(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114254756 B
(45)授权公告日 2024.09.10

(21)申请号 202011016938.4

(22)申请日 2020.09.24

(65)同一申请的已公布的文献号
申请公布号 CN 114254756 A

(43)申请公布日 2022.03.29

(73)专利权人 香港理工大学深圳研究院
地址 518057 广东省深圳市南山区高新园南区粤兴一道18号香港理工大学产学研大楼205室

(72)发明人 郭嵩 王号召 詹玉峰

(74)专利代理机构 深圳中一专利商标事务所 44237
专利代理师 雷浩

(51)Int.Cl.
G06N 20/00(2019.01)
H04L 67/10(2022.01)

(56)对比文件
CN 109754060 A,2019.05.14
US 2018253423 A1,2018.09.06

审查员 李宇文

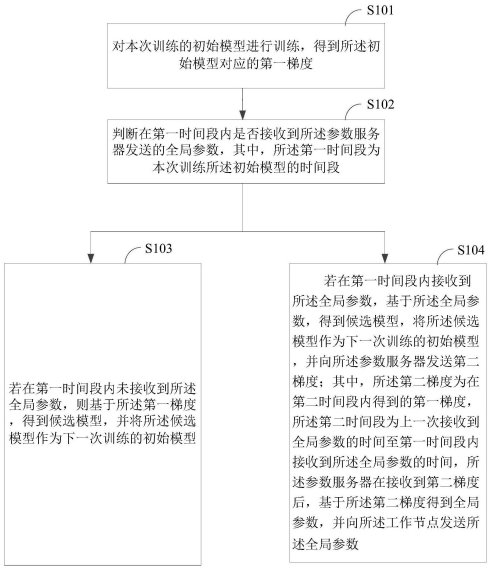
权利要求书2页 说明书12页 附图5页

(54)发明名称

一种分布式机器学习方法、装置、终端设备及存储介质

(57)摘要

本申请适用于计算机技术领域,提供了一种分布式机器学习方法、装置、终端设备及存储介质,该方法包括:对本次训练的初始模型进行训练,得到初始模型对应的第一梯度;判断在第一时间段内是否接收到参数服务器发送的全局参数,其中,第一时间段为本次训练初始模型的时间段;若在第一时间段内未接收到全局参数,则基于第一梯度,得到候选模型,并将候选模型作为下一次训练的初始模型;本申请在没有接收到全局参数时使用第一梯度得到候选模型,并对候选模型继续训练,使参数服务器在计算全局参数和向工作节点传输全局参数的时间内,工作节点一直处于训练的状态,不用必须接收到全局参数后再继续训练,节约了模型训练的时间,使模型训练速度更快。



CN 114254756 B

1.一种分布式机器学习方法,应用于分布式机器学习系统,所述分布式机器学习系统包括参数服务器和至少两个用于对模型进行训练的工作节点,所述工作节点与所述参数服务器相连,其特征在于,该方法包括:

对本次训练的初始模型进行训练,得到所述初始模型对应的第一梯度;

判断在第一时间段内是否接收到所述参数服务器发送的全局参数,其中,所述第一时间段为本次训练所述初始模型的时间段;

若在第一时间段内未接收到所述全局参数,则基于所述第一梯度,得到候选模型,并将所述候选模型作为下一次训练的初始模型。

2.如权利要求1所述的分布式机器学习方法,其特征在于,所述基于所述第一梯度,得到候选模型,包括:

基于所述第一梯度更新所述初始模型的参数,得到候选模型。

3.如权利要求1或2所述的分布式机器学习方法,其特征在于,在所述判断在第一时间段内是否接收到所述参数服务器发送的全局参数之后,还包括:

若在第一时间段内接收到所述全局参数,基于所述全局参数,得到候选模型,将所述候选模型作为下一次训练的初始模型,并向所述参数服务器发送第二梯度;

其中,所述第二梯度为在第二时间段内得到的第一梯度,所述第二时间段为上一次接收到全局参数的时间至第一时间段内接收到所述全局参数的时间,所述参数服务器在接收到第二梯度后,基于所述第二梯度得到全局参数,并向所述工作节点发送所述全局参数。

4.如权利要求3所述的分布式机器学习方法,其特征在于,所述基于所述全局参数,得到候选模型,包括:

基于全局参数更新所述初始模型的参数,得到候选模型。

5.如权利要求3所述的分布式机器学习方法,其特征在于,所述向所述参数服务器发送第二梯度,包括:

对所述第二梯度进行降维处理,得到目标梯度;

向所述参数服务器发送所述目标梯度。

6.如权利要求5所述的分布式机器学习方法,其特征在于,在所述对所述第二梯度进行降维处理,得到目标梯度之前,还包括:

判断所述第二梯度的个数是否大于1;

若所述第二梯度的个数大于1,则计算所有第二梯度的和,得到候选梯度;

相应的,对所述第二梯度进行降维处理,得到目标梯度,包括:

对所述候选梯度进行降维处理,得到目标梯度。

7.如权利要求2所述的分布式机器学习方法,其特征在于,所述基于所述第一梯度更新所述初始模型的参数,包括:

基于所述第一梯度,利用梯度下降法更新所述初始模型的参数。

8.一种分布式机器学习系统,其特征在于,包括:参数服务器和至少两个用于对模型进行训练的工作节点,工作节点与所述参数服务器相连;

其中,所述工作节点包括:

模型训练模块,用于对本次训练的初始模型进行训练,得到所述初始模型对应的第一梯度;

判断模块,用于判断在第一时间段内是否接收到所述参数服务器发送的全局参数,其中,所述第一时间段为工作节点训练当前模型的时间段;

参数更新模块,用于若在第一时间段内未接收到所述全局参数,则基于所述第一梯度,得到候选模型,并将所述候选模型作为当前模型进行下一次模型训练。

9.一种终端设备,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7任一项所述的分布式机器学习方法。

10.一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至7任一项所述的分布式机器学习方法。

一种分布式机器学习方法、装置、终端设备及存储介质

技术领域

[0001] 本申请属于计算机技术领域,尤其涉及一种分布式机器学习方法、装置、终端设备及存储介质。

背景技术

[0002] 机器学习是计算机利用已有的数据,通过对初始模型进行训练,得到训练后的模型,并利用训练后的模型预测需要的数据。目前多采用分布式机器学习系统对模型进行训练。采用分布式机器学习系统训练模型的具体方法为:将训练样本分别输入并行的多台子服务器中,利用多台子服务器同时对模型进行训练,然后将训练得到的梯度发送至参数服务器,参数服务器利用梯度对全局参数进行更新,并更新后的全局参数返回至各个子服务器中,子服务器利用参数服务器返回的全局参数更新模型参数,并进行下一次训练,依此循环直到训练结束。

[0003] 上述方法在子服务器较多时,由于子服务器的数据处理能力存在差异,参数服务器需要接收到所有子服务器发送的梯度后才可以进行全局参数的更新,且子服务器需要在接收到全局参数后才能进行下一次的模型训练,由于全局参数返回时间较长,使模型训练的时间延长,降低了模型训练的效率。

发明内容

[0004] 本申请实施例提供了一种分布式机器学习方法、装置、终端设备及存储介质,可以解决目前模型训练效率低的问题。

[0005] 第一方面,本申请实施例提供了一种分布式机器学习方法,应用于分布式机器学习系统,所述分布式机器学习系统包括参数服务器和至少两个用于对模型进行训练的工作节点,所述工作节点与所述参数服务器相连,包括:

[0006] 对本次训练的初始模型进行训练,得到所述初始模型对应的第一梯度;

[0007] 判断在第一时间段内是否接收到所述参数服务器发送的全局参数,其中,所述第一时间段为本次训练所述初始模型的时间段;

[0008] 若在第一时间段内未接收到所述全局参数,则基于所述第一梯度,得到候选模型,并将所述候选模型作为下一次训练的初始模型。

[0009] 第二方面,本申请实施例提供了一种分布式机器学习系统,包括:参数服务器和至少两个用于对模型进行训练的工作节点,工作节点与所述参数服务器相连;

[0010] 其中,所述工作节点包括:

[0011] 模型训练模块,用于对本次训练的初始模型进行训练,得到所述初始模型对应的第一梯度;

[0012] 判断模块,用于判断在第一时间段内是否接收到所述参数服务器发送的全局参数,其中,所述第一时间段为工作节点训练当前模型的时间段;

[0013] 参数更新模块,用于若在第一时间段内未接收到所述全局参数,则基于所述第一

梯度,得到候选模型,并将所述候选模型作为当前模型进行下一次模型训练。

[0014] 第三方面,本申请实施例提供了一种终端设备,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上述第一方面中任一项所述的分布式机器学习方法。

[0015] 第四方面,本申请实施例提供了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现上述第一方面中任一项所述的分布式机器学习方法。

[0016] 第五方面,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,使得终端设备执行上述第一方面中任一项所述的分布式机器学习方法。

[0017] 可以理解的是,上述第二方面至第五方面的有益效果可以参见上述第一方面中的相关描述,在此不再赘述。

[0018] 本申请实施例与现有技术相比存在的有益效果是:本申请通过工作节点对本次训练的初始模型进行训练,得到初始模型对应的第一梯度;然后判断在本次训练初始模型的时间段内是否接收到参数服务器发送的全局参数;若在第一时间段内未接收到全局参数,则基于第一梯度,得到候选模型,并将候选模型作为下一次训练的初始模型;本申请在没有接收到全局参数时使用第一梯度得到候选模型,并对候选模型继续训练,使参数服务器在计算全局参数和向工作节点传输全局参数的时间内,工作节点一直处于训练的状态,不用必须接收到全局参数后再继续训练,节约了模型训练的时间,使模型训练速度更快。

附图说明

[0019] 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性的前提下,还可以根据这些附图获得其他的附图。

[0020] 图1是本申请一实施例提供的分布式机器学习系统的示意图;

[0021] 图2是本申请一实施例提供的分布式机器学习方法的流程示意图;

[0022] 图3是本申请一实施例提供的对第二梯度进行处理的流程示意图;

[0023] 图4是本申请一实施例提供的工作节点训练模型的方法的流程示意图;

[0024] 图5是本申请一实施例提供的工作节点的结构示意图;

[0025] 图6是本申请一实施例提供的终端设备的结构示意图;

[0026] 图7是本申请一实施例提供的计算机的部分结构的框图。

具体实施方式

[0027] 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。

[0028] 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。

[0029] 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。

[0030] 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]”。

[0031] 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。

[0032] 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。

[0033] 如图1所示,分布式机器学习系统包括参数服务器和4个用于对模型进行训练的工作节点,工作节点与所述参数服务器之间通过链路连接。

[0034] 将样本集分成4份,分别为数据块D1、数据块D2、数据块D3和数据块D4,将数据块D1分配至工作节点W1,数据块D2分配至工作节点W2,将数据块D3分配至工作节点W3,将数据块D4分配至工作节点W4。4个工作节点分别计算与之对应的数据块的梯度,并将计算的梯度发送至参数服务器。参数服务器接收到4个工作节点传输的梯度后,对接收到的梯度进行处理,得到全局参数,并将全局参数发送至各个工作节点,各个工作节点利用接收到的全局参数对模型进行参数更新,并对更新参数后的模型进行训练。

[0035] 在分布式机器学习系统中工作节点将梯度传输至参数服务器、参数服务器计算全局参数、以及参数服务器将全局参数传输至工作节点的时间内,工作节点处于不工作状态,工作节点必须接收到全局参数后才开始继续工作,导致工作节点在两次训练之间的时间的浪费。

[0036] 因此,本申请提供的分布式机器学习方法,工作节点一直处于工作状态,提高了模型训练的效率。

[0037] 以下结合图1对本申请实施例的分布式机器学习方法进行详细说明。

[0038] 图2示出了本申请提供的分布式机器学习方法的示意性流程图,参照图2,对该方法的详述如下:

[0039] S101,对本次训练的初始模型进行训练,得到所述初始模型对应的第一梯度。

[0040] 在本实施例中,工作节点是对模型进行迭代训练的节点,通过不断的更新模型参数进行模型的不断训练。工作节点基于训练数据对模型进行训练。训练数据可以是预先存储在工作节点中的数据,也可以是从外部设备中传输至工作节点的数据。

[0041] 在本实施例中,工作节点对模型进行训练,每训练一次得到一个第一梯度。梯度的本意是一个向量(矢量),表示某一函数在该点处的方向导数沿着该方向取得最大值,即函数在该点处沿着该方向(此梯度的方向)变化最快,变化率最大(为该梯度的模)。将在工作

节点中的当前模型作为本次训练的初始模型。工作节点对初始模型进行训练,可以得到初始模型对应的第一梯度。

[0042] S102,判断在第一时间段内是否接收到所述参数服务器发送的全局参数,其中,所述第一时间段为本次训练所述初始模型的时间段。

[0043] 在本实施例中,全局参数是参数服务器基于接收到的各个工作节点发送的第一梯度进行初始全局参数更新,得到更新后的初始全局参数作为全局参数,并向各个节点发送全局参数。

[0044] 在本实施例中,第一时间段为本次对初始模型开始训练的时间至对初始模型结束训练的时间,例如,从8点30分开始本次对初始模型训练,到8点40分对初始模型训练结束,则第一时间段为从8点30分至8点40分。

[0045] S103,若在第一时间段内未接收到所述全局参数,则基于所述第一梯度,得到候选模型,并将所述候选模型作为下一次训练的初始模型。

[0046] 在一种可能的实现方式中,步骤S103的实现过程可以包括:

[0047] 基于所述第一梯度更新所述初始模型的参数,得到候选模型。

[0048] 具体的,基于所述第一梯度,利用梯度下降法更新所述初始模型的参数。

[0049] 在本实施例中,如果工作节点在本次对初始模型进行训练的期间,该工作节点未接收到参数服务器传输的全局参数,说明参数服务器还未计算完,还没有得到全局参数,因此工作节点不可以使用全局参数更新初始模型中的参数。现有技术是继续等待直到接收到参数服务器发送的全局参数后,使用全局参数更新模型的参数。本申请根据本次对初始模型训练得到的第一梯度更新初始模型中的参数,然后将更新参数后的初始模型作为候选模型,对候选模型继续进行训练。本申请不用等待全局参数,工作节点可以一直处于工作状态。

[0050] 可选的,可设置在各工作节点开始工作的预设时间后,参数服务器将预先存储的一个全局参数发送至各工作节点,以使得工作节点基于全局参数继续按照上述步骤S101-S103继续训练模型。例如,在各工作节点开始工作3分钟后,参数服务器向各工作节点发送全局参数a。

[0051] 可选的,在工作节点开始对模型进行第一次训练时,参数服务器可以根据预存的数据计算第一个全局参数,并在计算完后将第一个全局参数发送至工作节点。工作节点在接收到第一个全局参数后需要将接收到第一个全局参数的时间之前计算的所有第一梯度发送至参数服务器。

[0052] 可选的,各个工作节点第一次对模型训练得到的第一梯度,可以将第一次得到的第一梯度上传至参数服务器,以启动参数服务器计算全局参数。那么,各工作节点第二次上传的梯度,为第一次接收到全局参数的时间之前计算的所有第一梯度。

[0053] 可选的,在步骤S101之后,还可以包括:

[0054] 判断本次训练是否为第一次对模型进行训练,如果是第一次进行模型训练,则向参数服务器发送第一梯度。如果不是第一次进行模型训练,则继续步骤S102的判断。

[0055] 本申请实施例中,通过工作节点对本次训练的初始模型进行训练,得到初始模型对应的第一梯度;然后判断在本次训练初始模型的时间段内是否接收到参数服务器发送的全局参数;若在第一时间段内未接收到全局参数,则基于第一梯度,得到候选模型,并将候

选模型作为下一次训练的初始模型;本申请在没有接收到全局参数时使用第一梯度得到候选模型,并对候选模型继续训练,使参数服务器在计算全局参数和向工作节点传输全局参数的时间内,工作节点一直处于训练的状态,不用必须接收到全局参数后再继续训练,节约了模型训练的时间,使模型训练速度更快。本申请使用工作节点的模型训练与数据传输的并行处理机制,充分利用数据传输的时间和全局参数的计算时间,工作节点一直处于工作状态,加快了模型的训练。

[0056] 在一种可能的实现方式中,在步骤S103之后,还可以包括:

[0057] S104,若在第一时间段内接收到所述全局参数,基于所述全局参数,得到候选模型,将所述候选模型作为下一次训练的初始模型,并向所述参数服务器发送第二梯度;其中,所述第二梯度为在第二时间段内得到的第一梯度,所述第二时间段为上一次接收到全局参数的时间至第一时间段内接收到所述全局参数的时间,所述参数服务器在接收到第二梯度后,基于所述第二梯度得到全局参数,并向所述工作节点发送所述全局参数。

[0058] 具体的,基于全局参数更新所述初始模型的参数,得到候选模型。

[0059] 在本实施例中,如果在本次对初始模型训练期间,工作节点接收到参数服务器发送的全局参数,则需要根据全局参数更新初始模型的参数,将更新参数后的初始模型作为候选模型,对候选模型进行训练,将对候选模型的训练作为本次训练。

[0060] 在本实施例中,如果工作节点接收到参数服务器发送的全局参数,则说明参数服务器已经基于接收到的工作节点在本次初始训练之前传输的第一梯度完成了全局参数的计算,需要将新的第一梯度传输至参数服务器。

[0061] 具体的,由于在上一次接收到全局参数的时间至第一时间段内接收到所述全局参数的时间内是没有向参数服务器发送第一梯度的,因此,需要将工作节点在上一次接收到全局参数的时间至第一时间段内得到的所有第一梯度上传至参数服务器。因为在上一次接收到全局参数的时间至第一时间段内工作节点对模型进行训练的次数可能大于1,因此,需要将每次训练得到的第一梯度均上传至参数服务器。还可以将需要上传的多个第二梯度组成梯度集合,对梯度集合进行编码后发送至参数服务器。

[0062] 在本实施例中,第二时间段为工作节点接收到相邻两次全局参数的时间段,例如,工作节点b第2次接收到全局参数的时间为5点10分,第3次接收到全局参数的时间为5点30分,则第二时间段为5点10分至5点30分。

[0063] 作为举例,如果上一次接收到全局参数的时间为8点10分,第一时间段内接收到所述全局参数的时间为8点40分。在8点10分至8点40分之间工作节点进行了4次模型训练,分别得到4个第一梯度,则将4个第一梯度均传输至参数服务器。

[0064] 在本实施例中,参数服务器需要获取到预设个数的工作节点发送的第一梯度后,再次进行全局参数的计算。

[0065] 需要说明是,如果参数服务器计算的第一个全局参数是根据预存在参数服务器中的数据计算的,则工作节点在接收到第一个全局参数后,需要将接收到第一个全局参数的时间之前计算的所有第一梯段发送至参数服务器。也就是第二时间段为工作节点的启动时间至接收到第一个全局参数的时间。

[0066] 如图3所述,在一种可能的实现方式中,步骤S104中向所述参数服务器发送第二梯度的实现过程可以包括:

[0067] S1041,对所述第二梯度进行降维处理,得到目标梯度;

[0068] S1042,向所述参数服务器发送所述目标梯度。

[0069] 在本实施例中,对第二梯度进行降维处理,降维处理是将高维数据化为低维度数据的操作,例如,第二梯度为32bit的浮点数,可以将第二梯度转换成1bit的数据。对第二梯度进行降维处理可以减少通信数据,降低通信时间。

[0070] 在一种可能的实现方式中,在步骤S1041之前,上述方法还可以包括:

[0071] 判断所述第二梯度的个数是否大于1;

[0072] 若所述第二梯度的个数大于1,则计算所有第二梯度的和,得到候选梯度。

[0073] 在本实施例中,如果第二梯度的个数大于1,可以将所有的第二梯度相加,得到候选梯度,对候选梯度进行降维,并将降维后的候选梯度传输至参数服务器。

[0074] 在一种可能的实现方式中,分布式机器学习方法的实现过程还可以包括:

[0075] 如果分布式机器学习系统包括两个工作节点,分别为W1和W2,一个参数服务器。

[0076] S201,W1对初始模型M1进行第一次训练,得到第一梯度TW1;W2对初始模型M2进行第一次训练,得到第一梯度TW2,并分别将TW1和TW2发送至参数服务器;

[0077] S202,参数服务器接收到TW1和TW2后进行第一个全局参数的计算。

[0078] 与此同时,W1利用TW1更新初始模型M1的参数,得到候选模型M11,并对候选模型M11进行训练,得到第一梯度TW11。

[0079] W2利用TW2更新初始模型M2的参数,得到候选模型M22,并对候选模型M22进行训练,得到第一梯度TW22。

[0080] S203,W1利用TW11更新初始模型M11的参数,得到候选模型M111,并对候选模型M111进行训练,得到第一梯度TW111。在对M111进行训练期间,W1接收到参数服务器发送的全局参数Q1。

[0081] W2利用TW22更新初始模型M22的参数,得到候选模型M222,并对候选模型M222进行训练,得到第一梯度TW222。

[0082] S204,W1基于获取到的全局参数Q1更新M111中的参数,得到候选模型M1111,并对候选模型进行训练,得到第一梯度TW1111。与此同时,W1将TW1、TW11发送至参数服务器。

[0083] W2利用TW222更新初始模型M222的参数,得到候选模型M2222,并对候选模型M2222进行训练,得到第一梯度TW2222。在对M2222进行训练期间,W2接收到参数服务器发送的全局参数Q1。

[0084] S205,W1利用TW1111更新初始模型M1111的参数,得到候选模型M11111,并对候选模型M11111进行训练,得到第一梯度TW11111。

[0085] W2利用全局参数Q1更新初始模型M2222的参数,得到候选模型M22222,并对候选模型M22222进行训练,得到第一梯度TW22222。与此同时,W2将TW2、TW22和TW222发送至参数服务器。

[0086] S206,参数服务器基于TW1、TW11、TW2、TW22和TW222计算全局参数Q2。W1和W2继续依照上述方法对模型进行训练,直到完成模型的训练。

[0087] 在一种可能的实现方式中,分布式机器学习方法的实现过程还可以包括:

[0088] 如图4所示,分布式机器学习系统包括两个工作节点,分别为W1和W2,一个参数服务器。箭头表示工作节点对模型训练过程,箭头的长短表示对模型训练所用的时间。方形表

示工作节点将第一梯度传输至参数服务器的时间至工作节点接收到参数服务器发送的参数服务的时间。一个虚线格内表示参数服务器完成一次全局参数的计算及传输。

[0089] S301,W1对初始模型M1进行第一次训练,得到第一梯度TW1;W2对初始模型M2进行第一次训练,得到第一梯度TW2;

[0090] S302,W1利用TW1更新初始模型M1的参数,得到候选模型M11,并对候选模型M11进行训练,得到第一梯度TW11。在对M11进行训练期间,W1接收到参数服务器发送的全局参数Q1。

[0091] W2利用TW2更新初始模型M2的参数,得到候选模型M22,并对候选模型M22进行训练,得到第一梯度TW22。

[0092] S303,W1将接收到全局参数Q1时间之前得到的第一梯度发送至参数服务器,即将TW1发送至参数服务器。

[0093] W1根据全局参数Q1更新M11中的参数,得到候选模型M111,并对候选模型M111进行训练,得到第一梯度TW111。与此同时,在对候选模型M111进行训练期间,接收到参数服务器发送的全局参数Q2。

[0094] W2根据第一梯度TW22更新M22中的参数,得到候选模型M222,并对候选模型M222进行训练,得到第一梯度TW222。在对M222进行训练期间,W2接收到参数服务器发送的全局参数Q1。

[0095] S304,W1将接收到全局参数Q1时间至接收到全局参数Q2时间内得到的第一梯度发送至参数服务器,即将TW11发送至参数服务器。W2将接收到全局参数Q1时间之前得到的第一梯度发送至参数服务器,即将TW2和TW22发送至参数服务器。

[0096] W1根据全局参数Q2更新M111中的参数,得到候选模型M1111,并对候选模型M1111进行训练,得到第一梯度TW1111。与此同时,在对候选模型M1111进行训练期间,接收到参数服务器发送的全局参数Q3。

[0097] W2根据全局参数Q1更新M222中的参数,得到候选模型M2222,并对候选模型M2222进行训练,得到第一梯度TW2222。

[0098] S305,W1将接收到全局参数Q2时间至接收到全局参数Q3时间内得到的第一梯度发送至参数服务器,即将TW111发送至参数服务器。

[0099] W1根据全局参数Q3更新M1111中的参数,得到候选模型M11111,并对候选模型M11111进行训练,得到第一梯度TW11111。

[0100] W2根据第一梯度TW2222更新M2222中的参数,得到候选模型M22222,并对候选模型M22222进行训练,得到第一梯度TW22222。与此同时,在对候选模型M22222进行训练期间,接收到参数服务器发送的全局参数Q2。W2将TW222和TW2222发送至参数服务器。

[0101] S306,依照上述方法对模型进行训练,直到训练结束。

[0102] 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,而不应对本申请实施例的实施过程构成任何限定。

[0103] 对应于上文实施例所述的分布式机器学习方法,本申请实施例提供的分布式机器学习系统,包括:参数服务器和至少两个用于对模型进行训练的工作节点,工作节点与所述参数服务器相连。

[0104] 参照图5,工作节点400可以包括:模型训练模块410、判断模块420和参数更新模块430。

[0105] 其中,模型训练模块410,用于对本次训练的初始模型进行训练,得到所述初始模型对应的第一梯度;

[0106] 判断模块420,用于判断在第一时间段内是否接收到所述参数服务器发送的全局参数,其中,所述第一时间段为工作节点训练当前模型的时间段;

[0107] 参数更新模块430,用于若在第一时间段内未接收到所述全局参数,则基于所述第一梯度,得到候选模型,并将所述候选模型作为当前模型进行下一次模型训练。

[0108] 在一种可能的实现方式中,参数更新模块430具体可以用于:

[0109] 基于所述第一梯度更新所述初始模型的参数,得到候选模型。

[0110] 在一种可能的实现方式中,与判断模块420相连的还包括:

[0111] 数据更新模块,用于若在第一时间段内接收到所述全局参数,基于所述全局参数,得到候选模型,将所述候选模型作为下一次训练的初始模型,并向所述参数服务器发送第二梯度;

[0112] 其中,所述第二梯度为在第二时间段内得到的第一梯度,所述第二时间段为上一次接收到全局参数的时间至第一时间段内接收到所述全局参数的时间,所述参数服务器在接收到第二梯度后,基于所述第二梯度得到全局参数,并向所述工作节点发送所述全局参数。

[0113] 在一种可能的实现方式中,数据更新模块具体可以用于:

[0114] 基于全局参数更新所述初始模型的参数,得到候选模型。

[0115] 在一种可能的实现方式中,数据更新模块具体可以用于:

[0116] 对所述第二梯度进行降维处理,得到目标梯度;

[0117] 向所述参数服务器发送所述目标梯度。

[0118] 在一种可能的实现方式中,数据更新模块具体可以用于:

[0119] 判断所述第二梯度的个数是否大于1;

[0120] 若所述第二梯度的个数大于1,则计算所有第二梯度的和,得到候选梯度;

[0121] 相应的,对所述第二梯度进行降维处理,得到目标梯度,包括:

[0122] 对所述候选梯度进行降维处理,得到目标梯度。

[0123] 在一种可能的实现方式中,参数更新模块430具体还可以用于:

[0124] 基于所述第一梯度,利用梯度下降法更新所述初始模型的参数。

[0125] 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。

[0126] 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,即将所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。

[0127] 本申请实施例还提供了一种终端设备,参见图6,该终端设备500可以包括:至少一个处理器510、存储器520以及存储在所述存储器520中并可在所述至少一个处理器510上运行的计算机程序,所述处理器510执行所述计算机程序时实现上述任意各个方法实施例中的步骤,例如图2所示实施例中的步骤S101至步骤S103。或者,处理器510执行所述计算机程序时实现上述各装置实施例中各模块/单元的功能,例如图5所示模块410至430的功能。

[0128] 示例性的,计算机程序可以被分割成一个或多个模块/单元,一个或者多个模块/单元被存储在存储器520中,并由处理器510执行,以完成本申请。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机程序段,该程序段用于描述计算机程序在终端设备500中的执行过程。

[0129] 本领域技术人员可以理解,图6仅仅是终端设备的示例,并不构成对终端设备的限定,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,例如输入输出设备、网络接入设备、总线等。

[0130] 处理器510可以是中央处理单元(Central Processing Unit,CPU),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。

[0131] 存储器520可以是终端设备的内部存储单元,也可以是终端设备的外部存储设备,例如插接式硬盘,智能存储卡(Smart Media Card,SMC),安全数字(Secure Digital,SD)卡,闪存卡(Flash Card)等。所述存储器520用于存储所述计算机程序以及终端设备所需的其他程序和数据。所述存储器520还可以用于暂时地存储已经输出或者将要输出的数据。

[0132] 总线可以是工业标准体系结构(Industry Standard Architecture,ISA)总线、外部设备互连(Peripheral Component,PCI)总线或扩展工业标准体系结构(Extended Industry Standard Architecture,EISA)总线等。总线可以分为地址总线、数据总线、控制总线等。为便于表示,本申请附图中的总线并不限定仅有一根总线或一种类型的总线。

[0133] 本申请实施例提供的分布式机器学习方法可以应用于计算机、平板电脑、笔记本电脑、上网本、个人数字助理(personal digital assistant,PDA)等终端设备上,本申请实施例对终端设备的具体类型不作任何限制。

[0134] 以所述终端设备为计算机为例。图7示出的是与本申请实施例提供的计算机的部分结构的框图。参考图7,计算机包括:通信电路610、存储器620、输入单元630、显示单元640、音频电路660、无线保真(wireless fidelity,WiFi)模块660、处理器670以及电源680等部件。

[0135] 下面结合图7对计算机的各个构成部件进行具体的介绍:

[0136] 通信电路610可用于收发信息或通话过程中,信号的接收和发送,特别地,将图像采集设备发送的图像样本接收后,给处理器670处理;另外,将图像采集指令发送给图像采集设备。通常,通信电路包括但不限于天线、至少一个放大器、收发信机、耦合器、低噪声放

大器(Low Noise Amplifier,LNA)、双工器等。此外,通信电路610还可以通过无线通信与网络和其他设备通信。上述无线通信可以使用任一通信标准或协议,包括但不限于全球移动通讯系统(Global System of Mobile communication,GSM)、通用分组无线服务(General Packet Radio Service,GPRS)、码分多址(Code Division Multiple Access,CDMA)、宽带码分多址(Wideband Code Division Multiple Access,WCDMA)、长期演进(Long Term Evolution,LTE))、电子邮件、短消息服务(Short Messaging Service,SMS)等。

[0137] 存储器620可用于存储软件程序以及模块,处理器670通过运行存储在存储器620的软件程序以及模块,从而执行计算机的各种功能应用以及数据处理。存储器620可主要包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需的应用程序(比如声音播放功能、图像播放功能等)等;存储数据区可存储根据计算机的使用所创建的数据(比如音频数据、电话本等)等。此外,存储器620可以包括高速随机存取存储器,还可以包括非易失性存储器,例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

[0138] 输入单元630可用于接收输入的数字或字符信息,以及产生与计算机的用户设置以及功能控制有关的键信号输入。具体地,输入单元630可包括触控面板631以及其他输入设备632。触控面板631,也称为触摸屏,可收集用户在其上或附近的触摸操作(比如用户使用手指、触笔等任何适合的物体或附件在触控面板631上或在触控面板631附近的操作),并根据预先设定的程式驱动相应的连接装置。可选的,触控面板631可包括触摸检测装置和触摸控制器两个部分。其中,触摸检测装置检测用户的触摸方位,并检测触摸操作带来的信号,将信号传送给触摸控制器;触摸控制器从触摸检测装置上接收触摸信息,并将它转换成触点坐标,再送给处理器670,并能接收处理器670发来的命令并加以执行。此外,可以采用电阻式、电容式、红外线以及表面声波等多种类型实现触控面板631。除了触控面板631,输入单元630还可以包括其他输入设备632。具体地,其他输入设备632可以包括但不限于物理键盘、功能键(比如音量控制按键、开关按键等)、轨迹球、鼠标、操作杆等中的一种或多种。

[0139] 显示单元640可用于显示由用户输入的信息或提供给用户的信息以及计算机的各种菜单。显示单元640可包括显示面板641,可选的,可以采用液晶显示器(Liquid Crystal Display,LCD)、有机发光二极管(Organic Light-Emitting Diode,OLED)等形式来配置显示面板641。进一步的,触控面板631可覆盖显示面板641,当触控面板631检测到在其上或附近的触摸操作后,传送给处理器670以确定触摸事件的类型,随后处理器670根据触摸事件的类型在显示面板641上提供相应的视觉输出。虽然在图7中,触控面板631与显示面板641是作为两个独立的部件来实现计算机的输入和输入功能,但是在某些实施例中,可以将触控面板631与显示面板641集成而实现计算机的输入和输出功能。

[0140] 音频电路660可提供用户与计算机之间的音频接口。音频电路660可将接收到的音频数据转换后的电信号,传输到扬声器由扬声器转换为声音信号输出;另一方面,传声器将收集的声音信号转换为电信号,由音频电路660接收后转换为音频数据,再将音频数据输出处理器670处理后,经通信电路610以发送给比如另一计算机,或者将音频数据输出至存储器620以便进一步处理。

[0141] WiFi属于短距离无线传输技术,计算机通过WiFi模块660可以帮助用户收发电子邮件、浏览网页和访问流式媒体等,它为用户提供了无线的宽带互联网访问。虽然图7示出

了WiFi模块660,但是可以理解的是,其并不属于计算机的必须构成,完全可以根据需要在不改变发明的本质的范围内而省略。

[0142] 处理器670是计算机的控制中心,利用各种接口和线路连接整个计算机的各个部分,通过运行或执行存储在存储器620内的软件程序和/或模块,以及调用存储在存储器620内的数据,执行计算机的各种功能和处理数据,从而对计算机进行整体监控。可选的,处理器670可包括一个或多个处理单元;优选的,处理器670可集成应用处理器和调制解调处理器,其中,应用处理器主要处理操作系统、用户界面和应用程序等,调制解调处理器主要处理无线通信。可以理解的是,上述调制解调处理器也可以不集成到处理器670中。

[0143] 计算机还包括给各个部件供电的电源680(比如电池),优选的,电源680可以通过电源管理系统与处理器670逻辑相连,从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。

[0144] 本申请实施例还提供了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机程序,所述计算机程序被处理器执行时实现可实现上述分布式机器学习方法各个实施例中的步骤。

[0145] 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移动终端上运行时,使得移动终端执行时实现可实现上述分布式机器学习方法各个实施例中的步骤。

[0146] 所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取存储介质中。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,所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,所述计算机程序包括计算机程序代码,所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。所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。在某些司法管辖区,根据立法和专利实践,计算机可读介质不可以是电载波信号和电信信号。

[0147] 在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中没有详述或记载的部分,可以参见其它实施例的相关描述。

[0148] 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。

[0149] 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,所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。

[0150] 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。

[0151] 以上所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。

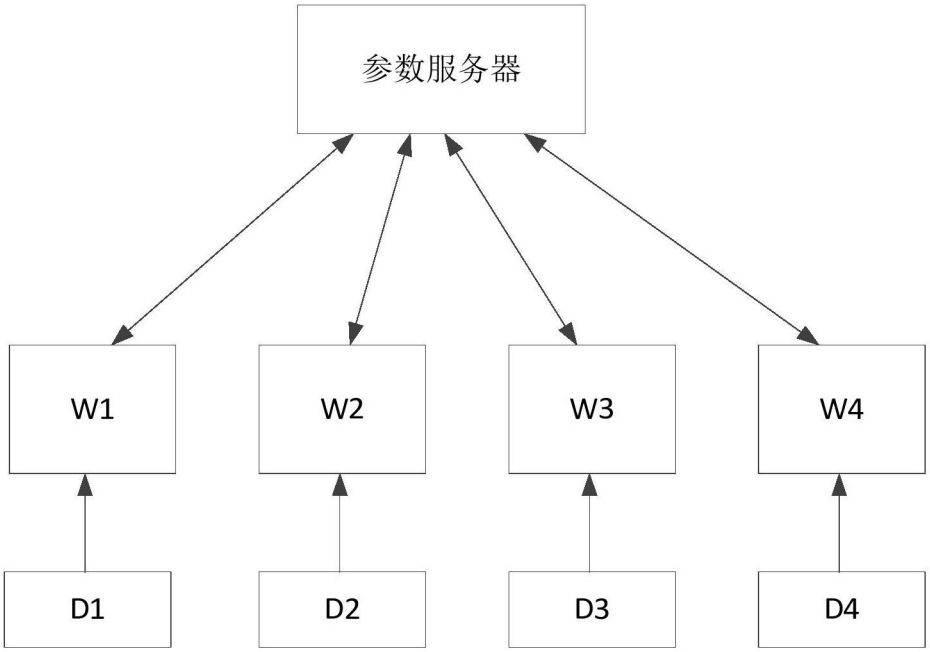


图1

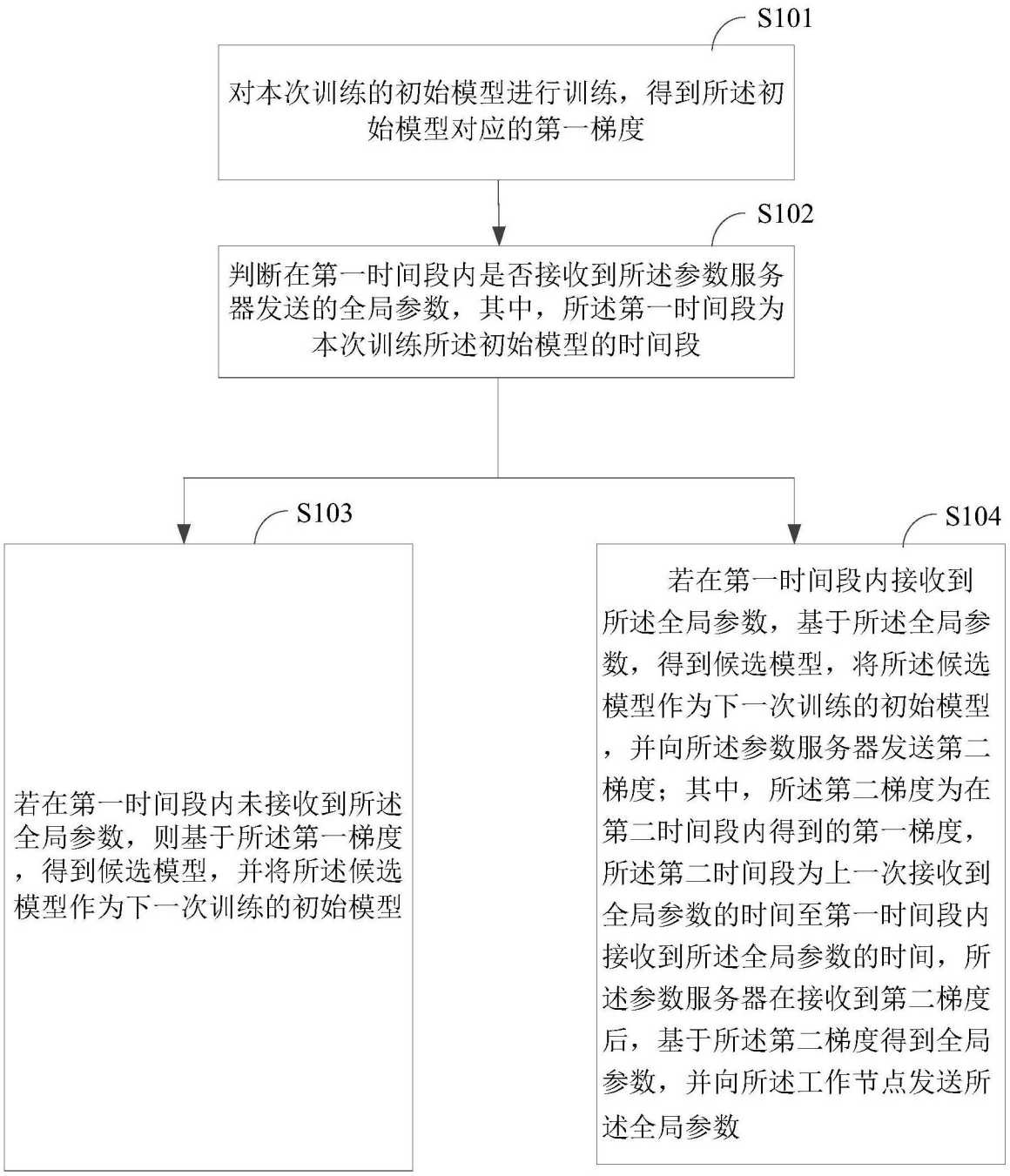

S101
对本次训练的初始模型进行训练，得到所述初始模型对应的第一梯度
S102
判断在第一时间段内是否接收到所述参数服务器发送的全局参数，其中，所述第一时间段为本次训练所述初始模型的时间段
S103
若在第一时间段内未接收到所述全局参数，则基于所述第一梯度，得到候选模型，并将所述候选模型作为下一次训练的初始模型
S104
若在第一时间段内接收到所述全局参数，基于所述全局参数，得到候选模型，将所述候选模型作为下一次训练的初始模型，并向所述参数服务器发送第二梯度；其中，所述第二梯度为在第二时间段内得到的第一梯度，所述第二时间段为上一次接收到全局参数的时间至第一时间段内接收到所述全局参数的时间，所述参数服务器在接收到第二梯度后，基于所述第二梯度得到全局参数，并向所述工作节点发送所述全局参数

图2

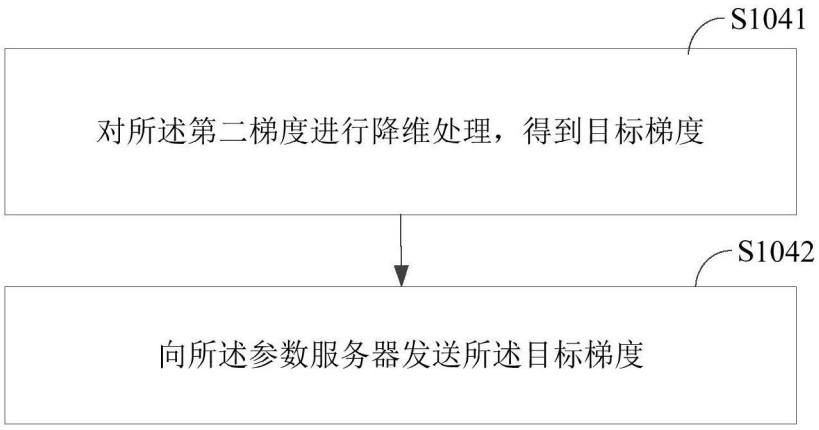

S1041
对所述第二梯度进行降维处理，得到目标梯度
S1042
向所述参数服务器发送所述目标梯度

图3

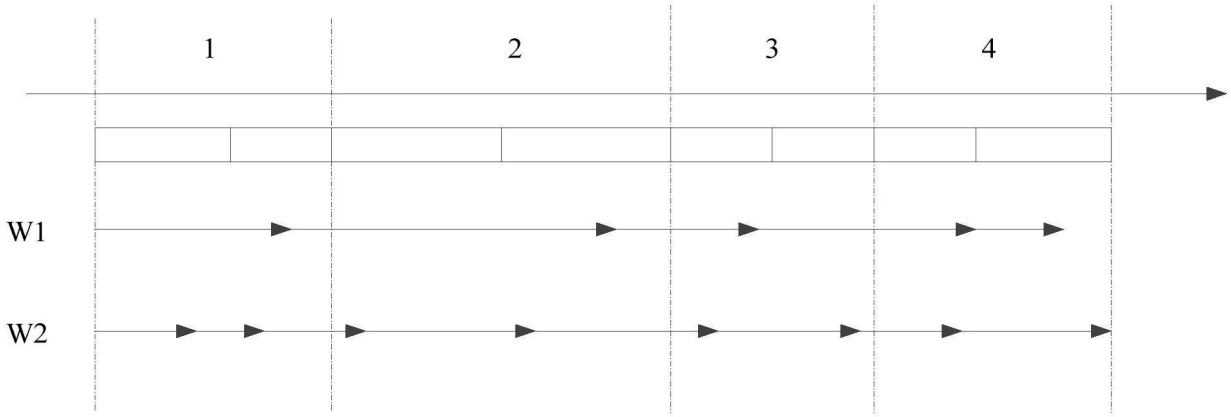

1
2
3
4
W1
W2

图4

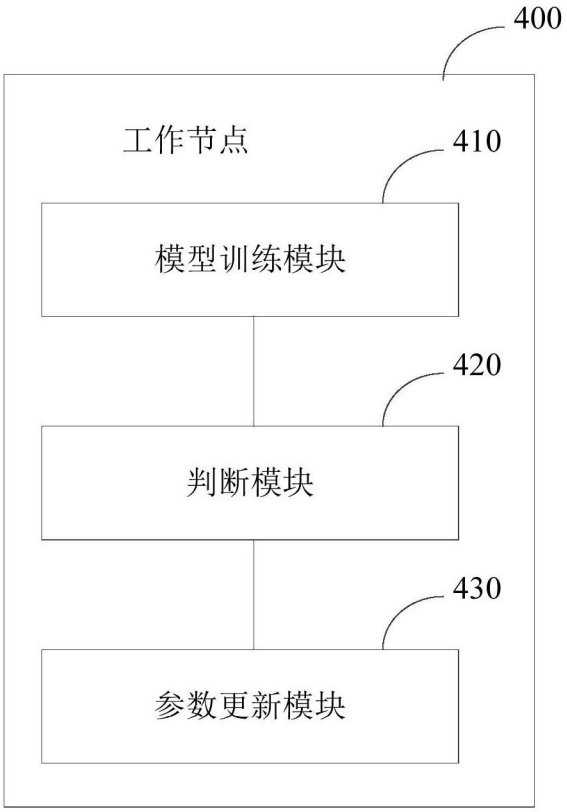

400
工作节点
410
模型训练模块
420
判断模块
430
参数更新模块

图5

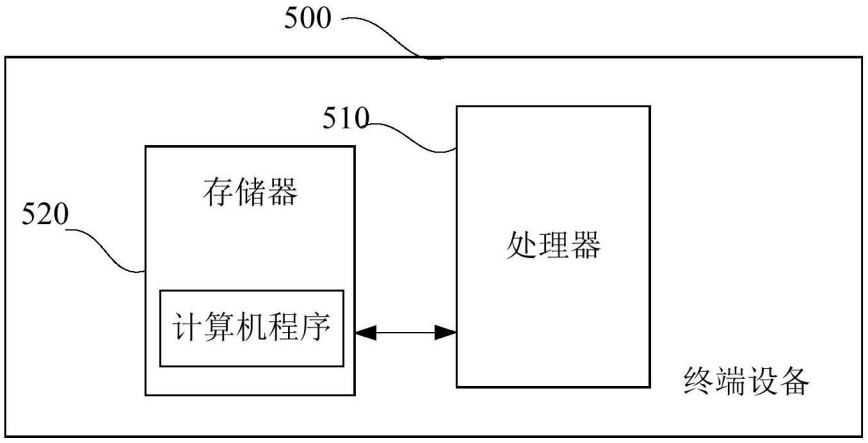

500
510
存储器
520
处理器
计算机程序
终端设备

图6

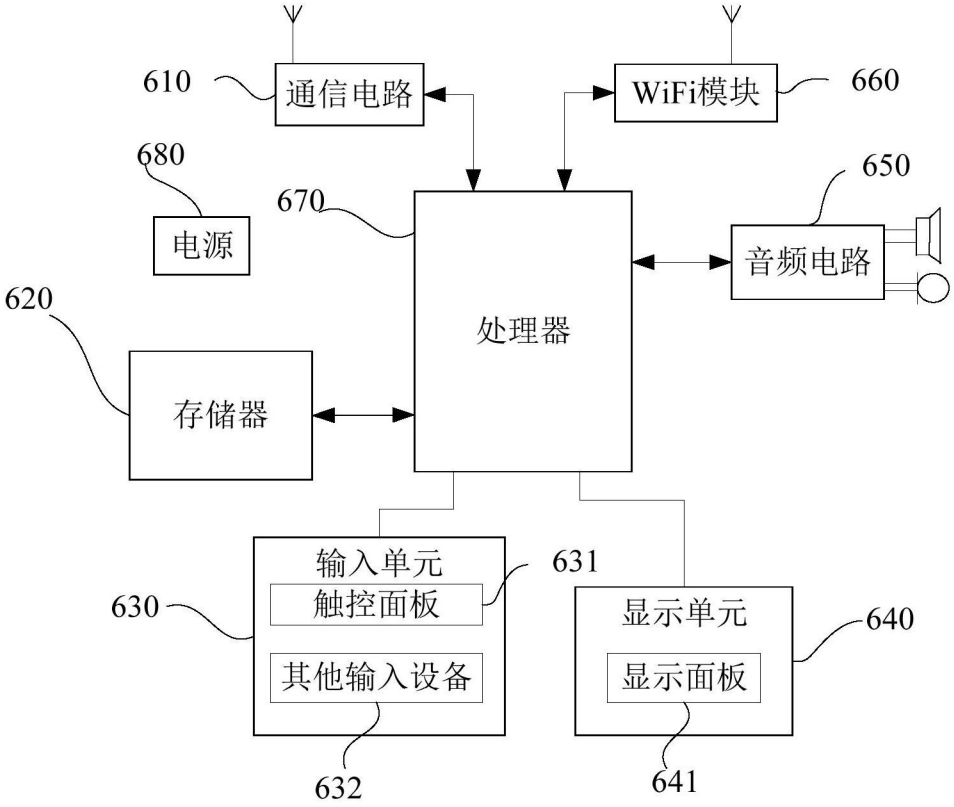

610
通信电路
WiFi模块
660
680
电源
670
650
音频电路
620
处理器
存储器
输入单元
631
630
触控面板
显示单元
640
其他输入设备
显示面板
632
641

图7