(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111401552 B
(45)授权公告日 2023.04.07

(21)申请号 202010166667.4

(22)申请日 2020.03.11

(65)同一申请的已公布的文献号
申请公布号 CN 111401552 A

(43)申请公布日 2020.07.10

(73)专利权人 浙江大学
地址 310013 浙江省杭州市西湖区余杭塘路866号

(72)发明人 刘胜利 余官定 殷锐 袁建涛

(74)专利代理机构 杭州天勤知识产权代理有限公司 33224
专利代理师 曹兆霞

(51)Int.Cl.
G06N 3/098(2023.01)
G06N 3/045(2023.01)

(56)对比文件
CN 109934802 A,2019.06.25
CN 110633805 A,2019.12.31
CN 110782042 A,2020.02.11
CN 110874484 A,2020.03.10
US 2015063651 A1,2015.03.05
US 2019227980 A1,2019.07.25
刘俊旭.机器学习的隐私保护研究综述.《计算机研究与发展》.2020,346-362.
Haijian Sun .Adaptive Federated Learning With Gradient Compression in Uplink NOMA.《arXiv》.2020,1-10.

审查员 罗彩珠

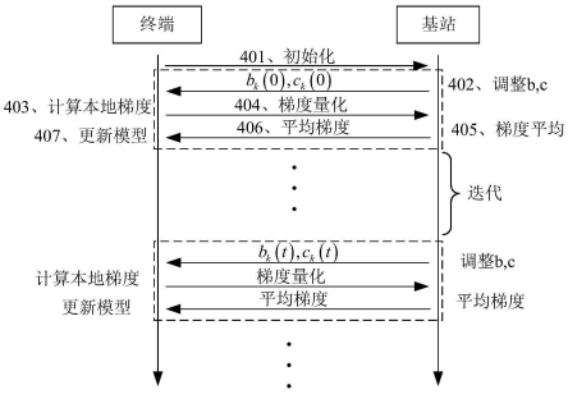
权利要求书3页 说明书8页 附图4页

(54)发明名称

基于调整批量大小与梯度压缩率的联邦学习方法和系统

(57)摘要

本发明公开了一种基于调整批量大小与梯度压缩率的联邦学习方法和系统,用于提高模型训练性能,包括:在联邦学习场景中,多个终端共享上行无线信道资源,基于本地终端的训练数据,与边缘服务器共同完成神经网络模型的训练;在模型训练过程中,终端在本地计算中采用批量的方法计算梯度,在上行传输过程中,传输前需要对梯度进行压缩;根据各终端的计算能力与其所处的信道状态,调整批量大小以及梯度压缩率,以在保证训练时间与不降低模型正确率的同时,提高模型训练的收敛速率。



CN 111401552 B

1.一种基于调整批量大小与梯度压缩率的联邦学习方法,实现所述联邦学习的系统包括边缘服务器、与所述边缘服务器无线通信的多个终端,所述终端根据本地数据进行模型学习,其特征在于,所述联邦学习方法包括:

所述边缘服务器根据当前批量大小和梯度压缩率,并结合终端的计算能力和边缘服务器与终端之间的通信能力调整终端的批量大小与梯度压缩率,包括:

(a)根据当前批量大小和梯度压缩率,并结合终端的计算能力和边缘服务器与终端之间的通信能力计算当前学习时延;

(b)比较所述当前学习时延与规定学习时延,当前学习时延大于规定学习时延时,减小批量大小和增大梯度压缩率;当前学习时延小于规定学习时延时,增大批量大小和减小梯度压缩率;

(c)重复步骤(a)和步骤(b),直到当前学习时延等于规定学习时延为止,与规定学习时延相等的当前学习时延对应的批量大小与梯度压缩率即为调整后的批量大小与梯度压缩率;

所述边缘服务器将调整后的批量大小与梯度压缩率传输至终端;

所述终端按照接收的批量大小进行模型学习,并将模型学习获得的梯度信息按照接收的提取压缩率压缩后输出至边缘服务器;

所述边缘服务器对接收的所有梯度信息求平均后,将梯度平均值同步到终端;

所述终端根据接收的梯度平均值更新模型。

2.如权利要求1所述的基于调整批量大小与梯度压缩率的联邦学习方法,其特征在于,步骤(a)中,当前学习时延的计算过程为:

根据终端学习能力和批量大小计算终端以当前批量大小计算梯度时所需要的时间;

根据通信能力和梯度压缩率计算梯度信息经过压缩后经过无线信道上传至边缘服务器所经历的时间;

计算所有梯度信息汇总后进行平均获得平均梯度信息所需要的时间;

计算边缘服务器将平均梯度信息下发至各个终端所需要经历的时间;

计算各终端收到平均梯度信息后对模型进行更新所需要的时间;

这五部分时间之和即为当前学习时延。

3.如权利要求1所述的基于调整批量大小与梯度压缩率的联邦学习方法,其特征在于,所述梯度压缩率包括对梯度信息进行梯度量化获得的压缩率,表示为:

$$c=\frac{\left\|Q(x)-x\right\|_2^2}{\left\|x\right\|_2^2}$$

其中,x表示梯度信息,Q(x)表示量化函数,c表示压缩率,$\left\|\cdot\right\|_2^2$表示二范数的平方;

对应模型训练的收敛速率表示为:

$$\gamma=\frac{\alpha_1+\beta_1 C}{\sqrt{BS}}$$

其中,α_1,β_1是与模型相关的系数,且$\alpha_1>0$,$\beta_1>0$,B是各终端批量大小的总和,C是各终端梯度压缩率的平均值,S是训练步数。

4.如权利要求1所述的基于调整批量大小与梯度压缩率的联邦学习方法,其特征在于,所述梯度压缩率包括对梯度信息进行稀疏化处理获得的压缩率,即选择梯度矩阵中的最大m个梯度进行传输,此时将梯度的传输量定义为压缩率,表示为:

$$c=\frac{M}{m}$$

其中,M表示模型的大小;

对应模型训练的收敛速率表示为:

$$\gamma=\alpha_2\frac{1}{\sqrt{BS}}+\beta_2\frac{C^3-C}{S}$$

其中,α_2,β_2是与模型相关的系数,且$\alpha_2>0$,$\beta_2>0$,B是各终端批量大小的总和,C是各终端梯度压缩率的平均值,S是训练步数。

5.如权利要求3所述的基于调整批量大小与梯度压缩率的联邦学习方法,其特征在于,当梯度压缩率采用梯度量化方式获得的压缩率时,通过增大或减小量化函数采用的量化位数来增大或减小梯度压缩率。

6.如权利要求4所述的基于调整批量大小与梯度压缩率的联邦学习方法,其特征在于,当梯度压缩率采用梯度稀疏化方式获得的压缩率时,通过增大或减小梯度个数来增大或减小梯度压缩率。

7.如权利要求1所述的基于调整批量大小与梯度压缩率的联邦学习方法,其特征在于,采用以下公式获得梯度平均值:

$$G(w)=\frac{1}{K}\sum_{k=1}^{K}g_k(w)$$

其中,$g_k(w)$表示终端计算的梯度,G(w)表示梯度平均值,w表示模型参数,k表示终端索引,K表示终端的总个数。

8.如权利要求1所述的基于调整批量大小与梯度压缩率的联邦学习方法,其特征在于,采用以下公式根据梯度平均值更新模型:

$$w_{t+1}=w_t-\alpha G(w_t)$$

其中,t是迭代次数,$G(w_t)$表示第t次迭代时梯度平均值,w_t、w_{t+1}分别表示第t次和第t+1次迭代时终端的模型参数。

9.一种基于调整批量大小与梯度压缩率的联邦学习系统,包括连接到基通信端的边缘服务器,与所述边缘服务器无线通信的多个终端,其特征在于,

所述边缘服务器根据当前批量大小和梯度压缩率,并结合终端的计算能力和边缘服务器与终端之间的通信能力调整终端的批量大小与梯度压缩率,包括:(a)根据当前批量大小和梯度压缩率,并结合终端的计算能力和边缘服务器与终端之间的通信能力计算当前学习时延;

(b)比较所述当前学习时延与规定学习时延,当前学习时延大于规定学习时延时,减小批量大小和增大梯度压缩率;当前学习时延小于规定学习时延时,增大批量大小和减小梯度压缩率;

(c)重复步骤(a)和步骤(b),直到当前学习时延等于规定学习时延为止,与规定学习时延相等的当前学习时延对应的批量大小与梯度压缩率即为调整后的批量大小与梯度压缩

率；

所述边缘服务器将调整后的批量大小与梯度压缩率传输至终端；

所述终端按照接收的批量大小进行模型学习，并将模型学习获得的梯度信息按照接收的提取压缩率压缩后输出至边缘服务器；

所述边缘服务器对接收的所有梯度信息求平均后，将梯度平均值同步到终端；

所述终端根据接收的梯度平均值更新模型。

基于调整批量大小与梯度压缩率的联邦学习方法和系统

技术领域

[0001] 本发明涉及人工智能与通信领域，具体涉及一种基于调整批量大小与梯度压缩率的联邦学习方法和系统。

背景技术

[0002] 近年来，随着硬件及软件水平的不断提升，人工智能(Artificial Intelligence，AI)技术又迎来了发展的高峰期。它从海量的数据中挖掘出关键信息以实现各种应用，如人脸识别，语音识别，数据挖掘等。然而对于数据隐私性比较敏感的场景，如医院的病人信息，银行的客户信息等，数据通常很难获取，俗称信息孤岛。如果仍然采用现有的人工智能训练方法，由于没有足够的数据，很难得到有效的结果。

[0003] 由谷歌提出的联邦学习(Federated Learning，FL)就是用来解决信息孤岛问题，主要面向数据隐私性和安全性比较敏感的业务，以及将要到来的自动驾驶，物联网等场景。它将模型训练分散至若干个终端进行，终端不需要将原始数据发送至边缘服务器，取而代之的是模型的参数或者梯度信息。基于传统的随机梯度下降的训练方法，运用在联邦学习场景的平均梯度方法仍然可以获得很好的学习性能。在5G的高可靠低时延的无线通信场景中，自动驾驶的实现，物联网的智能分析、决策等都离不开联邦学习。

[0004] 在传统的联邦学习场景中，由于终端与服务器之间采用有线连接，通信开销与本地计算时延都可以忽略不计。但随着移动通信网络的发展，以及移动智能设备的快速增长，为了快速实现物联网应用以及自动驾驶，人工智能模型训练可以放置在移动智能终端处，传统的有线通信也可以改为无线通信，使得训练终端的加入与退出变得非常便捷，即将联邦学习与无线通信网络相结合是未来的发展方向。

[0005] 但将无线通信应用到联邦学习的场景，将会出现许多问题。首先，本地计算时延增加。虽然本地终端的计算能力在不断增加，已经可以部署人工智能模型，但是其与台式机或者服务器的差距仍然是比较大的，本地梯度计算带来的计算时延是不可以被忽略的。另外一方面，由于无线通信带宽资源紧缺，并且无线信道不稳定，大量的模型梯度信息传输会带来巨大的通信开销，引起很大的传输时延。

[0006] 为了解决本地计算与梯度传输带来较高的训练时延的问题，本地批量数据处理与梯度压缩技术在模型训练过程中得到应用。每一轮训练交互中，终端可以只根据一部分的数据计算模型的梯度，以减少本地计算产生的时延，同时在传输的过程中，对梯度信息进行压缩，以较少的数据量表示原有的梯度信息，减少通信传输时间。但批量处理方式与梯度压缩都会对模型的收敛速率产生影响。

[0007] 因此，在控制模型训练时延的同时，也需要考虑模型的收敛速率。如何采用合理的方式调整批量大小与梯度压缩率来保证训练时延、提高收敛速率是急需解决的问题。

发明内容

[0008] 本发明的目的是提供一种基于调整批量大小与梯度压缩率的联邦学习方法和系

统，该联邦学习方法通过调整批量大小与梯度压缩率，在规定学习时间内提升了模型的收敛速率，且在联邦学习时不需要传输原始数据，更好地保护用户的隐私性和安全性。

[0009] 为实现上述发明目的本发明提供以下技术方案：

[0010] 第一方面，提供了一种基于调整批量大小与梯度压缩率的联邦学习方法，实现所述联邦学习的系统包括边缘服务器、与所述边缘服务器无线通信的多个终端，所述终端根据本地数据进行模型学习，所述联邦学习方法包括：

[0011] 所述边缘服务器根据当前批量大小和梯度压缩率，并结合终端的计算能力和边缘服务器与终端之间的通信能力调整终端的批量大小与梯度压缩率，并将调整后的批量大小与梯度压缩率传输至终端；

[0012] 所述终端按照接收的批量大小进行模型学习，并将模型学习获得的梯度信息按照接收的提取压缩率压缩后输出至边缘服务器；

[0013] 所述边缘服务器对接收的所有梯度信息求平均后，将梯度平均值同步到终端；

[0014] 所述终端根据接收的梯度平均值更新模型。

[0015] 本发明中，为了减少终端计算时间以及设备硬件的需求，本地终端采用批量的方式计算梯度，即每次计算选择一定的批量大小进行梯度计算。为了减少终端至边缘服务器的传输信息，节省通信开销，降低通信时间，终端在上传梯度信息前，需要对梯度信息进行压缩。

[0016] 在一个可能的实现方式中，所述根据当前批量大小和梯度压缩率，并结合终端的计算能力和边缘服务器与终端之间的通信能力调整终端的批量大小与梯度压缩率包括：

[0017] (a)根据当前批量大小和梯度压缩率，并结合终端的计算能力和边缘服务器与终端之间的通信能力计算当前学习时延；

[0018] (b)比较所述当前学习时延与规定学习时延，当前学习时延大于规定学习时延时，减小批量大小和增大梯度压缩率；当前学习时延小于规定学习时延时，增大批量大小和减小梯度压缩率；

[0019] (c)重复步骤(a)和步骤(b)，直到当前学习时延等于规定学习时延为止，与规定学习时延相等的当前学习时延对应的批量大小与梯度压缩率即为调整后的批量大小与梯度压缩率。

[0020] 当前学习时延高于规定学习时间，即当前训练不能够按时完成，此时应该减小批量大小以及增大压缩率，以节约本地终端计算和无线传输的时间；当前学习时延小于规定学习时间，即当前训练可以完成，但收敛速率还可以进一步提升，此时可以适当增加批量大小以及减小梯度的压缩率，以提高收敛性能。

[0021] 特别地，当学习时间特别短时，即训练任务非常紧张，各终端批量大小应尽可能地小，压缩率应该尽可能地大，以保证学习时延；当学习时间特别长，即该任务没有严格的学习时间要求，各终端批量大小应该尽可能地大，压缩率应该尽可能地小，以保证收敛性能最好。

[0022] 步骤(a)中，当前学习时延的计算过程为：

[0023] 根据终端学习能力和批量大小计算终端以当前批量大小计算梯度时所需要的时间；

[0024] 根据通信能力和梯度压缩率计算梯度信息经过压缩后经过无线信道上传至边缘

服务器所经历的时间；

[0025] 计算所有梯度信息汇总后进行平均获得平均梯度信息所需要的时间；

[0026] 计算边缘服务器将平均梯度信息下发至各个终端所需要经历的时间；

[0027] 计算各终端收到平均梯度信息后对模型进行更新所需要的时间；

[0028] 这五部分时间之和即为当前学习时延。

[0029] 根据当前学习时延的计算过程可得，当前学习时延可以通过调整批量大小与梯度压缩率进行改变，当批量大小增大或者梯度压缩率减小时，当前学习时延增加，当批量大小减小或者梯度压缩率增大时，当前学习时延减小。

[0030] 在一种可能的实现方式中，对梯度信息进行压缩时，可以将梯度用量化损失来进行量化，此时量化损失被定义为压缩率，即所述梯度压缩率包括对梯度信息进行梯度量化获得的压缩率，表示为：

[0031] $$c=\frac{\left\|Q(x)-x\right\|_2^2}{\left\|x\right\|_2^2}$$

[0032] 其中，x表示梯度信息，Q(x)表示量化函数，c表示压缩率，$\left\|\cdot\right\|_2^2$表示二范数的平方。

[0033] 对应模型训练的收敛速率可以表示为：

[0034] $$\gamma=\frac{\alpha_1+\beta_1 C}{\sqrt{BS}}$$

[0035] 其中，α_1，β_1是与模型相关的系数，且$\alpha_1>0$，$\beta_1>0$，B是各终端批量大小的总和，C是各终端梯度压缩率的平均值，S是训练步数。

[0036] 当梯度压缩率采用梯度量化方式获得的压缩率时，通过增大或减小量化函数采用的量化位数来增大或减小梯度压缩率。

[0037] 在一种可能的实现方式中，所述梯度压缩率包括对梯度信息进行稀疏化处理获得的压缩率，即选择梯度矩阵中的最大m个梯度进行传输，此时将梯度的传输量定义为压缩率，表示为：

[0038] $$c=\frac{M}{m}$$

[0039] 其中，M表示模型的大小。

[0040] 对应模型训练的收敛速率可以表示为：

[0041] $$\gamma=\alpha_2\frac{1}{\sqrt{BS}}+\beta_2\frac{C^3-C}{S}$$

[0042] 其中，α_2，β_2是与模型相关的系数，且$\alpha_2>0$，$\beta_2>0$，B是各终端批量大小的总和，C是各终端梯度压缩率的平均值，S是训练步数。

[0043] 当梯度压缩率采用梯度稀疏化方式获得的压缩率时，通过增大或减小梯度个数来增大或减小梯度压缩率。

[0044] 在一个可能的实现方式中，采用以下公式获得梯度平均值：

[0045] $$G(w)=\frac{1}{K}\sum_{k=1}^{K}g_k(w)$$

[0046] 其中，$g_k(w)$表示终端计算的梯度，G(w)表示梯度平均值，w表示模型参数，k表示终端索引，K表示终端的总个数。

[0047] 在一个可能的实现方式中，采用以下公式根据梯度平均值更新模型：

[0048] $w_{t+1}=w_t-\alpha G(w_t)$

[0049] 其中，t是迭代次数，$G(w_t)$表示第t次迭代时梯度平均值，w_t、w_{t+1}分别表示第t次和第t+1次迭代时终端的模型参数。

[0050] 第二方面，提供了一种基于调整批量大小与梯度压缩率的联邦学习系统，包括连接到基通信端的边缘服务器，与所述边缘服务器无线通信的多个终端，

[0051] 所述边缘服务器根据当前批量大小和梯度压缩率，并结合终端的计算能力和边缘服务器与终端之间的通信能力调整终端的批量大小与梯度压缩率，并将调整后的批量大小与梯度压缩率传输至终端；

[0052] 所述终端按照接收的批量大小进行模型学习，并将模型学习获得的梯度信息按照接收的提取压缩率压缩后输出至边缘服务器；

[0053] 所述边缘服务器对接收的所有梯度信息求平均后，将梯度平均值同步到终端；

[0054] 所述终端根据接收的梯度平均值更新模型。

[0055] 从以上技术方案可以看出，本申请实施例提供的基于调整批量大小与梯度压缩率的联邦学习方法和系统具有以下优点：与直接传输数据相比，传输梯度可以充分保护用户数据的隐私性和安全性；对梯度进行压缩可以减缓通信传输的压力，减低时延；根据学习时间的需求动态调整批量大小以及压缩率，可以在保证训练时间以及模型训练正确率的同时，提高模型的收敛速率。

附图说明

[0056] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图做简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动前提下，还可以根据这些附图获得其他附图。

[0057] 图1为本申请实施例中基于无线通信网络的联邦学习系统示意图；

[0058] 图2为本申请实施例中梯度压缩中量化方法示意图；

[0059] 图3为本申请实施例中提供的一个实施例的批量大小和压缩率调整流程图；

[0060] 图4为本申请实施例中提供的一个实施例的整体训练交互示意图；

[0061] 图5为本申请实施例中梯度压缩中稀疏化方法示意图；

[0062] 图6为本申请实施例中提供的另一个实施例的批量大小和压缩率调整流程图；

[0063] 图7为本申请实施例中提供的另一个实施例的整体训练交互示意图。

具体实施方式

[0064] 为使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例对本发明进行进一步的详细说明。应当理解，此处所描述的具体实施方式仅仅用以解释本发明，并不限定本发明的保护范围。

[0065] 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”、“第三”、“第

四”等(如果存在)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。

[0066] 图1为本申请实施例中基于无线通信网络的联邦学习系统,该联邦学习系统包括连接至基站等通信端的边缘服务器和终端,终端共享上行无线信道资源,基于本地终端的训练数据,与边缘服务器共同完成神经网络模型的训练。利用该联邦学习系统可以实现基于调整批量大小与梯度压缩率的联邦学习方法。具体地,对于本地计算引起的计算时延,可以根据终端的计算能力,调整批量大小,以减小计算时延;另一方面,对于通信瓶颈问题,可以通过减少传输的信息量,以减少通信开销,即梯度压缩技术。根据终端所在的无线信道的状态以调整梯度压缩率,从而减小通信时延。为了满足训练时间的需求,本发明中批量大小与压缩率调整方法可以在保证训练时延的同时,提高模型训练的收敛速率。

[0067] 实施例一

[0068] 本实施例提供的基于调整批量大小与梯度压缩率的联邦学习方法适用于多个移动终端与一个连接通信热点(如基站)的边缘服务器共同训练一个人工智能模型的场景,对于其他无线通信技术,可以以相同的工作模式工作,因此在本实施例中,主要考虑移动通信技术的情况。

[0069] 本实施例中,各个终端采用批量的方法进行本地计算,梯度压缩方法为量化,特别的,量化采用固定长度量化,其量化过程如图2,梯度信息经由一定比特量化并编码后,作为梯度信息传输至边缘服务器。当采用高量化位数表示梯度时,可以尽可能地保留原始的梯度信息,同时也增加了传输的信息量;当采用低量化位数对梯度进行量化时,量化后的梯度信息与原始的梯度信息存在偏差,但是传输的信息量减少。此时,梯度的压缩率可以用量化误差表示,即

[0070] $$c=\frac{\left\|Q(x)-x\right\|_2^2}{\left\|x\right\|_2^2}$$

[0071] 其中,Q(x)表示量化函数,x表示梯度。

[0072] 采用梯度量化时,模型训练的收敛速率可以表示为:

[0073] $$\gamma=\frac{\alpha_1+\beta_1 C}{\sqrt{BS}}$$

[0074] 其中,α_1,β_1是与模型相关的系数,且$\alpha_1>0$,$\beta_1>0$,B是各终端批量大小的总和,C是各终端梯度压缩率的平均值,S是训练步数。

[0075] 在本实施例中,最终的目标要在满足训练时延的基础上,根据各个终端的计算能力以及所处无线信道的状态,调整批量大小以及压缩率的大小,以提高模型训练的收敛速率。

[0076] 具体地,批量大小与压缩率的调整算法如图3,包括以下部分:

[0077] 301、计算当前的训练时延,其中,训练时延包含五个部分:

[0078] (1)本地以批量大小b计算梯度时,所需要的时间;

[0079] (2)梯度信息经过压缩后,经过无线信道上传至边缘服务器所经历的时间;

[0080] (3)所以梯度信息汇总后进行平均汇总所需要的时间;

[0081] (4)边缘服务器将计算好的平均梯度信息发送至各个终端所需要经历的时间;

[0082] (5)各终端收到梯度信息后,对模型进行更新的时间;

[0083] 302、由于本实施例的目的是为了在保证训练时间的前提下,提高收敛速率,因此需要将当前的训练时间与规定的训练时间T^{max}相比较,分为三种情况,以做出对应的改变。

[0084] 303、当训练时间小于规定的训练时间时,应该适当增大批量大小与量化的位数,以提高收敛的速率。

[0085] 304、当训练时间大于规定的训练时间时,应该适当减小批量大小与量化的位数,及时完成训练,满足时间要求。

[0086] 305、当训练时间等于规定的训练时间时,此时的批量大小与压缩率可以用于训练。

[0087] 将此调整过程用于联邦学习的实际训练中,终端与基站的具体的交互过程如图4,具体包括如下内容:

[0088] 401、初始化,各个终端需要上传相关的信息,如计算能力,以及无线信道所在的状态等信息至基站。

[0089] 402、基站根据各个终端上传的信息,利用图3中所述的调整方法,得到批量大小b与梯度压缩率c。

[0090] 403、各终端计算本地的梯度信息。

[0091] 404、根据梯度压缩率,即量化误差,得到相应的量化位数,对梯度进行量化编码,并使用无线信道传输。

[0092] 405、基站收到梯度信息,对梯度进行平均,并发送至各终端。

[0093] 406、各终端下载平均梯度信息。

[0094] 407、各终端使用该平均梯度信息对模型进行更新。

[0095] 使用该发明方法,即可以获得最好的模型性能。

[0096] 实施例二

[0097] 本实施例提供的调整方法适用于多个移动终端与一个连接通信热点(如基站)的边缘服务器共同训练一个人工智能模型的场景,对于其他无线通信技术,可以以相同的工作模式工作,因此在本实施例中,主要考虑移动通信的情况。

[0098] 本实施例中,各个终端采用批量的方法进行本地计算,梯度压缩方法为稀疏化,特别的,稀疏化的方式是选择部分较大的梯度进行传输,稀疏化的过程如图5,梯度信息经过稀疏化之后,将选中的梯度信息及其编号传输至边缘服务器。当保留更多的梯度信息时,可以减少梯度信息的损失,同时也增加了传输的信息量;当保留较少的梯度信息时,更多的梯度信息被丢失,但是传输的信息量减少。此时,梯度的压缩率可以表示为模型总量与传输的梯度个数的比值,即

[0099] $$c=\frac{M}{m}$$

[0100] 其中,M表示模型的大小。

[0101] 采用梯度稀疏化时，模型训练的收敛速率可以表示为：

[0102] $$\gamma=\alpha_2\frac{1}{\sqrt{BS}}+\beta_2\frac{C^3-C}{S}$$

[0103] 其中，α_2，β_2是与模型相关的系数，且$\alpha_2>0$，$\beta_2>0$，B是各终端批量大小的总和，C是各终端梯度压缩率的平均值，S是训练步数。

[0104] 在本实施例中，最终的目标要在满足训练时延的基础上，根据各个终端的计算能力以及所处无线信道的状态，调整批量大小以及压缩率的大小，以提高模型训练的收敛速率。

[0105] 具体地，批量大小与压缩率的调整算法如图6，包括以下部分：

[0106] 601、计算当前的训练时延，其中，训练时延包含五个部分：

[0107] (1)本地以批量大小b计算梯度时，所需要的时间；

[0108] (2)梯度信息经过压缩后，经过无线信道上传至边缘服务器所经历的时间；

[0109] (3)所以梯度信息汇总后进行平均汇总所需要的时间；

[0110] (4)边缘服务器将计算好的平均梯度信息发送至各个终端所需要经历的时间；

[0111] (5)各终端收到梯度信息后，对模型进行更新的时间；

[0112] 602、由于本实施例的目的是为了在保证训练时间的前提下，提高收敛速率，因此需要将当前的训练时间与规定的训练时间T^{max}相比较，分为三种情况，以做出对应的改变。

[0113] 603、当训练时间小于规定的训练时间时，应该适当增大批量大小与传输梯度的个数，以提高收敛的速率。

[0114] 604、当训练时间大于规定的训练时间时，应该适当减小批量大小与传输梯度的个数，及时完成训练，满足时间要求。

[0115] 605、当训练时间等于规定的训练时间时，此时的批量大小与压缩率可以用于训练。

[0116] 将此调整过程用于联邦学习的实际训练中，终端与基站的具体的交互过程如图7，具体包括如下内容：

[0117] 701、初始化，各个终端需要上传相关的信息，如计算能力，以及无线信道所在的状态等信息至基站。

[0118] 702、基站根据各个终端上传的信息，利用图6中所述的调整方法，得到批量大小b与梯度压缩率c。

[0119] 703、各终端计算本地的梯度信息。

[0120] 704、根据梯度压缩率，得到应该传输的梯度的个数，对梯度进行排序选择，选出最大的梯度及其位置，并使用无线信道传输。

[0121] 705、基站收到梯度信息，对梯度进行平均，并发送至各终端。

[0122] 706、各终端下载平均梯度信息。

[0123] 707、各终端使用该平均梯度信息对模型进行更新。

[0124] 使用该发明方法，即可以获得最好的模型性能。

[0125] 该联邦学习方法通过调整批量大小与梯度压缩率，在规定学习时间内提升了模型的收敛速率，且在联邦学习时不需要传输原始数据，更好地保护用户的隐私性和安全性。

[0126] 实施例三

[0127] 实施例三提供了一种基于调整批量大小与梯度压缩率的联邦学习系统，包括连接到基通信端的边缘服务器，与所述边缘服务器无线通信的多个终端，

[0128] 所述边缘服务器根据当前批量大小和梯度压缩率，并结合终端的计算能力和边缘服务器与终端之间的通信能力调整终端的批量大小与梯度压缩率，并将调整后的批量大小与梯度压缩率传输至终端；

[0129] 所述终端按照接收的批量大小进行模型学习，并将模型学习获得的梯度信息按照接收的提取压缩率压缩后输出至边缘服务器；

[0130] 所述边缘服务器对接收的所有梯度信息求平均后，将梯度平均值同步到终端；

[0131] 所述终端根据接收的梯度平均值更新模型。

[0132] 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，该联邦学习系统与实施例一和实施例二提供的基于调整批量大小与梯度压缩率的联邦学习方法实现具体过程相同，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0133] 该联邦学习系统通过调整批量大小与梯度压缩率，在规定学习时间内提升了模型的收敛速率，且在联邦学习时不需要传输原始数据，更好地保护用户的隐私性和安全性。

[0134] 其中，上述提到的无线通信方式，可以是现有的移动通信网络，即LTE(Long-term Evolution)或5G网络，或者是WiFi网络。

[0135] 其中，上述提到的边缘服务器的处理器能力远超于本地终端的计算能力，并且具备独立进行模型训练的能力。处理器可以是一个通用的中央处理器(Central Processing Unit,CPU)，图形处理器(Graphics Processing Unit,GPU)，微处理器，特定应用集成电路(Application Specific Integrated Circuit,ASIC)，或一个或多个用于上述的模型训练的程序执行的集成电路。

[0136] 其中，上述提到的本地终端，可以是现代智能手机、平板电脑、笔记本、自动驾驶汽车等可以支撑模型训练的移动终端，配有无线通信系统，可以接入移动通信网络、WiFi等主流的无线通信网络。

[0137] 以上所述的具体实施方式对本发明的技术方案和有益效果进行了详细说明，应理解的是以上所述仅为本发明的最优选实施例，并不用于限制本发明，凡在本发明的原则范围内所做的任何修改、补充和等同替换等，均应包含在本发明的保护范围之内。

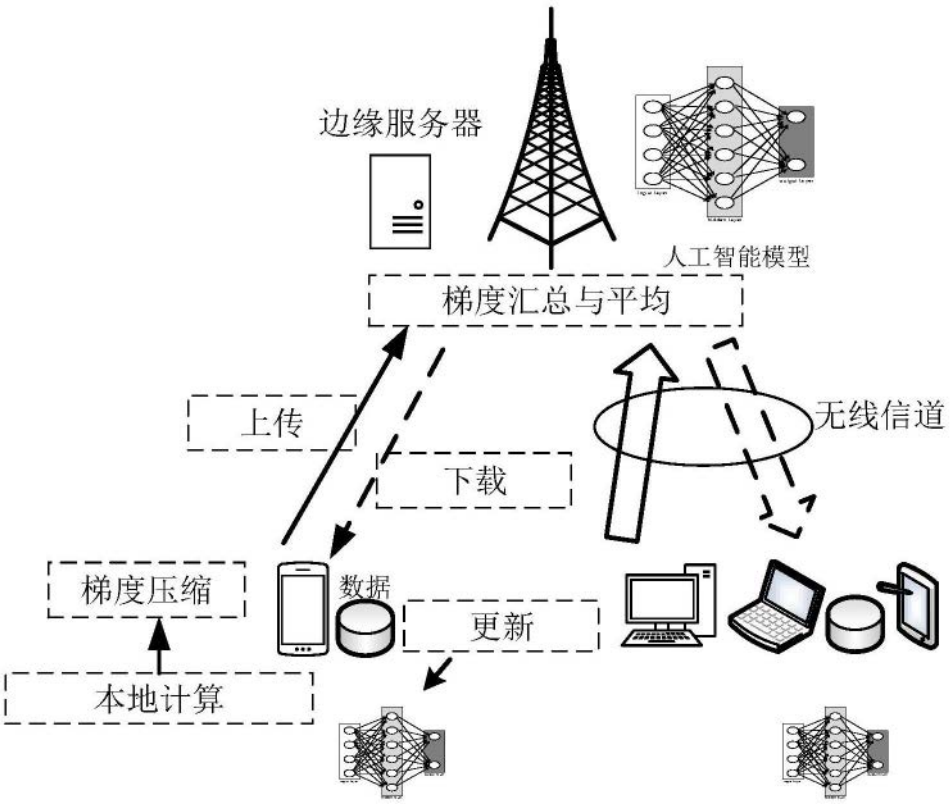


图1

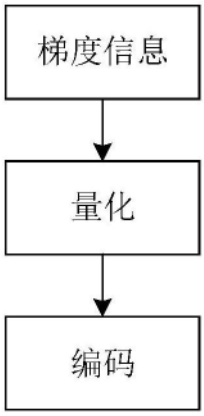


图2

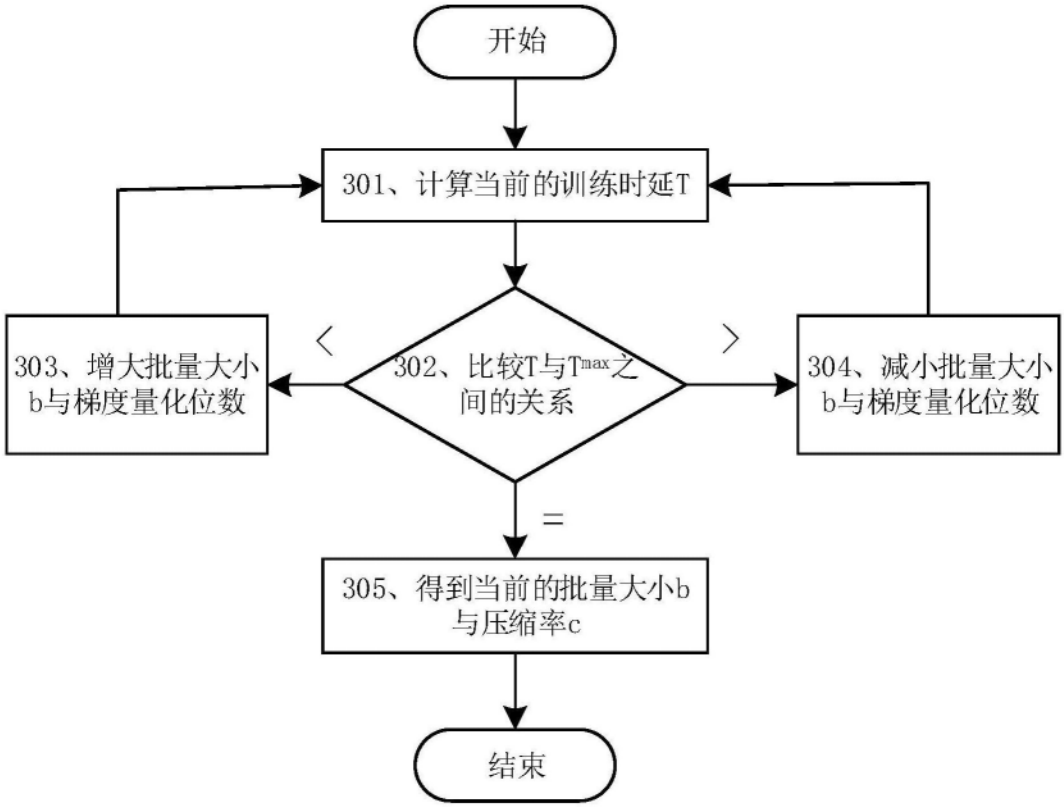

开始
301、计算当前的训练时延T
302、比较T与Tmax之间的关系
<
>
=
303、增大批量大小b与梯度量化位数
304、减小批量大小b与梯度量化位数
305、得到当前的批量大小b与压缩率c
结束

图3

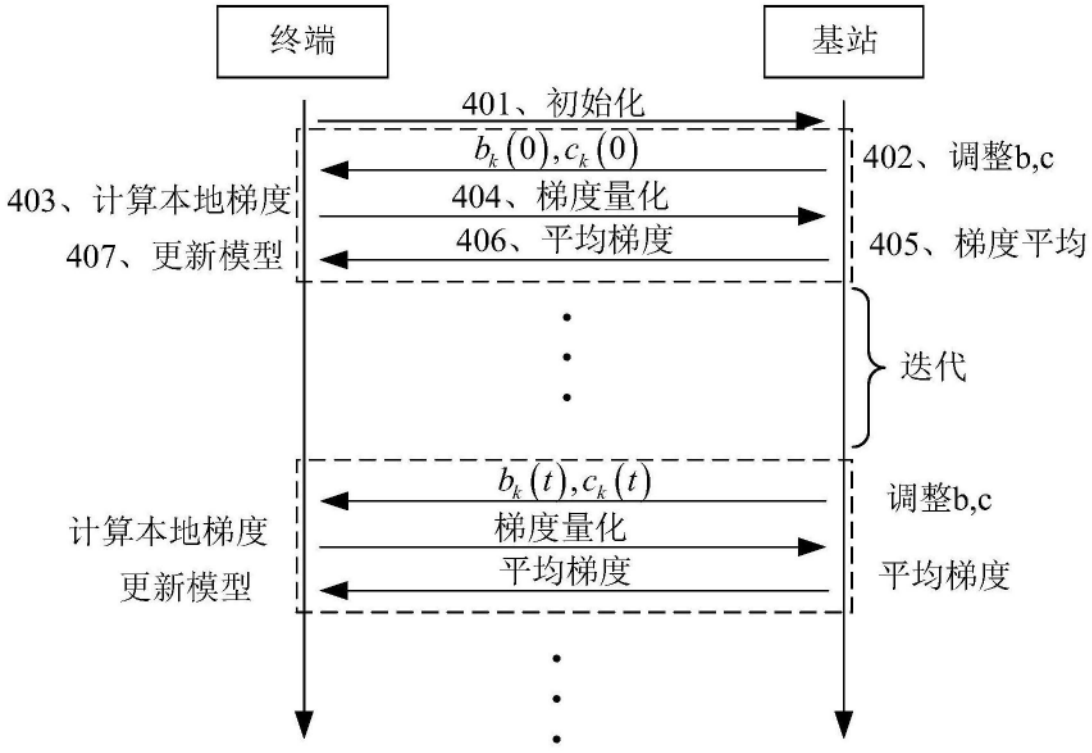

终端
基站
401、初始化
$b_k(0), c_k(0)$
402、调整b,c
403、计算本地梯度
404、梯度量化
406、平均梯度
405、梯度平均
407、更新模型
迭代
$b_k(t), c_k(t)$
调整b,c
计算本地梯度
梯度量化
平均梯度
更新模型
平均梯度

图4

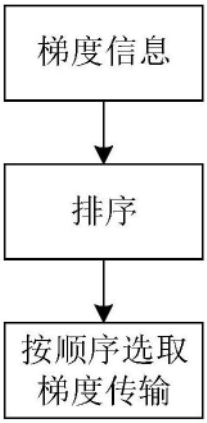

梯度信息
排序
按顺序选取梯度传输

图5

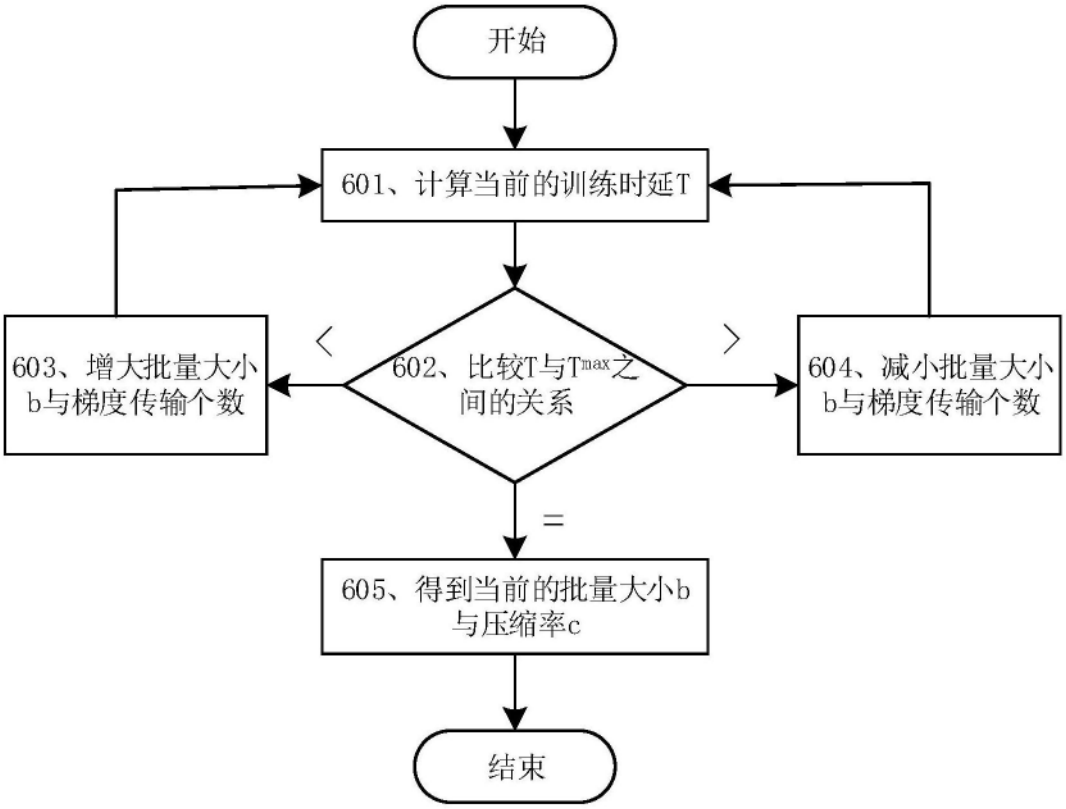

开始
601、计算当前的训练时延T
602、比较T与Tmax之间的关系
<
>
=
603、增大批量大小b与梯度传输个数
604、减小批量大小b与梯度传输个数
605、得到当前的批量大小b与压缩率c
结束

图6

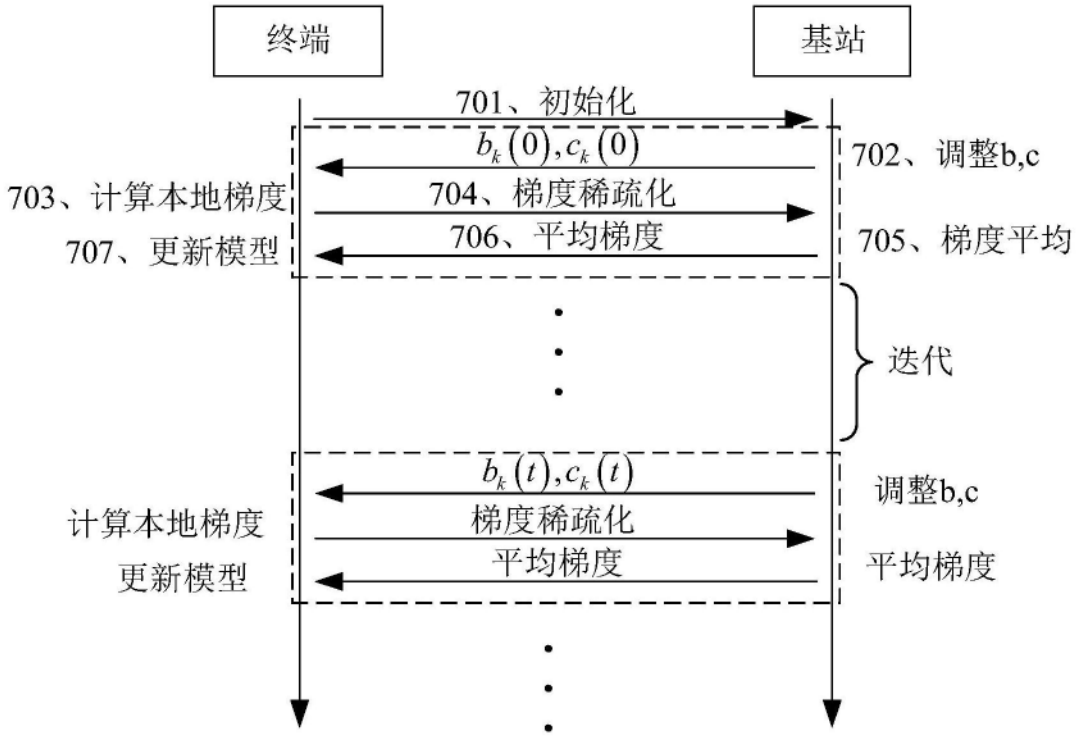

终端
基站
701、初始化
$b_k(0), c_k(0)$
702、调整b,c
703、计算本地梯度
704、梯度稀疏化
706、平均梯度
705、梯度平均
707、更新模型
迭代
$b_k(t), c_k(t)$
调整b,c
计算本地梯度
梯度稀疏化
平均梯度
更新模型
平均梯度

图7