梯度的一个梯度分量。用上述当前节点梯度减去上述旧节点梯度,得到的差值即为上述变化量。

[0047] 可选地,上述步骤201具体包括:

[0048] 分别计算当前节点梯度中各个维度的梯度分量与旧节点梯度中对应的各个维度的梯度分量之间的变化量。

[0049] 具体地,上述当前节点梯度和上述旧节点梯度具有相同数量的维度,通过上述当前节点梯度中各个维度的梯度分量减去上述旧节点梯度中对应的各个维度的梯度分量,得到的差值即为上述变化量。因此,上述变化量是一个向量,且该向量的维度数与上述当前节点梯度以及上述旧节点梯度的维度数相同,该向量中的每一维元素都指示了上述当前节点梯度与上述旧节点梯度之间相应维度的梯度分量的变化量。例如,我们有当前节点梯度 $M = (6, 7, 8, 9)$ 和旧节点梯度 $N = (1, 2, 3, 4)$,通过当前节点梯度 M 的梯度分量“6”、“7”、“8”及“9”分别减去旧节点梯度 N 的梯度分量“1”、“2”、“3”及“4”,得到变化量 $Q = (5, 5, 5, 5)$ 。

[0050] 步骤202,对上述变化量进行压缩得到压缩变化量;

[0051] 在本申请实施例中,需要对上述变化量的数据量进行一个压缩,在保证上述变化量的精度损失尽量少的前提下,使上述变化量的数据量最小。因此,上述压缩变化量在计算节点与参数服务器之间进行传输时的通信开销较小。

[0052] 可选地,上述步骤202具体包括:

[0053] A1、根据存储在上述计算节点上的第二旧总梯度计算压缩阈值;

[0054] A2、将各个维度的梯度分量的变化量分别与上述压缩阈值进行比较;

[0055] A3、将目标维度的梯度分量的变化量保持不变,将非目标维度的梯度分量的变化量设置为0,得到压缩变化量。

[0056] 其中,上述第二旧总梯度为上述计算节点在迭代计算梯度过程中上一次计算得到的总梯度,上述目标维度为梯度分量的变化量大于上述压缩阈值的维度。总梯度即为综合各个计算节点的节点梯度计算得到的梯度,因此,在本申请实施例中,使用总梯度来更新机器学习模型的模型参数。通过上述第二旧总梯度计算得到的上述压缩阈值需要满足一个条件,就是在迭代计算梯度的过程中,保证每一次得到的损失函数的损失值相对上一次都在下降。

[0057] 得到上述压缩阈值后,将上述变化量中各个维度的梯度分量的变化量分别和上述压缩阈值进行比较,将上述变化量中目标维度的梯度分量的变化量保持不变,非目标维度的梯度分量的变化量设置为0,得到上述压缩变化量。例如,有一个变化量 $J = (3, 5, 8, 9)$,将

其中的第一维度的梯度分量的变化量“3”、第二维度的梯度分量的变化量“5”、第三维度的梯度分量的变化量“8”以及第四维度的梯度分量的变化量“9”分别和压缩阈值“6”进行比较,则将上述第一维度和第二维度确定为非目标维度,上述第三维度和第四维度为目标维度,将变化量J压缩后,得到的压缩变化量 $K=(0,0,8,9)$ 。

[0058] 可选地,上述步骤A1具体包括:

[0059] 根据预设的阈值计算公式计算上述压缩阈值,上述阈值计算公式为

$$[0060] \quad H = \lambda \frac{\|g_0\|^2}{\sqrt{D}}。$$

[0061] 其中H为压缩阈值, λ 为机器学习模型中预设的超参数, g_0 为存储在计算节点上的第二旧总梯度,D为第二旧总梯度的维度数。

[0062] 可选地,上述步骤202之后还包括:

[0063] 根据上述压缩变化量和上述旧节点梯度更新上述当前节点梯度。

[0064] 其中,更新后的当前节点梯度作为上述计算节点在迭代计算梯度过程中下一次计算得到的节点梯度对应的旧节点梯度。由于上述旧节点梯度加上压缩前的变化量等于更新前的当前节点梯度,而压缩后的变化量往往与压缩前的变化量存在差异,因而当上述旧节点梯度加上压缩后的变化量(压缩变化量)时,会得到一个不同于更新前的当前节点梯度的新节点梯度,该新节点梯度即可被看作为更新后的当前节点梯度。然后将该更新后的当前节点梯度保存在上述计算节点中,当上述计算节点进行下一次节点梯度的计算时,会将前次计算过程中所保留的更新后的当前节点梯度取出作为本次计算过程中的旧节点梯度。例如,某一计算节点在第i次节点梯度的计算时,旧节点梯度为(2,3,6),当前节点梯度为(4,7,13),压缩前的变化量为(2,4,7),压缩后的变化量为(0,4,7),则将旧节点梯度(2,3,6)和压缩后的变化量(0,4,7)相加,得到更新后的当前节点梯度(2,7,13),将该更新后的当前节点梯度(2,7,13)保存在上述计算节点中;当该计算节点进行第i+1次节点梯度的计算时,会将第i次计算过程中所保存的更新后的当前节点梯度(2,7,13)取出作为本次(也即第i+1次)节点梯度计算过程中的旧节点梯度。

[0065] 步骤203,将上述压缩变化量发送至参数服务器,以指示上述参数服务器根据各个计算节点发送的压缩变化量和存储在上述参数服务器上的第一旧总梯度计算得到当前总梯度。

[0066] 在本申请实施例中,上述第一旧总梯度为上述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度。需要说明的是,上述第一旧总梯度和上述第二旧总梯度相同,只是上述第一旧总梯度存储在上述参数服务器上,而上述第二旧总梯度存储在上述计算节点上。在每一次计算节点将压缩变化量发送给参数服务器后,参数服务器会根据接收到的各个计算节点发送的压缩变化量计算得到一个总梯度变化量,并且参数服务器将该总梯度变化量发送给各个计算节点,参数服务器和各个计算节点均根据该总梯度变化量和各自存储的旧总梯度(第一旧总梯度或第二旧总梯度)计算出各自的当前总梯度,每个当前总梯度都相同。参数服务器和各个计算节点均将各自计算得到的当前总梯度保存,并将该当前总梯度作为迭代计算梯度过程中下一次的旧总梯度。

[0067] 可选地,上述步骤203之后还包括:

[0068] B1、接收上述参数服务器发送的总梯度变化量；

[0069] B2、根据上述总梯度变化量更新上述计算节点的模型参数。

[0070] 具体地，上述计算节点接收到上述参数服务器发送的总梯度变化量，根据上述总梯度变化量更新上述计算节点上存储的机器学习模型的模型参数。需要说明的是，上述计算节点和上述参数服务器上存储有相同的机器学习模型，并且在迭代计算梯度的每一轮迭代中，上述计算节点上的机器学习模型的模型参数和上述参数服务器上的机器学习模型的模型参数会分别进行更新，以保证在下一轮迭代之前，上述计算节点和参数服务器上的模型参数是同步的。

[0071] 可选地，上述步骤B2具体包括：

[0072] C1、根据上述总梯度变化量和存储在上述计算节点上的第二旧总梯度计算得到上述当前总梯度；

[0073] C2、根据上述当前总梯度更新上述计算节点的模型参数。

[0074] 具体地，将上述总梯度变化量和上述第二旧总梯度相加，即可得到上述当前总梯度。然后使用上述当前总梯度完成对上述计算节点上的机器学习模型的模型参数更新。

[0075] 由上可见，本申请方案中将上述计算节点的当前节点梯度和旧节点梯度之间的变化量进行压缩得到压缩变化量，然后将所述压缩变化量发送给参数服务器，以指示上述参数服务器根据各个计算节点发送的压缩变化量和存储在上述参数服务器上的第一旧总梯度计算得到当前总梯度。通过本申请方案，在计算节点与参数服务器之间传输梯度变化量，可以在保证梯度精度损失较少的同时减小计算节点与参数服务器之间的通信量。

[0076] 应理解，上述实施例中各步骤的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不对本申请实施例的实施过程构成任何限定。

[0077] 图3示出了本申请实施例提供的一种梯度传输方法的流程图，该梯度传输方法应用于参数服务器，详述如下：

[0078] 步骤301，接收各个计算节点发送的压缩变化量；

[0079] 在本申请实施例中，上述参数服务器同时与多个计算节点相互通信，上述参数服务器接收与之通信连接的各个计算节点发送的压缩变化量。其中，压缩变化量由计算节点的当前节点梯度与旧节点梯度之间的变化量经过压缩得到，旧节点梯度为计算节点在迭代计算梯度过程中上一次计算得到的节点梯度。

[0080] 步骤302，根据各个压缩变化量计算得到总梯度变化量；

[0081] 在本申请实施例中，上述参数服务器计算各个压缩变化量的平均值，将该平均值作为上述总梯度变化量。

[0082] 步骤303，根据上述总梯度变化量和第一旧总梯度计算得到当前总梯度。

[0083] 在本申请实施例中，上述参数服务器将上述总梯度变化量和上述第一旧总梯度相加，即可得到上述当前总梯度。其中，上述第一旧总梯度为上述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度。

[0084] 可选地，上述步骤303具体包括：

[0085] 分别将上述总梯度变化量中各个维度的梯度分量的变化量与第一旧总梯度中对应的各个维度的梯度分量相加，得到上述当前总梯度。

[0086] 可选地,上述步骤303之后还包括:

[0087] 根据上述当前总梯度更新上述参数服务器的模型参数;

[0088] 将上述总梯度变化量发送给各个计算节点。

[0089] 具体地,上述参数服务器根据上述当前总梯度更新存储在上述参数服务器上的机器学习模型的模型参数。然后将上述总梯度变化量发送给与上述参数服务器通信连接的各个计算节点。

[0090] 由上可见,本申请方案中将上述计算节点的当前节点梯度和旧节点梯度之间的变化量进行压缩得到压缩变化量,然后将所述压缩变化量发送给参数服务器,以指示上述参数服务器根据各个计算节点发送的压缩变化量和存储在上述参数服务器上的第一旧总梯度计算得到当前总梯度。通过本申请方案,在计算节点与参数服务器之间传输梯度变化量,可以在保证梯度精度损失较少的同时减小计算节点与参数服务器之间的通信量。

[0091] 图4示出了本申请实施例提供的参数服务器与计算节点的交互流程示意图,详述如下:

[0092] 在步骤401中,上述计算节点计算上述计算节点的当前节点梯度与上述计算节点的旧节点梯度之间的变化量;

[0093] 在步骤402中,上述计算节点对上述变化量进行压缩得到压缩变化量;

[0094] 在步骤403中,上述计算节点发送上述压缩变化量至上述参数服务器;

[0095] 在步骤404中,上述参数服务器接收各个计算节点发送的压缩变化量;

[0096] 在步骤405中,上述参数服务器根据各个压缩变化量计算得到总梯度变化量;

[0097] 在步骤406中,上述参数服务器根据上述总梯度变化量和第一旧总梯度计算得到当前总梯度;

[0098] 在步骤407中,上述参数服务器发送上述总梯度变化量至各个计算节点;

[0099] 在步骤408中,上述计算节点根据上述总梯度变化量更新上述计算节点的模型参数。

[0100] 由上可见,本申请方案中将上述计算节点的当前节点梯度和旧节点梯度之间的变化量进行压缩得到压缩变化量,然后将所述压缩变化量发送给参数服务器,以指示上述参数服务器根据各个计算节点发送的压缩变化量和存储在上述参数服务器上的第一旧总梯度计算得到当前总梯度。通过本申请方案,在计算节点与参数服务器之间传输梯度变化量,可以在保证梯度精度损失较少的同时减小计算节点与参数服务器之间的通信量。

[0101] 图5示出了本申请实施例提供的一种梯度传输装置的结构示意图,为了便于说明,仅示出了与本申请实施例相关的部分。

[0102] 该梯度传输装置500包括:

[0103] 变化量计算单元501,用于计算上述计算节点的当前节点梯度与上述计算节点的旧节点梯度之间的变化量,其中,上述旧节点梯度为上述计算节点在迭代计算梯度过程中上一次计算得到的节点梯度;

[0104] 变化量压缩单元502,用于对上述变化量进行压缩得到压缩变化量;

[0105] 变化量发送单元503,用于将上述压缩变化量发送至参数服务器,以指示上述参数服务器根据各个计算节点发送的压缩变化量和存储在上述参数服务器上的第一旧总梯度计算得到当前总梯度,其中,上述第一旧总梯度为上述参数服务器在迭代计算梯度过程中

上一次计算得到的总梯度。

[0106] 可选地,上述变化量计算单元501还包括:

[0107] 分量变化量计算子单元,用于分别计算当前节点梯度中各个维度的梯度分量与旧节点梯度中对应的各个维度的梯度分量之间的变化量。

[0108] 可选地,上述变化量压缩单元502还包括:

[0109] 压缩阈值计算子单元,用于根据存储在上述计算节点上的第二旧总梯度计算压缩阈值,其中,上述第二旧总梯度为上述计算节点在迭代计算梯度过程中上一次计算得到的总梯度;

[0110] 阈值比较子单元,用于将各个维度的梯度分量的变化量分别与上述压缩阈值进行比较;

[0111] 目标维度压缩子单元,用于将目标维度的梯度分量的变化量保持不变,将非目标维度的梯度分量的变化量设置为0,得到压缩变化量,其中,上述目标维度为梯度分量的变化量大于上述压缩阈值的维度。

[0112] 可选地,上述压缩阈值计算子单元还包括:

[0113] 公式计算子单元,用于根据预设的阈值计算公式计算上述压缩阈值,上述阈值计

算公式为 $H = \lambda \frac{\|g_0\|^2}{\sqrt{D}}$, 其中H为压缩阈值, λ 为预设的超参数, g_0 为第二旧总梯度,D为第二旧

总梯度的维度数。

[0114] 可选地,上述梯度传输装置500还包括:

[0115] 更新梯度单元,用于根据上述压缩变化量和上述旧节点梯度更新上述当前节点梯度,其中,更新后的当前节点梯度作为上述计算节点在迭代计算梯度过程中下一次计算得到的节点梯度对应的旧节点梯度。

[0116] 可选地,上述梯度传输装置500还包括:

[0117] 总变化量接收单元,用于接收上述参数服务器发送的总梯度变化量;

[0118] 节点模型更新单元,用于根据上述总梯度变化量更新上述计算节点的模型参数。

[0119] 可选地,上述节点模型更新单元还包括:

[0120] 总梯度计算子单元,用于根据上述总梯度变化量和存储在上述计算节点上的第二旧总梯度计算得到上述当前总梯度;

[0121] 节点模型参数更新子单元,根据上述当前总梯度更新上述计算节点的模型参数。

[0122] 由上可见,本申请方案中将上述计算节点的当前节点梯度和旧节点梯度之间的变化量进行压缩得到压缩变化量,然后将所述压缩变化量发送给参数服务器,以指示上述参数服务器根据各个计算节点发送的压缩变化量和存储在上述参数服务器上的第一旧总梯度计算得到当前总梯度。通过本申请方案,在计算节点与参数服务器之间传输梯度变化量,可以在保证梯度精度损失较少的同时减小计算节点与参数服务器之间的通信量。

[0123] 图6示出了本申请实施例提供的一种参数服务器的结构示意图,为了便于说明,仅示出了与本申请实施例相关的部分。

[0124] 该参数服务器600包括:

[0125] 变化量接收单元601,用于接收各个计算节点发送的压缩变化量,其中,压缩变化

量由计算节点的当前节点梯度与旧节点梯度之间的变化量经过压缩得到,旧节点梯度为计算节点在迭代计算梯度过程中上一次计算得到的节点梯度;

[0126] 总变化量计算单元602,用于根据各个压缩变化量计算得到总梯度变化量;

[0127] 总梯度计算单元603,用于根据上述总梯度变化量和第一旧总梯度计算得到当前总梯度,其中,上述第一旧总梯度为上述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度。

[0128] 可选地,上述总梯度计算单元603还包括:

[0129] 分量相加子单元,用于分别将上述总梯度变化量中各个维度的梯度分量的变化量与第一旧总梯度中对应的各个维度的梯度分量相加,得到上述当前总梯度。

[0130] 可选地,上述参数服务器600还包括:

[0131] 服务器模型更新单元,用于根据上述当前总梯度更新上述参数服务器的模型参数;

[0132] 总变化量发送单元,用于将上述总梯度变化量发送给各个计算节点。

[0133] 由上可见,本申请方案中将上述计算节点的当前节点梯度和旧节点梯度之间的变化量进行压缩得到压缩变化量,然后将所述压缩变化量发送给参数服务器,以指示上述参数服务器根据各个计算节点发送的压缩变化量和存储在上述参数服务器上的第一旧总梯度计算得到当前总梯度。通过本申请方案,在计算节点与参数服务器之间传输梯度变化量,可以在保证梯度精度损失较少的同时减小计算节点与参数服务器之间的通信量。

[0134] 图7为本申请一实施例提供的终端设备的结构示意图。如图7所示,该实施例的终端设备7包括:至少一个第一处理器70(图7中仅示出一个)、第一存储器71以及存储在上述第一存储器71中并可在上述至少一个第一处理器70上运行的第一计算机程序72,上述第一处理器70执行上述第一计算机程序72时实现以下步骤:

[0135] 计算上述计算节点的当前节点梯度与上述计算节点的旧节点梯度之间的变化量,其中,上述旧节点梯度为上述计算节点在迭代计算梯度过程中上一次计算得到的节点梯度;

[0136] 对上述变化量进行压缩得到压缩变化量;

[0137] 将上述压缩变化量发送至参数服务器,以指示上述参数服务器根据各个计算节点发送的压缩变化量和存储在上述参数服务器上的第一旧总梯度计算得到当前总梯度,其中,上述第一旧总梯度为上述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度。

[0138] 假设上述为第一种可能的实施方式,则在第一种可能的实施方式作为基础而提供的第二种可能的实施方式中,上述计算上述计算节点的当前节点梯度与上述计算节点的旧节点梯度之间的变化量,包括:

[0139] 分别计算当前节点梯度中各个维度的梯度分量与旧节点梯度中对应的各个维度的梯度分量之间的变化量;

[0140] 相应地,上述对上述变化量进行压缩得到压缩变化量,包括:

[0141] 根据存储在上述计算节点上的第二旧总梯度计算压缩阈值,其中,上述第二旧总梯度为上述计算节点在迭代计算梯度过程中上一次计算得到的总梯度;

[0142] 将各个维度的梯度分量的变化量分别与上述压缩阈值进行比较;

[0143] 将目标维度的梯度分量的变化量保持不变,将非目标维度的梯度分量的变化量设置为0,得到压缩变化量,其中,上述目标维度为梯度分量的变化量大于上述压缩阈值的维度。

[0144] 在上述第二种可能的实施方式作为基础而提供的第三种可能的实施方式中,上述根据存储在上述计算节点上的第二旧总梯度计算压缩阈值,包括:

[0145] 根据预设的阈值计算公式计算上述压缩阈值,上述阈值计算公式为 $H = \lambda \frac{\|g_0\|^2}{\sqrt{D}}$,

其中H为压缩阈值, λ 为预设的超参数, g_0 为第二旧总梯度,D为第二旧总梯度的维度数。

[0146] 在上述第一种可能的实施方式作为基础而提供的第四种可能的实施方式中,上述对上述变化量进行压缩得到压缩变化量之后,还包括:

[0147] 根据上述压缩变化量和上述旧节点梯度更新上述当前节点梯度,其中,更新后的当前节点梯度作为上述计算节点在迭代计算梯度过程中下一次计算得到的节点梯度对应的旧节点梯度。

[0148] 在上述第一种可能的实施方式作为基础而提供的第五种可能的实施方式中,上述将上述压缩变化量发送至参数服务器之后,还包括:

[0149] 接收上述参数服务器发送的总梯度变化量;

[0150] 根据上述总梯度变化量更新上述计算节点的模型参数。

[0151] 在上述第五种可能的实施方式作为基础而提供的第六种可能的实施方式中,上述根据上述总梯度变化量更新上述计算节点的模型参数,包括:

[0152] 根据上述总梯度变化量和存储在上述计算节点上的第二旧总梯度计算得到上述当前总梯度;

[0153] 根据上述当前总梯度更新上述计算节点的模型参数。

[0154] 上述终端设备7可以是手机、桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。该终端设备可包括,但不仅限于,第一处理器70、第一存储器71。本领域技术人员可以理解,图7仅仅是终端设备7的举例,并不构成对终端设备7的限定,可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件,例如还可以包括输入输出设备、网络接入设备等。

[0155] 所称第一处理器70可以是中央处理单元(Central Processing Unit,CPU),该第一处理器70还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field-Programmable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0156] 上述第一存储器71在一些实施例中可以是上述终端设备7的内部存储单元,例如终端设备7的硬盘或内存。上述第一存储器71在另一些实施例中也可以是上述终端设备7的外部存储设备,例如上述终端设备7上配备的插接式硬盘,智能存储卡(Smart Media Card, SMC),安全数字(Secure Digital,SD)卡,闪存卡(Flash Card)等。进一步地,上述第一存储器71还可以既包括上述终端设备7的内部存储单元也包括外部存储设备。上述第一存储器

71用于存储操作系统、应用程序、引导装载程序(BootLoader)、数据以及其他程序等,例如上述第一计算机程序的程序代码等。上述第一存储器71还可以用于暂时地存储已经输出或者将要输出的数据。

[0157] 图8为本申请一实施例提供的参数服务器的结构示意图。如图8所示,该实施例的参数服务器8包括:至少一个第二处理器80(图8中仅示出一个)、第二存储器81以及存储在上述第二存储器81中并可在上述至少一个第二处理器80上运行的第二计算机程序82,上述第二处理器80执行上述第二计算机程序82时实现以下步骤:

[0158] 接收各个计算节点发送的压缩变化量,其中,压缩变化量由计算节点的当前节点梯度与旧节点梯度之间的变化量经过压缩得到,旧节点梯度为计算节点在迭代计算梯度过程中上一次计算得到的节点梯度;

[0159] 根据各个压缩变化量计算得到总梯度变化量;

[0160] 根据上述总梯度变化量和第一旧总梯度计算得到当前总梯度,其中,上述第一旧总梯度为上述参数服务器在迭代计算梯度过程中上一次计算得到的总梯度。

[0161] 假设上述为第一种可能的实施方式,则在第一种可能的实施方式作为基础而提供的第二种可能的实施方式中,上述根据上述总梯度变化量和第一旧总梯度计算得到当前总梯度,包括:

[0162] 分别将上述总梯度变化量中各个维度的梯度分量的变化量与第一旧总梯度中对应的各个维度的梯度分量相加,得到上述当前总梯度。

[0163] 在上述第一种可能的实施方式作为基础而提供的第三种可能的实施方式中,上述根据上述总梯度变化量和第一旧总梯度计算得到当前总梯度之后,还包括:

[0164] 根据上述当前总梯度更新上述参数服务器的模型参数;

[0165] 将上述总梯度变化量发送给各个计算节点。

[0166] 该参数服务器可包括,但不仅限于,第二处理器80、第二存储器81。本领域技术人员可以理解,图8仅仅是参数服务器8的举例,并不构成对参数服务器8的限定,可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件,例如还可以包括输入输出设备、网络接入设备等。

[0167] 所称第二处理器80可以是中央处理单元(Central Processing Unit,CPU),该第二处理器80还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field-Programmable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0168] 上述第二存储器81在一些实施例中可以是上述参数服务器8的内部存储单元,例如参数服务器8的硬盘或内存。上述第二存储器81在另一些实施例中也可以是上述参数服务器8的外部存储设备,例如上述参数服务器8上配备的插接式硬盘,智能存储卡(Smart Media Card,SMC),安全数字(Secure Digital,SD)卡,闪存卡(Flash Card)等。进一步地,上述第二存储器81还可以既包括上述参数服务器8的内部存储单元也包括外部存储设备。上述第二存储器81用于存储操作系统、应用程序、引导装载程序(BootLoader)、数据以及其他程序等,例如上述第二计算机程序的程序代码等。上述第二存储器81还可以用于暂时地

存储已经输出或者将要输出的数据。

[0169] 需要说明的是,上述装置/单元之间的信息交互、执行过程等内容,由于与本申请方法实施例基于同一构思,其具体功能及带来的技术效果,具体可参见方法实施例部分,此处不再赘述。

[0170] 所属领域的技术人员可以清楚地了解到,为了描述的方便和简洁,仅以上述各功能单元、模块的划分进行举例说明,实际应用中,可以根据需要而将上述功能分配由不同的功能单元、模块完成,即将上述装置的内部结构划分成不同的功能单元或模块,以完成以上描述的全部或者部分功能。实施例中的各功能单元、模块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,上述集成的单元既可以采用硬件的形式实现,也可以采用软件功能单元的形式实现。另外,各功能单元、模块的具体名称也只是为了便于相互区分,并不用于限制本申请的保护范围。上述系统中单元、模块的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0171] 本申请实施例还提供了一种计算机可读存储介质,上述计算机可读存储介质存储有计算机程序,上述计算机程序被处理器执行时实现可实现上述各个方法实施例中的步骤。

[0172] 本申请实施例提供了一种计算机程序产品,当计算机程序产品在终端设备上运行时,使得终端设备执行时实现可实现上述各个方法实施例中的步骤。

[0173] 上述集成的单元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存储介质中。基于这样的理解,本申请实现上述实施例方法中的全部或部分流程,可以通过计算机程序来指令相关的硬件来完成,上述的计算机程序可存储于一计算机可读存储介质中,该计算机程序在被处理器执行时,可实现上述各个方法实施例的步骤。其中,上述计算机程序包括计算机程序代码,上述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。上述计算机可读介质至少可以包括:能够将计算机程序代码携带到拍照装置/终端设备的任何实体或装置、记录介质、计算机存储器、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、电载波信号、电信信号以及软件分发介质。例如U盘、移动硬盘、磁碟或者光盘等。在一些实施例中,计算机可读介质不可以是电载波信号和电信信号。

[0174] 在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中未详述或记载的部分,可以参见其它实施例的相关描述。

[0175] 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的范围。

[0176] 在本申请所提供的实施例中,应该理解到,所揭露的装置/网络设备和方法,可以通过其它的方式实现。例如,以上所描述的装置/网络设备实施例仅仅是示意性的,例如,上述模块或单元的划分,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讯连接可以是通过一些接口,装置

或单元的间接耦合或通讯连接,可以是电性,机械或其它的形式。

[0177] 上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。

[0178] 以上上述实施例仅用以说明本申请的技术方案,而非对其限制;尽管参照前述实施例对本申请进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围,均应包含在本申请的保护范围之内。

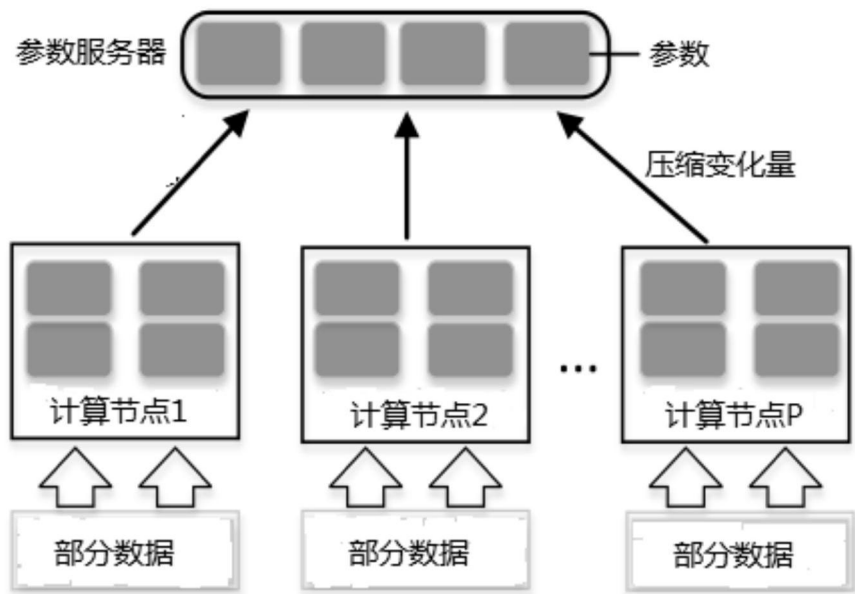


图1

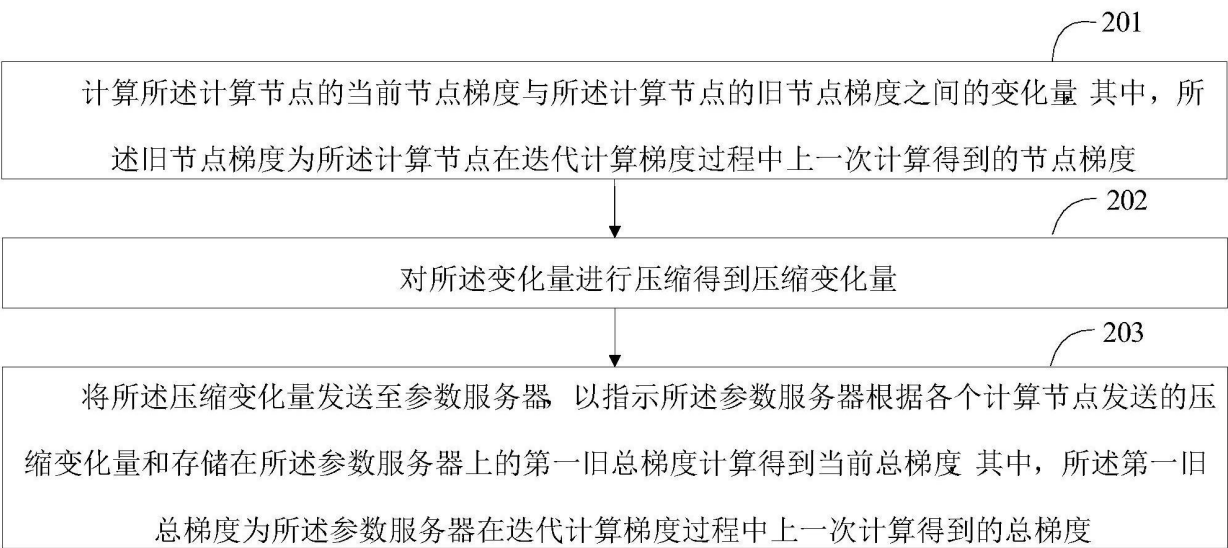


图2

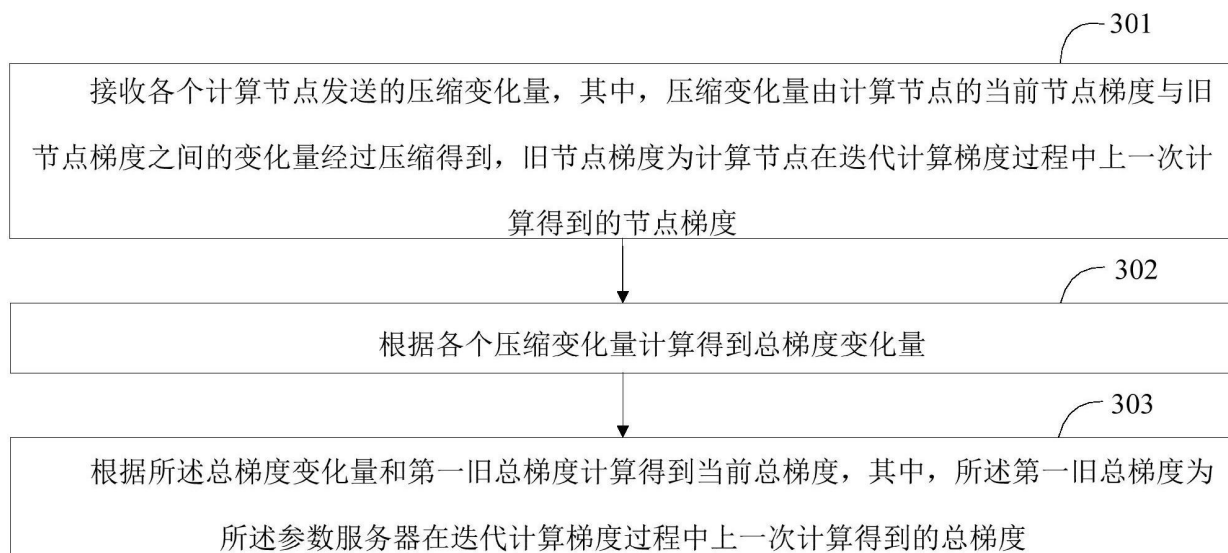


图3

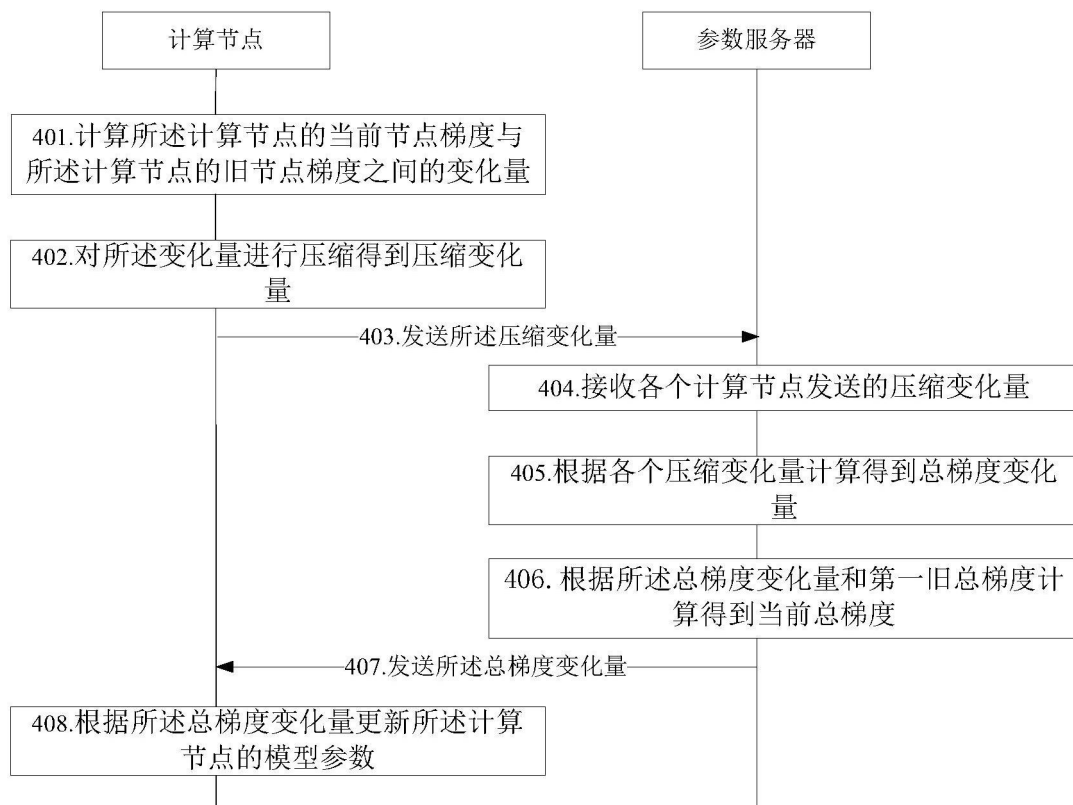


图4

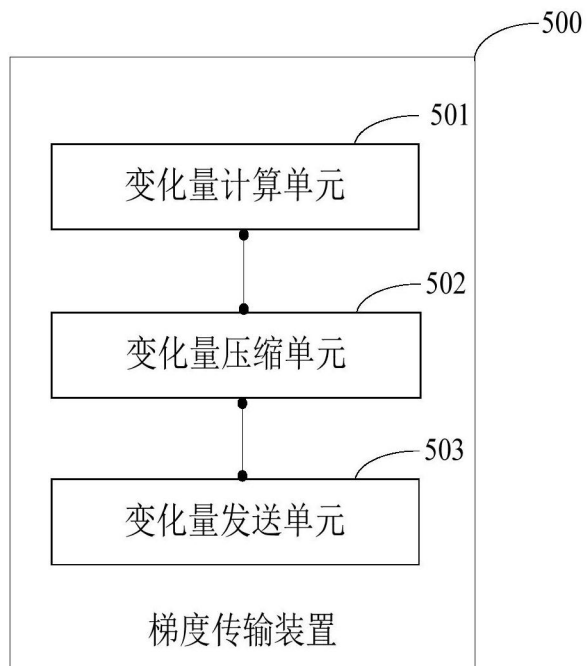


图5

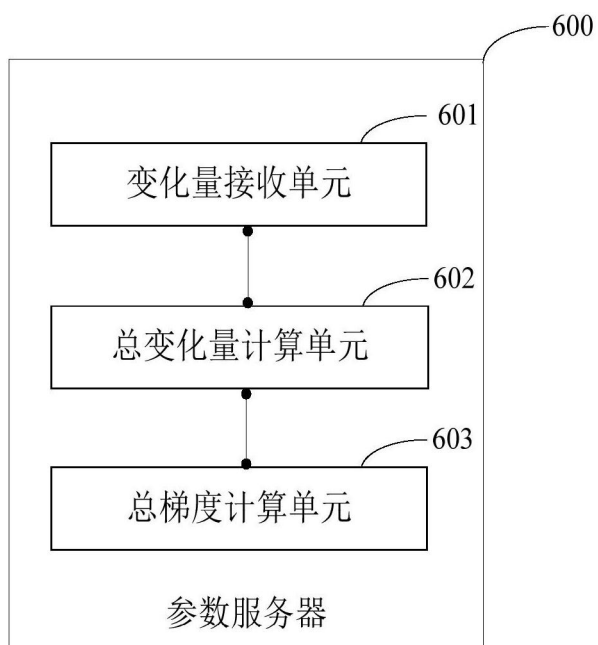


图6

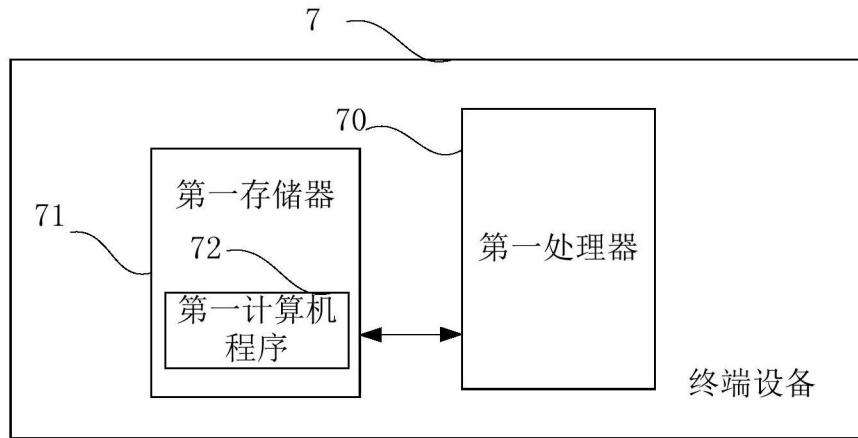


图7

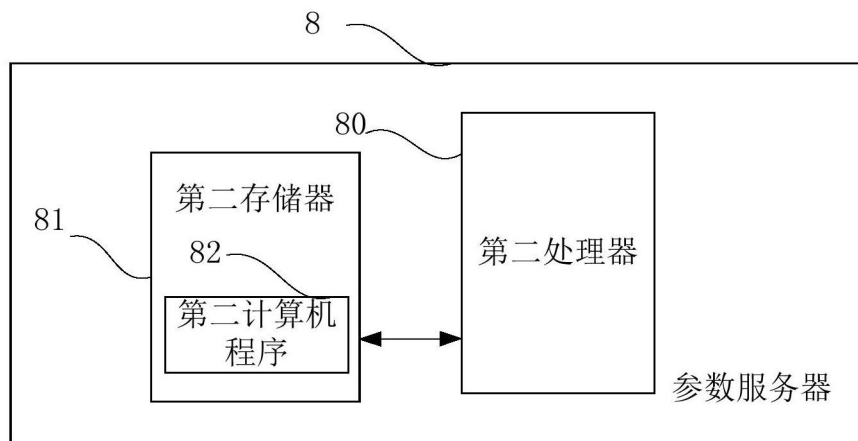


图8