

用于处理模型参数的方法和装置

5 相关申请的交叉引用

本申请要求于 2020 年 04 月 08 日提交的、申请号为 202010269954.8、发明名称为“用于处理模型参数的方法和装置”的中国专利申请的优先权，该申请的全文通过引用结合在本申请中。

10 技术领域

本申请实施例涉及计算机技术领域，具体涉及用于处理模型参数的方法和装置。

背景技术

15 随着机器学习技术的飞速发展，模型的种类越来越丰富，同时随着模型结构越来越复杂，模型的体量也越来越大。

在机器学习技术中，通常需要将各种信息（例如文档、视频、用户属性信息等）进行特征表示（Embedding）。在涉及大规模的特征表示的训练时，相关的方式通常是将大规模的特征表示分布在多个参数服务器（Parameter Server）上进行训练，在训练完成后重新组合成完整的模型。当模型进行推断（Inference）时，往往需要加载完整的模型。

发明内容

25 本申请实施例提出了用于处理模型参数的方法和装置。

第一方面，本申请实施例提供了一种用于处理模型参数的方法，该方法包括：获取待分片的模型参数集合，其中，模型参数集合包括特征表示对应的多维数组；获取用于存储待分片的模型参数集合的存储系统的属性信息，其中，用于存储待分片的模型参数集合的存储系统有别于待分片的模型参数集合对应的模型运行时所在的系统；根据属性信息，将待分片的模型参数集合存入存储系统。

5 在一些实施例中，上述获取待分片的模型参数集合，包括：获取待分片的模型参数集合对应的模型文件；对模型文件进行解析，生成对应的解析图，其中，解析图中包括变量和对变量的操作；根据预设的字段，从解析图中提取变量作为目标变量；根据目标变量对应的特征表示，获取待分片的模型参数集合。

10 在一些实施例中，上述根据目标变量对应的特征表示，获取待分片的模型参数集合，包括：获取根据目标变量对应的特征表示对应的多维数组的数据量；响应于确定获取的数据量大于预设数据量阈值，获取根据目标变量对应的特征表示对应的多维数组作为待分片的模型参数集合。

在一些实施例中，该方法还包括：接收待分片的模型参数集合对应的特征表示查询操作；根据特征表示查询操作，在存储系统中查询与特征表示查询操作匹配的结果；发送匹配的结果。

15 在一些实施例中，上述属性信息包括存储系统的存储空间分布；以及上述根据属性信息，将待分片的模型参数集合存入存储系统，包括：根据属性信息，生成模型参数分配表，其中，模型参数分配表用于表征待分片的模型参数集合与存储系统的存储位置之间的对应关系；根据模型参数分配表，将待分片模型参数集合存入存储系统对应的存储位置，以及生成查询接口。

20 在一些实施例中，上述属性信息包括键值(key-value, KV)存储系统的标识；以及上述根据属性信息，将待分片的模型参数集合存入存储系统，包括：获取前次更新的模型参数集合作为参考模型参数集合；确定待分片的模型参数集合与参考模型参数集合之间的差异，生成更新模型参数集合；根据属性信息，将更新模型参数集合存入存储系统。

25 在一些实施例中，该方法还包括：接收待分片的模型参数集合对应的特征表示查询操作；根据特征表示查询操作，生成相应数量的远程调用请求；向键值存储系统发送远程调用（Remote Procedure Call, RPC）请求；响应于接收到键值存储系统发送的包括与特征表示查询操作匹配的结果的响应信息，发送匹配的结果。

30 第二方面，本申请实施例提供了一种用于处理模型参数的装置，

该装置包括：第一获取单元，被配置成获取待分片的模型参数集合，其中，模型参数集合包括特征表示对应的多维数组；第二获取单元，被配置成获取用于存储待分片的模型参数集合的存储系统的属性信息，其中，用于存储待分片的模型参数集合的存储系统有别于待分片的模型参数集合对应的模型运行时所在的系统；存储单元，被配置成根据属性信息，将待分片的模型参数集合存入存储系统。

在一些实施例中，上述第一获取单元包括：第一获取子单元，被配置成获取待分片的模型参数集合对应的模型文件；解析子单元，被配置成对模型文件进行解析，生成对应的解析图，其中，解析图中包括变量和对变量的操作；提取子单元，被配置成根据预设的字段，从解析图中提取变量作为目标变量；第二获取子单元，被配置成根据目标变量对应的特征表示，获取待分片的模型参数集合。

在一些实施例中，上述第二获取子单元包括：第一获取模块，被配置成获取根据目标变量对应的特征表示对应的多维数组的数据量；第二获取模块，被配置成响应于确定获取的数据量大于预设数据量阈值，获取根据目标变量对应的特征表示对应的多维数组作为待分片的模型参数集合。

在一些实施例中，该装置还包括：第一接收单元，被配置成接收待分片的模型参数集合对应的特征表示查询操作；第一查询单元，被配置成根据特征表示查询操作，在存储系统中查询与特征表示查询操作匹配的结果；第一发送单元，被配置成发送匹配的结果。

在一些实施例中，上述属性信息包括存储系统的存储空间分布；以及上述存储单元包括：第一生成子单元，被配置成根据属性信息，生成模型参数分配表，其中，模型参数分配表用于表征待分片的模型参数集合与存储系统的存储位置之间的对应关系；第一存储子单元，被配置成根据模型参数分配表，将待分片模型参数集合存入存储系统对应的存储位置，以及生成查询接口。

在一些实施例中，上述属性信息包括键值存储系统的标识；以及上述存储单元包括：第三获取子单元，被配置成获取前次更新的模型参数集合作为参考模型参数集合；第二生成子单元，被配置成确定待

分片的模型参数集合与参考模型参数集合之间的差异，生成更新模型参数集合；第二存储子单元，被配置成根据属性信息，将更新模型参数集合存入存储系统。

5 在一些实施例中，该装置还包括第二接收单元，被配置成接收待分片的模型参数集合对应的特征表示查询操作；生成单元，被配置成根据特征表示查询操作，生成相应数量的远程调用请求；第二发送单元，被配置成向键值存储系统发送远程调用请求；第三发送单元，被配置成响应于接收到键值存储系统发送的包括与特征表示查询操作匹配的结果的响应信息，发送匹配的结果。

10 第三方面，本申请实施例提供了一种服务器，该服务器包括：一个或多个处理器；存储装置，其上存储有一个或多个程序；当一个或多个程序被一个或多个处理器执行，使得一个或多个处理器实现如第一方面中任一实现方式描述的方法。

15 第四方面，本申请实施例提供了一种计算机可读介质，其上存储有计算机程序，该程序被处理器执行时实现如第一方面中任一实现方式描述的方法。

第五方面，本申请实施例提供了一种计算机程序产品，包括计算机程序，该计算机程序在被处理器执行时实现如第一方面中任一实现方式描述的方法。

20 第六方面，本申请实施例提供了一种计算机程序，该计算机程序在被处理器执行时实现如第一方面中任一实现方式描述的方法。

本申请实施例提供的用于处理模型参数的方法和装置，通过获取待分片的模型参数集合。其中，模型参数集合包括特征表示对应的多维数组。而后，获取用于存储待分片的模型参数集合的存储系统的属性信息。其中，用于存储待分片的模型参数集合的存储系统有别于待分片的模型参数集合对应的模型运行时所在的系统。最后，根据属性信息，将待分片的模型参数集合存入存储系统。从而实现了将占据较大空间的模型参数集合与模型运行所在的系统相分离，进而优化了大规模特征表示的参数存储方式，提升了资源利用效率。

30

附图说明

通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本申请的其它特征、目的和优点将会变得更明显：

图 1 是本申请的一个实施例可以应用于其中的示例性系统架构图；

图 2 是根据本申请的用于处理模型参数的方法的一个实施例的流程图；

图 3 是根据本申请的实施例的用于处理模型参数的方法的一个应用场景的示意图；

图 4 是根据本申请的用于处理模型参数的方法的又一个实施例的流程图；

图 5 是根据本申请的用于处理模型参数的装置的一个实施例的结构示意图；

图 6 是适于用来实现本申请的实施例的电子设备的结构示意图。

具体实施方式

下面结合附图和实施例对本申请作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关发明，而非对该发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。

需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

图 1 示出了可以应用本申请的用于处理模型参数的方法或用于处理模型参数的装置的示例性架构 100。

如图 1 所示，系统架构 100 可以包括终端设备 101、102、103，网络 104、106 和服务器集群 105、服务器 107。网络 104 用以在终端设备 101、102、103 和服务器集群 105 之间提供通信链路的介质。网络 106 用以在服务器集群 105 和服务器 107 之间提供通信链路的介质。网络 104、106 可以包括各种连接类型，例如有线、无线通信链路或者

光纤电缆等等。

终端设备 101、102、103 通过网络 104 与服务器集群 105 交互，以接收或发送消息等。终端设备 101、102、103 上可以安装有各种通讯客户端应用，例如网页浏览器应用、购物类应用、搜索类应用、即时通信工具、机器学习类软件等。

终端设备 101、102、103 可以是硬件，也可以是软件。当终端设备 101、102、103 为硬件时，可以是具有显示屏并且支持机器学习类软件运行的各种电子设备，包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。当终端设备 101、102、103 为软件时，可以安装在上述所列举的电子设备中。其可以实现成多个软件或软件模块（例如用来提供分布式服务的软件或软件模块），也可以实现成单个软件或软件模块。在此不做具体限定。

服务器集群 105 和服务器 107 可以是提供各种服务的服务器，例如为终端设备 101、102、103 上进行的机器学习模型的训练和推断提供支持的后台服务器。在采用分布式训练 Tensorflow 的模型时，服务器集群 105 可以采用 PS（Parameter Server，参数服务器）架构。服务器集群 105 可以对模型进行分布式训练，并将模型参数通过网络 106 发送给服务器 107 进行存储。在模型推断时，服务器集群 105 可以通过网络 106 向服务器 107 发送获取相关模型参数的请求，服务器 107 可以返回相应的结果至服务器集群 105，以使模型完成推断过程。

需要说明的是，服务器可以是硬件，也可以是软件。当服务器为硬件时，可以实现成多个服务器组成的分布式服务器集群，也可以实现成单个服务器。当服务器为软件时，可以实现成多个软件或软件模块（例如用来提供分布式服务的软件或软件模块），也可以实现成单个软件或软件模块。在此不做具体限定。

需要说明的是，本申请实施例所提供的用于处理模型参数的方法一般由服务器 107 执行，相应地，用于处理模型参数的装置一般设置于服务器 107 中。

应该理解，图 1 中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务

继续参考图 2，示出了根据本申请的用于处理模型参数的方法的一个实施例的流程 200。该用于处理模型参数的方法包括以下步骤：

步骤 201，获取待分片的模型参数集合。

5 在本实施例中，用于处理模型参数的方法的执行主体（如图 1 所示的服务器 107）可以通过有线连接方式或者无线连接方式获取待分片的模型参数集合。其中，上述待分片的模型参数集合可以包括各种用于进行分布式存储的模型参数集合，例如特征表示对应的多维数组。上述多维数组可以作为上述各种特征表示的集合。

10 需要说明的是，上述模型参数通常是指模型通过训练进行调整的参数，即有别于模型的“超参数”。作为示例，上述多维数组可以是 Tensorflow 中的“查询表”（lookup table）。

在本实施例中，作为示例，上述执行主体可以获取预先存储于本地的待分片的模型参数集合。作为又一示例，上述执行主体也可以获
15 取与之通信连接的电子设备（例如图 1 所示的服务器集群 105）发送的待分片的模型参数集合。

在本实施例的一些可选的实现方式中，上述待分片的模型参数集合中可以包括联合训练模型的子模型的模型参数。其中，上述联合训练模型可以包括各种在不暴露原始样本数据的前提下由多个参与方之
20 间通过传递模型中间值（例如梯度）共同完成训练的模型。上述联合训练模型的子模型的模型参数可以包括各参与方通过联合训练对本地的模型的参数进行调整而得到的模型参数。

在本实施例的一些可选的实现方式中，上述执行主体还可以通过如下步骤获取待分片的模型参数集合：

25 第一步，获取待分片的模型参数集合对应的模型文件。

在这些实现方式中，上述执行主体可以通过各种方式获取待分片的模型参数集合对应的模型文件。其中，上述模型文件通常可以包括上述待分片的模型参数集合。作为示例，上述模型文件可以是利用 Tensorflow 加载的 checkpoint 文件。实践中，上述执行主体通常可以
30 从通信连接的电子设备（例如图 1 所示的服务器 105 集群）获取上述

模型文件。

第二步，对模型文件进行解析，生成对应的解析图。

在这些实现方式中，上述执行主体可以通过各种方式对上述第一步所获取的模型文件进行解析，生成模型文件对应的解析图。其中，
5 上述解析图中可以包括变量和对变量的操作。作为示例，在 Tensorflow 中，对模型文件解析后的图（graph）中可以包括节点和边。上述解析图中可以显示有原模型文件中的变量（variable）和对变量的操作（operation）。

第三步，根据预设的字段，从解析图中提取变量作为目标变量。

10 在这些实现方式中，上述执行主体可以根据预设的字段，通过各种方式从上述第二步所生成的解析图中提取变量作为目标变量。其中，上述预设的字段可以与模型参数集合相匹配。作为示例，上述模型参数集合可以包括“查询表”，那么上述预设的字段例如可以为“embedding lookup”。从而，上述执行主体可以从解析图中提取与特征
15 表示相关联的变量作为目标变量。

第四步，根据目标变量对应的特征表示，获取待分片的模型参数集合。

在这些实现方式中，根据上述第三步所提取的目标变量对应的特征表示，上述执行主体可以通过各种方式获取待分片的模型参数集合。
20 作为示例，上述执行主体可以获取上述目标变量对应的特征表示所对应的多维数组作为上述待分片的模型参数集合。

可选地，基于上述可选的实现方式，上述执行主体还可以按照如下步骤获取待分片的模型参数集合：

S1、获取根据目标变量对应的特征表示对应的多维数组的数据量。

25 在这些实现方式中，上述执行主体可以首先获取根据目标变量对应的特征表示对应的多维数组的数据量。其中，上述数据量可以用于表征所占用的存储空间。

S2、响应于确定获取的数据量大于预设数据量阈值，获取根据目标变量对应的特征表示对应的多维数组作为待分片的模型参数集合。

30 在这些实现方式中，响应于确定上述步骤 S1 所获取的数据量大于

预设数据量阈值，上述执行主体可以获取根据目标变量对应的特征表示对应的多维数组作为待分片的模型参数集合。上述预设数据量阈值例如可以是 1G、2G。

5 基于上述可选的实现方式，上述执行主体可以在模型参数集合所占据的空间大于预设阈值的情况下将模型参数集合与模型其他部分分别存储，从而减少由于模型参数集合本身数据量较小而引起的数据传输，使得提高资源利用率的效果更加明显。

步骤 202，获取用于存储待分片的模型参数集合的存储系统的属性信息。

10 在本实施例中，上述执行主体可以通过各种方式获取用于存储上述存储待分片的模型参数集合的存储系统的属性信息。其中，上述用于存储待分片的模型参数集合的存储系统的属性信息可以包括但不限于以下至少一项：存储系统的存储空间分布，键值存储系统的标识。其中，上述存储系统的存储空间分布可以包括存储系统中各存储单元的存储容量，例如 Tensorflow 的 PS 架构中参数服务器的数量或各参数服务器的存储容量信息。上述键值存储系统的标识可以包括各种用于标识键值存储系统的信息。其中，键值存储系统可以位于上述执行主体的本地或远程，在此不作限定。作为示例，上述键值存储系统可以位于本地磁盘或 AEP (Apache Pass，一种非易失存储设备)。作为
15 又一示例，上述键值存储系统也可以位于与上述执行主体通信连接的电子设备。可选地，上述键值存储系统还可以包括本地缓存 (cache)，从而可以提高查询效率。

20 在本实施例中，上述用于存储待分片的模型参数集合的存储系统通常有别于上述待分片的模型参数集合对应的模型运行时所在的系统。作为示例，上述待分片的模型参数集合对应的模型运行时所在的系统例如可以是如图 1 所示的服务器集群 105。上述用于存储待分片的模型参数集合的存储系统例如可以是如图 1 所述的服务器 107，也可以是与服务器 107 通信连接的其他电子设备 (例如云存储系统)。

25 步骤 203，根据属性信息，将待分片的模型参数集合存入存储系统。
30

在本实施例中，根据上述步骤 202 所获取的属性信息，上述执行主体可以通过各种方式将上述步骤 201 所获取的待分片的模型参数集合存入上述步骤 202 所获取的属性信息对应的存储系统。

在本实施例的一些可选的实现方式中，基于上述属性信息包括存储系统的存储空间分布，上述执行主体可以按照如下步骤将待分片的模型参数集合存入存储系统：

第一步，根据属性信息，生成模型参数分配表。

在这些实现方式中，根据所获取的属性信息，上述执行主体可以通过各种方式生成模型参数分配表。其中，上述模型参数分配表可以用于表征上述待分片的模型参数集合与上述存储系统的存储位置之间的对应关系。作为示例，上述执行主体可以借助 Tensorflow 生成布置计划（Placement Plan）作为上述模型参数分配表。其中，上述布置计划可以根据用户预先设置的 Placement 而生成，也可以根据 Tensorflow 的 Placer 工具而生成。

第二步，根据模型参数分配表，将待分片模型参数集合存入存储系统对应的存储位置，以及生成查询接口。

在这些实现方式中，根据上述第一步生成的模型参数分配表，上述执行主体可以通过各种方式将上述待分片模型参数集合存入存储系统对应的存储位置。其中，上述存储位置可以与上述存储系统对应。作为示例，在 Tensorflow 的 PS 架构中，上述执行主体可以按照作为上述模型参数分配表的布置计划向各参数服务器发送指令，以使各参数服务器通过加载（load）相应的多维数组将相应的待分片模型参数集合存入作为存储系统对应的存储位置的各参数服务器存储系统。之后，上述执行主体还可以生成上述存储系统的查询接口。作为示例，上述执行主体可以利用 Tensorflow 中的工具，根据布置计划在相应位置插入 Send/Receive 以及其他节点。

基于上述可选的实现方式，上述执行主体可以直接利用 Tensorflow 原有的 PS 架构，将大规模的特征表示分布式存储于各参数服务器，减少了模型推断过程中模型本身的大小对系统内存的占用，在无需对原有架构进行更多改动的前提下提高了系统资源的利用率。

在本实施例的一些可选的实现方式中，上述执行主体还可以继续执行如下步骤：

第一步，接收待分片的模型参数集合对应的特征表示查询操作。

在这些实现方式中，上述执行主体可以接收待分片的模型参数集合对应的特征表示查询操作。其中，上述特征表示查询操作通常可以是上述待分片的模型参数集合对应的模型运行时所在的系统发送的。作为示例，上述特征表示查询操作例如可以是 Tensorflow 中的“tf.nn.embedding_lookup”函数。

第二步，根据特征表示查询操作，在存储系统中查询与特征表示查询操作匹配的结果。

在这些实现方式中，根据上述第一步接收到的特征表示查询操作，上述执行主体可以通过各种方式在上述存储系统中查询与上述特征表示查询操作匹配的结果。其中，上述匹配的结果通常可以为特征表示。作为示例，上述执行主体可以利用存储系统的检索系统查询与上述特征表示查询操作匹配的结果。

第三步，发送匹配的结果。

在这些实现方式中，上述执行主体可以将上述第二步查询到的匹配的结果发送至目标设备。其中，上述目标设备可以是预先指定的任意设备。上述目标设备也可以是发送上述特征表示查询操作的设备。

基于上述可选的实现方式，上述执行主体可以通过存储系统的查询功能返回与模型参数集合对应的特征表示，从而可以利用存储系统实现与模型整体加载相同的功能，因而也适用于模型推断过程。

继续参见图 3，图 3 是根据本申请实施例的用于处理模型参数的方法的应用场景的一个示意图。在图 3 的应用场景中，用户 301 通过终端设备 302 发送模型训练指令 303 至后台服务器 304。后台服务器 304 上可以对预先存储的初始商品预测模型进行训练，得到商品预测模型。其中，上述商品预测模型可以是用于表征用户信息与用户感兴趣的商品信息之间对应关系的神经网络模型。服务器 305 从后台服务器 304 获取上述训练好的商品预测模型中用于进行特征表示的多维数

组 306。而后，服务器 305 获取服务器集群 307 的属性信息 308。根据上述属性信息 308，服务器 305 将上述多维数组 306 存入上述服务器集群 307。可选地，用户 301 还可以通过终端设备 302 发送用户信息 309 至后台服务器 304 进行商品预测。后台服务器 304 在运行上述商品预测模型期间向服务器 305 发送表征查询上述用户信息 309 对应的特征表示的查询请求 310。服务器 305 将接收的查询请求 310 发送至服务器集群 307。服务器集群 307 向服务器 305 返回包括与上述用户信息 309 对应的特征表示的响应信息 311。服务器 305 将包括与上述用户信息 309 对应的特征表示的响应信息 311 返回至后台服务器 304。后台服务器 304 将接收到的特征表示输入上述商品预测模型中相应的网络层，最终得到与用户信息 309 相对应的用户感兴趣的商品信息 312。后台服务器 304 可以将上述商品信息 312 发送至终端设备 302 以显示给用户 301 查看。

目前，现有技术之一通常是将特征表示作为模型参数，在运行时与模型其他参数一同加载至内存，导致运行模型时的服务规模受制于单机内存的大小，造成诸如 CPU、网络资源方面的利用率很低，并且限制了模型本身的大小。而本申请的上述实施例提供的方法，通过将待分片的模型参数集合存储于有别于模型运行时所在的系统，实现了将占据较大空间的模型参数集合与模型运行所在的系统相分离，从而优化了大规模特征表示的参数存储方式，提升了资源利用效率。

进一步参考图 4，其示出了用于处理模型参数的方法的又一个实施例的流程 400。该用于处理模型参数的方法的流程 400，包括以下步骤：

步骤 401，获取待分片的模型参数集合。

步骤 402，获取用于存储待分片的模型参数集合的存储系统的属性信息。

在本实施例中，上述属性信息可以包括键值存储系统的标识。

上述步骤 401、步骤 402 分别与前述实施例中的步骤 201、步骤 202 一致，上文针对步骤 201 和步骤 202 的描述也适用于步骤 401 和

步骤 402，此处不再赘述。

步骤 403，获取前次更新的模型参数集合作为参考模型参数集合。

在本实施例中，用于处理模型参数的方法的执行主体（例如图 1 所示的服务器 107）可以通过各种方式获取前次更新的模型参数集合作为参考模型参数集合。其中，上述前次更新的模型参数集合可以是与上述待分片的模型参数集合对应于同一模型的模型参数集合。实践中，模型的训练往往需要进行迭代。上述前次更新的模型参数集合可以是上述待分片的模型参数集合最近一次迭代前所存储的模型参数集合。

10 步骤 404，确定待分片的模型参数集合与参考模型参数集合之间的差异，生成更新模型参数集合。

在本实施例中，上述执行主体可以通过各种方式确定上述待分片的模型参数集合和步骤 403 所获取的参考模型参数集合之间的差异，生成更新模型参数集合。其中，上述更新模型参数集合用于表征最近两次模型参数集合之间的更新部分。作为示例，上述执行主体可以将上述待分片的模型参数集合和步骤 403 所获取的参考模型参数集合之间进行差分，生成上述更新模型参数集合。

步骤 405，根据属性信息，将更新模型参数集合存入存储系统。

在本实施例中，根据步骤 402 所获取的属性信息，上述执行主体可以通过各种方式将上述更新模型参数集合存入上述存储系统。其中，上述存储系统可以包括带有容灾、可扩展容机制的远程键值存储系统。通过上述轻量、高性能的键值存储系统，使得整个系统的后续运行维护更加方便。作为示例，上述执行主体可以通过各种方式将上述更新模型参数集合映射至上述键值存储系统。

25 在本实施例的一些可选的实现方式中，上述执行主体还可以继续执行以下步骤：

第一步，接收待分片的模型参数集合对应的特征表示查询操作。

上述第一步可以与前述实施例中步骤 203 的可选的实现方式中相应步骤一致，此处不再赘述。

30 第二步，根据特征表示查询操作，生成相应数量的远程调用请求。

在这些实现方式中，根据特征表示查询操作，上述执行主体可以通过各种方式生成相应数量的远程调用请求。其中，上述远程调用请求可以用于表征在上述目标系统中查询与上述特征表示查询操作相匹配的结果。作为示例，上述执行主体可以将 embedding lookup 操作替换为上述存储系统的远程调用请求。其中，上述每个特征表示查询操作可以对应有相应数量的远程调用请求。

可选地，为了防止过多远程调用请求所占用的大量网络资源，上述执行主体还可以在接收到预设数目个特征表示查询操作后生成一个远程调用请求。其中，所生成的远程调用请求中包括前述各特征表示查询操作对应的请求内容。从而可以通过减少连接次数和对网络带宽的占用来提升整个系统性能。

第三步，向键值存储系统发送远程调用请求。

在这些实现方式中，上述执行主体可以通过各种方式向上述键值存储系统发送远程调用请求，以使上述键值存储系统查询与上述特征表示查询操作匹配的结果。其中，上述匹配的结果通常可以包括特征表示。

第四步，响应于接收到键值存储系统发送的包括与特征表示查询操作匹配的结果的响应信息，发送匹配的结果。

在这些实现方式中，响应于接收到上述键值存储系统发送的上述响应信息，上述执行主体可以向目标设备发送上述匹配的结果。其中，上述目标设备可以是预先指定的任意设备。上述目标设备也可以是发送上述特征表示查询操作的设备。

基于上述可选的实现方式，上述执行主体可以将现有技术中完整加载模型模式下的特征表示查询操作转化为向键值存储系统的发送的远程调用请求。从而使得模型运行时单机加载原生模型（例如 Tensorflow Model）时无需将包括大规模特征表示的多维数组一同加载至内存，一方面使得模型的特征表示不再受到运行模型的机型限制，另一方面也大幅提升了单机能够支持模型的数量，从而提升系统的利用率和服务性能。

从图 4 中可以看出，本实施例中的用于处理模型参数的方法的流

程 400 体现了确定待分片的模型参数集合与前次更新模型参数集合之间的差异，生成更新模型参数集合的步骤，以及将更新模型参数集合存入存储系统的步骤。由此，本实施例描述的方案可以在更新参数时无需全部更新而仅更新需要改变的部分，从而减少了网络传输流量，
5 提高了系统的服务性能。

进一步参考图 5，作为对上述各图所示方法的实现，本申请提供了用于处理模型参数的装置的一个实施例，该装置实施例与图 2 所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

10 如图 5 所示，本实施例提供的用于处理模型参数的装置 500 包括第一获取单元 501、第二获取单元 502 和存储单元 503。其中，第一获取单元 501，被配置成获取待分片的模型参数集合，其中，模型参数集合包括特征表示对应的多维数组；第二获取单元 502，被配置成获取用于存储待分片的模型参数集合的存储系统的属性信息，其中，
15 用于存储待分片的模型参数集合的存储系统有别于待分片的模型参数集合对应的模型运行时所在的系统；存储单元 503，被配置成根据属性信息，将待分片的模型参数集合存入存储系统。

在本实施例中，用于处理模型参数的装置 500 包括第一获取单元 501、第二获取单元 502 和存储单元 503 的具体处理及其所带来的技术
20 效果可分别参考图 2 对应实施例中的步骤 201、步骤 202 和步骤 203 的相关说明，在此不再赘述。

在本实施例的一些可选的实现方式中，上述第一获取单元 501 可以包括第一获取子单元（图中未示出）、解析子单元（图中未示出）、提取子单元（图中未示出）、第二获取子单元（图中未示出）。其中，
25 上述第一获取子单元，可以被配置成获取待分片的模型参数集合对应的模型文件。上述解析子单元，可以被配置成对模型文件进行解析，生成对应的解析图。其中，上述解析图中可以包括变量和对变量的操作。上述提取子单元，可以被配置成根据预设的字段，从解析图中提取变量作为目标变量。上述第二获取子单元，可以被配置成根据目标
30 变量对应的特征表示，获取待分片的模型参数集合。

在本实施例的一些可选的实现方式中，上述第二获取子单元 502 可以包括第一获取模块（图中未示出）、第二获取模块（图中未示出）。其中，上述第一获取模块，可以被配置成获取根据目标变量对应的特征表示对应的多维数组的数据量。上述第二获取模块，可以被配置成
5 响应于确定获取的数据量大于预设数据量阈值，获取根据目标变量对应的特征表示对应的多维数组作为待分片的模型参数集合。

在本实施例的一些可选的实现方式中，上述用于处理模型参数的装置 500 还可以包括：第一接收单元（图中未示出）、第一查询单元（图中未示出）、第一发送单元（图中未示出）。其中，上述第一接收单元，
10 可以被配置成接收待分片的模型参数集合对应的特征表示查询操作。上述第一查询单元，可以被配置成根据特征表示查询操作，在存储系统中查询与特征表示查询操作匹配的结果。上述第一发送单元，可以被配置成发送匹配的结果。

在本实施例的一些可选的实现方式中，上述属性信息可以包括存储系统的存储空间分布。上述存储单元 503 可以包括第一生成子单元（图中未示出）、第一存储子单元（图中未示出）。其中，上述第一获取单元 501 可以包括第一获取子单元（图中未示出）、解析子单元（图中未示出）、提取子单元（图中未示出）、第二获取子单元（图中未示出）。其中，上述第一生成子单元，可以被配置成根据属性信息，生成
20 模型参数分配表。其中，模型参数分配表可以用于表征待分片的模型参数集合与存储系统的存储位置之间的对应关系。上述第一存储子单元，可以被配置成根据模型参数分配表，将待分片模型参数集合存入存储系统对应的存储位置，以及生成查询接口。

在本实施例的一些可选的实现方式中，上述属性信息可以包括键值存储系统的标识。上述存储单元 503 可以包括第三获取子单元（图中未示出）、第二生成子单元（图中未示出）、第二存储子单元（图中未示出）。其中，上述第三获取子单元，可以被配置成获取前次更新的模型参数集合作为参考模型参数集合。上述第二生成子单元，可以被
25 配置成确定待分片的模型参数集合与参考模型参数集合之间的差异，生成更新模型参数集合。上述第二存储子单元，可以被配置成根据属
30

性信息，将更新模型参数集合存入存储系统。

在本实施例的一些可选的实现方式中，用于处理模型参数的装置还可以包括第二接收单元（图中未示出）、生成单元（图中未示出）、第二发送单元（图中未示出）、第三发送单元（图中未示出）。其中，上述第二接收单元，可以被配置成接收待分片的模型参数集合对应的特征表示查询操作。上述生成单元，可以被配置成根据特征表示查询操作，生成相应数量的远程调用请求。上述第二发送单元，可以被配置成向键值存储系统发送远程调用请求。上述第三发送单元，可以被配置成响应于接收到键值存储系统发送的包括与特征表示查询操作匹配的

10 配的结果的响应信息，发送匹配的结果。

本申请的上述实施例提供的装置，通过第一获取单元 501 获取待分片的模型参数集合，其中，模型参数集合包括特征表示对应的多维数组；第二获取单元 502 获取用于存储待分片的模型参数集合的存储系统的属性信息，其中，用于存储待分片的模型参数集合的存储系统有别于待分片的模型参数集合对应的模型运行时所在的系统；存储单元 503 根据属性信息，将待分片的模型参数集合存入存储系统。从而实现了将占据较大空间的模型参数集合与模型运行所在的系统相分离，进而优化了大规模特征表示的参数存储方式，提升了资源利用效率。

15

20

下面参考图 6，其示出了适于用来实现本申请实施例的电子设备（例如如图 1 中的服务器）600 的结构示意图。本申请实施例中的终端设备可以包括但不限于诸如移动电话、笔记本电脑、数字广播接收器、PDA（个人数字助理）、PAD（平板电脑）、PMP（便携式多媒体播放器）、车载终端（例如车载导航终端）等等的移动终端以及诸如数字 TV、台式计算机等等的固定终端。图 6 示出的服务器仅仅是一个示例，不应对本申请实施例的功能和使用范围带来任何限制。

25

如图 6 所示，电子设备 600 可以包括处理装置（例如中央处理器、图形处理器等）601，其可以根据存储在只读存储器（ROM）602 中的程序或者从存储装置 608 加载到随机访问存储器（RAM）603 中的程

30

序而执行各种适当的动作和处理。在 RAM 603 中，还存储有电子设备 600 操作所需的各种程序和数据。处理装置 601、ROM 602 以及 RAM 603 通过总线 604 彼此相连。输入/输出 (I/O) 接口 605 也连接至总线 604。

5 通常，以下装置可以连接至 I/O 接口 605：包括例如触摸屏、触摸板、键盘、鼠标等的输入装置 606；包括例如液晶显示器 (LCD, Liquid Crystal Display)、扬声器、振动器等的输出装置 607；包括例如磁带、硬盘等的存储装置 608；以及通信装置 609。通信装置 609 可以允许电子设备 600 与其他设备进行无线或有线通信以交换数据。虽然图 6 示出了具有各种装置的电子设备 600，但是应理解的是，并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。图 6 中示出的每个方框可以代表一个装置，也可以根据需要代表多个装置。

15 特别地，根据本申请的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本申请的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信装置 609 从网络上被下载和安装，或者从存储装置 608 被安装，或者从 ROM 602 被安装。在该计算机程序被处理装置 601 执行时，执行本申请的实施例的方法中限定的上述功能。

20 需要说明的是，本申请的实施例所述的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。

25 计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器 (RAM)、只读存储器 (ROM)、可擦式可编程只读存储器 (EPROM 或闪存)、光纤、便携式紧凑磁盘只读存储器 (CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本申请的实施例中，

30 计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序

可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本申请的实施例中，计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：电线、光缆、RF（Radio Frequency，射频）等等，或者上述的任意合适的组合。

上述计算机可读介质可以是上述服务器中所包含的；也可以是单独存在，而未装配入该服务器中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被该服务器/行时，使得该服务器：获取待分片的模型参数集合，其中，模型参数集合包括特征表示对应的多维数组；获取用于存储待分片的模型参数集合的存储系统的属性信息，其中，用于存储待分片的模型参数集合的存储系统有别于待分片的模型参数集合对应的模型运行时所在的系统；根据属性信息，将待分片的模型参数集合存入存储系统。

可以以一种或多种程序设计语言或其组合来编写用于执行本申请实施例的操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如 Java、Smalltalk、C++，还包括常规的过程式程序设计语言—诸如“C”语言或类似的设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。

附图中的流程图和框图，图示了按照本申请的各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这

上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。

- 5 例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

- 10 描述于本申请实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元也可以设置在处理器中，例如，可以描述为：一种处理器，包括第一获取单元、第二获取单元、存储单元。其中，这些单元的名称在某种情况下并不构成对该单元本身的限定，例如，第一获取单元还可以被描述为“获取待分片的模型参数集合的单元，其中，模型参数集合包括特征表示对应的多维数组”。
- 15

- 以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本申请的实施例中所涉及的发明范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述发明构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请实施例中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。
- 20