

模型的数据处理系统、方法和存储介质

5 本申请要求于 2022 年 12 月 27 日提交中国专利局、申请号为 202211679737.1、申请名称为“模型的数据处理系统、方法和存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及数据处理领域，具体而言，涉及一种模型的数据处理系统、方法和存储介质。

10 背景技术

随着场景越来越复杂，使用的深度学习模型也会越来越多，在多语种机器翻译或多语言识别等场景中，在线服务等深度学习模型数量可达到 100 个以上。

15 目前，单机的中央处理器的通用内存以及单张显卡的图形处理器（Graphics Processing Unit，简称为 GPU）的独占内存是有限的，无法同时将所有深度学习模型在单机或者单卡上部署，因此，在资源有限的情况下，会存在模型执行过程中数据处理所占用的存储空间多的技术问题。

针对上述的问题，目前尚未提出有效的解决方案。

发明内容

20 本申请实施例提供了一种模型的数据处理系统、方法和存储介质，以至少解决模型执行过程中数据处理所占用的存储空间多的技术问题。

25 根据本申请实施例的一个方面，提供了一种模型的数据处理的系统。该系统可以包括：第一存储设备，用于基于任意一个执行单元在深度学习模型中的预定执行顺序，确定当前待执行的目标执行单元，其中，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；至少一第一目标存储设备，用于将第一目标存储空间中存储的目标执行单元的权重数据，依次经由存储设备集合中至少一存储设备返回存储至第一存储空间，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间；执行设备，用于基于第一存储空间中存储的目标执行单元的权重数据，执行目标执行单元。

30 根据本申请实施例的一个方面，提供了一种模型的数据处理的方法。该方法可以应用于图形处理器，该方法可以包括：确定待执行的深度学习模型；基于深度学习模型中包含的任意一个执行单元的预定执行顺序，确定当前待执行的目标执行单元，其中，任意一个

执行单元的权重数据待加载到第一存储设备的第一存储空间；获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回的目标执行单元的权重数据，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间；将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据用于执行目标执行单元。

根据本申请实施例的另一个方面，提供了另一种模型的数据处理的方法。该方法可以包括：响应作用于操作界面上的模型执行指令，调用当前待执行的目标执行单元，其中，目标执行单元为基于深度学习模型中包含的任意一个执行单元在深度学习模型中的预定执行顺序而确定，任意一个执行单元的权重数据加载到第一存储设备的第一存储空间；响应作用于操作界面上的对象执行指令，基于加载到第一存储空间中的目标执行单元的权重数据，执行目标执行单元，其中，每个执行单元的权重数据为按照每个执行单元的预定执行顺序，存储至存储设备集合中至少一第一目标存储设备的第一目标存储空间，目标执行单元的权重数据依次经由存储设备集合中包含的至少一个存储设备返回存储至第一存储空间。

根据本申请实施例的另一个方面，提供了另一种模型的数据处理的方法。该方法可以包括：通过调用第一接口获取当前待执行的目标执行单元，其中，第一接口包括第一参数，第一参数的参数值为目标执行单元，目标执行单元为基于深度学习模型中包含的任意一个执行单元的预定执行顺序而确定，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回存储的目标执行单元的权重数据，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间；将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据，用于执行目标执行单元，得到执行结果；通过调用第二接口输出执行结果，其中，第二接口包括第二参数，第二参数的参数值为执行结果。

根据本申请实施例的一个方面，提供了一种模型的数据处理的装置。该装置可以包括：第一确定单元，用于确定待执行的深度学习模型；第二确定单元，用于基于深度学习模型中包含的任意一个执行单元的预定执行顺序，确定当前待执行的目标执行单元，其中，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；第一获取单元，用于获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回的目标执行单元的权重数据，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间；第一存储单元，用于将目标执行单元的权重数据存储

至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据用于执行目标执行单元。

根据本申请实施例的另一个方面，提供了另一种模型的数据处理的装置。该装置可以包括：第一调用单元，用于响应作用于操作界面上的模型执行指令，调用当前待执行的目标执行单元，其中，目标执行单元为基于深度学习模型中包含的任意一个执行单元在深度学习模型中的预定执行顺序而确定，任意一个执行单元的权重数据加载到第一存储设备的第一存储空间；执行单元，用于响应作用于操作界面上的对象执行指令，基于加载到第一存储空间中的目标执行单元的权重数据，执行目标执行单元，其中，每个执行单元的权重数据为按照每个执行单元的预定执行顺序，存储至存储设备集合中至少一第一目标存储设备的第一目标存储空间，目标执行单元的权重数据依次经由存储设备集合中包含的至少一个存储设备返回存储至第一存储空间。

根据本申请实施例的另一个方面，提供了另一种模型的数据处理的装置。该装置可以包括：第二获取单元，用于通过调用第一接口获取当前待执行的目标执行单元，其中，第一接口包括第一参数，第一参数的参数值为目标执行单元，目标执行单元为基于深度学习模型中包含的任意一个执行单元的预定执行顺序而确定，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；第三获取单元，用于获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回存储的目标执行单元的权重数据，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间；第二存储单元，用于将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据，用于执行目标执行单元，得到执行结果；输出单元，用于通过调用第二接口输出执行结果，其中，第二接口包括第二参数，第二参数的参数值为执行结果。

根据本申请实施例的另一方面，还提供了一种计算机可读存储介质，该计算机可读存储介质包括存储的程序，其中，在程序运行时控制存储介质所在设备执行上述任意一项的模型的数据处理的确定方法。

根据本申请实施例的另一方面，还提供了一种处理器，该处理器用于运行程序，其中，在程序运行时执行上述任意一项的模型的数据处理的确定方法。

在本申请实施例中，提出了一种模型数据处理系统，通过第一存储设备基于任意一个执行单元在深度学习模型中的预定执行顺序，确定当前待执行的目标执行单元，其中，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；通过至少一第一目标存储设备，将第一目标存储空间中存储的目标执行单元的权重数据，依次经由存储设备集合中至少一存储设备返回存储至第一存储空间，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数

据存储至第一目标存储空间；通过执行设备，基于第一存储空间中存储的目标执行单元的权重数据，执行目标执行单元。也就是说，本申请实施例获取存储设备集合，将每个执行单元中的权重数据依次存储至存储设备集合中的第一目标存储空间中，第一目标存储空间中的权重数据可以依次经由存储设备集合中至少一存储设备返回至第一存储空间中，最终

5 达到将权重数据存储在第一存储设备的目的，在目标执行单元执行前，可以获取第一存储空间中存储的目标执行单元对应的权重数据，基于权重数据执行目标执行单元，从而实现了降低模型执行过程中数据处理所占用的存储空间的技术效果，解决了模型执行过程中数据处理所占用的存储空间多的技术问题。

容易注意到的是，上面的通用描述和后面的详细描述仅仅是为了对本申请进行举例和

10 解释，并不构成对本申请的限定。

附图说明

此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

图 1 是根据本申请实施例的一种用于实现模型的数据处理方法的计算机终端（或移动

15 设备）的硬件结构框图；

图 2 是根据本申请实施例的一种计算环境的结构框图；

图 3 是根据本申请实施例的一种服务网络的结构框图；

图 4 是根据本申请实施例的一种模型的数据处理系统的示意图；

图 5 是根据本申请实施例的一种模型的数据处理方法的流程图；

20 图 6 是根据本申请实施例的另一种模型的数据处理方法的流程图；

图 7 是根据本申请实施例的另一种模型的数据处理方法的流程图；

图 8 是根据本申请实施例的一种计算机设备对私有网络的访问的示意图；

图 9 是根据本申请实施例的一种模型的最小执行单元连接的示意图；

图 10 是根据相关技术的一种占用显存空间的示意图；

25 图 11 是根据本申请实施例的一种权重分级换入换出的示意图；

图 12（a）是根据相关技术的一种执行时间的示意图；

图 12（b）是根据本申请的一种执行时间的示意图；

图 12（c）是根据本申请的一种灵活决定做分级存储的执行时间的示意图；

图 13 是根据本申请实施例的一种模型的数据处理装置的示意图；

图 14 是根据本申请实施例的另一种模型的数据处理装置的示意图;

图 15 是根据本申请实施例的另一种模型的数据处理装置的示意图;

图 16 是根据本申请实施例的一种计算机终端的结构框图。

具体实施方式

5 为了使本技术领域的人员更好地理解本申请方案，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，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。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都应当属于本申请保护的范围。

10 需要说明的是，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，以便这里描述的本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施。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。

首先，在对本申请实施例进行描述的过程中出现的部分名词或术语适用于如下解释：

深度学习模型，可以为一类通过数据训练得到的算法；

推理，可以指深度学习模型从给定输入、运行到得到输出的过程；

深度学习推理框架，可以指对深度学习模型做推理的软件模块；

20 内存，可以称为内存储器 and 主存储器，可以用于存放中央处理器中的运算数据，在本申请实施例中可以指中央处理器的通用内存；

显存，可以称为帧缓存，可以用于存储显示芯片处理过或者即将读取的渲染数据，在下述实施例中可以特指 GPU 的独占内存；

单卡，可以指单张显卡；

25 执行单元（OP），可以为模型的最小执行单元；

张量（tensor），可以为模型的最小数据存储单元；

自注意力机制的深度学习（Transformer）模型，可以在数据处理的过程中，用于按照输入数据各部分重要性的不同分配不同的权重；

30 卷积神经网络（Convolutional Neural Networks，简称为 CNN）模型，可以广泛用于图像识别；

响应时间 (Run Time, 简称为 RT), 可以指模型一次推理所消耗的时间。

实施例 1

根据本申请实施例, 提供了一种模型的数据处理方法, 需要说明的是, 在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行, 并且, 虽然在流程图 5 中示出了逻辑顺序, 但是在某些情况下, 可以以不同于此处的顺序执行所示出或描述的步骤。

本申请实施例一所提供的方法实施例可以在移动终端、计算机终端或者类似的运算装置中执行。图 1 是根据本申请实施例的一种用于实现模型的数据处理方法的计算机终端(或移动设备)的硬件结构框图。如图 1 所示, 计算机终端 A (或移动设备) 可以包括一个或多个 (图中采用 102a、102b, ……., 102n 来示出) 处理器 102 (处理器 102 可以包括但不限于微处理器 MCU 或可编程逻辑器件 FPGA 等的处理装置)、用于存储数据的存储器 104、以及用于通信功能的传输模块 106。除此以外, 还可以包括: 显示器、输入/输出接口 (I/O 接口)、通用串行总线 (Universal Serial Bus, USB) 端口 (可以作为 BUS 总线的端口中的一个端口被包括)、网络接口、电源和/或相机。本领域普通技术人员可以理解, 图 1 所示 15 的结构仅为示意, 其并不对上述电子装置的结构造成限定。例如, 计算机终端 A 还可包括比图 1 中所示更多或者更少的组件, 或者具有与图 1 所示不同的配置。

应当注意到的是上述一个或多个处理器 102 和/或其他的数据处理电路在本文中通常可以被称为“模型的数据处理电路”。该模型的数据处理电路可以全部或部分的体现为软件、硬件、固件或其他任意组合。此外, 模型的数据处理电路可为单个独立的处理模块, 20 或全部或部分的结合到计算机终端 A (或移动设备) 中的其他元件中的任意一个内。如本申请实施例中所涉及到的, 该模型的数据处理电路可以作为一种处理器控制 (例如与接口连接的可变电阻终端路径的选择)。

存储器 104 可用于存储应用程序的软件程序以及模块, 如本申请实施例中的模型的数据处理方法对应的程序指令/数据存储装置, 处理器 102 通过运行存储在存储器 104 内的软件 25 件程序以及模块, 从而执行各种功能应用以及数据处理, 即实现上述的模型的数据处理方法。存储器 104 可包括高速随机存储器, 还可包括非易失性存储器, 如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中, 存储器 104 可进一步包括相对于处理器 102 远程设置的存储器, 这些远程存储器可以通过网络连接至计算机终端 A。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

30 传输装置 106 用于经由一个网络接收或者发送数据。上述的网络具体实例可包括计算机终端 A 的通信供应商提供的无线网络。在一个实例中, 传输装置 106 包括一个网络适配器 (Network Interface Controller, NIC), 其可通过基站与其他网络设备相连从而可与互联网进行通讯。在一个实例中, 传输装置 106 可以为射频 (Radio Frequency, RF) 模块, 其

用于通过无线方式与互联网进行通讯。

显示器可以例如触摸屏式的液晶显示器 (Liquid Crystal Display, LCD), 该液晶显示器可使得用户能够与计算机终端 A (或移动设备) 的用户界面进行交互。

图 1 示出的硬件结构框图, 不仅可以作为上述计算机终端 A (或移动设备) 的示例性框图, 还可以作为上述服务器的示例性框图, 一种可选实施例中, 图 2 以框图示出了使用上述图 1 所示的计算机终端 A (或移动设备) 作为计算环境 201 中计算节点的一种实施例。图 2 是根据本申请实施例的一种计算环境的结构框图, 如图 2 所示, 计算环境 201 包括运行在分布式网络上的多个 (图中采用 210-1, 210-2, ..., 来示出) 计算节点 (如服务器)。每个计算节点都包含本地处理和内存资源, 终端用户 202 可以在计算环境 201 中远程运行应用程序或存储数据。应用程序可以作为计算环境 301 中的多个服务 220-1, 220-2, 220-3 和 220-4 进行提供, 分别代表服务 “A”, “D”, “E” 和 “H”。

终端用户 202 可以通过客户端上的 web 浏览器或其他软件应用程序提供和访问服务, 在一些实施例中, 可以将终端用户 202 的供应和/或请求提供给入口网关 230。入口网关 230 可以包括一个相应的代理来处理针对服务 (计算环境 201 中提供的一个或多个服务) 的供应和/或请求。

服务是根据计算环境 201 支持的各种虚拟化技术来提供或部署的。在一些实施例中, 可以根据基于虚拟机 (Virtual Machine, 简称为 VM) 的虚拟化、基于容器的虚拟化和/或类似的方式提供服务。基于虚拟机的虚拟化可以通过初始化虚拟机来模拟真实的计算机, 在不直接接触任何实际硬件资源的情况下执行程序 and 应用程序。在虚拟机虚拟化机器的同时, 根据基于容器的虚拟化, 可以启动容器来虚拟化整个操作系统 (Operating System, 简称为 OS), 以便多个工作负载可以在单个操作系统实例上运行。

在基于容器虚拟化的一个实施例中, 服务的若干容器可以被组装成一个 Pod (例如, Kubernetes Pod)。举例来说, 如图 2 所示, 服务 220-2 可以配备一个或多个 Pod 240-1, 240-2, ..., 240-N (统称为 Pod)。每个 Pod 可以包括代理 245 和一个或多个容器 242-1, 242-2, ..., 242-M (统称为容器)。Pod 中一个或多个容器处理与服务的一个或多个相应功能相关的请求, 代理 245 通常控制与服务相关的网络功能, 如路由、负载均衡等。其他服务也可以提供类似于 Pod 的 Pod。

在操作过程中, 执行来自终端用户 202 的用户请求可能需要调用计算环境 201 中的一个或多个服务, 执行一个服务的一个或多个功能可能需要调用另一个服务的一个或多个功能。如图 2 所示, 服务 “A” 220-1 从入口网关 230 接收终端用户 202 的用户请求, 服务 “A” 220-1 可以调用服务 “D” 220-2, 服务 “D” 220-2 可以请求服务 “E” 220-3 执行一个或多个功能。

上述的计算环境可以是云计算环境, 资源的分配由云服务提供上管理, 允许功能的开

发无需考虑实现、调整或扩展服务器。该计算环境允许开发人员在不构建或维护复杂基础设施的情况下执行响应事件的代码。服务可以被分割完成一组可以自动独立伸缩的功能，而不是扩展单个硬件设备来处理潜在的负载。

5 另一种可选实施例中，图 3 以框图示出了使用上述图 1 所示的计算机终端 A（或移动设备）作为服务网格的一种实施例。图 3 是根据本申请实施例的一种服务网格的结构框图，如图 3 所示，该服务网格 300 主要用于方便多个微服务之间进行安全和可靠的通信，微服务是指将应用程序分解为多个较小的服务或者实例，并分布在不同的集群/机器上运行。

10 如图 3 所示，微服务可以包括应用服务实例 A 和应用服务实例 B，应用服务实例 A 和应用服务实例 B 形成服务网格 300 的功能应用层。在一种实施方式中，应用服务实例 A 以容器/进程 308 的形式运行在机器/工作负载容器组 314（Pod），应用服务实例 B 以容器/进程 310 的形式运行在机器/工作负载容器组 316（Pod）。

在一种实施方式中，应用服务实例 A 可以是商品查询服务，应用服务实例 B 可以是商品下单服务。

15 如图 3 所示，应用服务实例 A 和网络代理（sidecar）303 共存于机器工作负载容器组 614，应用服务实例 B 和网络代理 305 共存于机器工作负载容器 314。网络代理 303 和网络代理 305 形成服务网格 300 的数据平面层（dataplane）。其中，网络代理 303 和网络代理 305 分别以容器/进程 304，容器/进程 306 的形式运行，可以接收请求 312，以用于进行商品查询服务，并且网络代理 303 和应用服务实例 A 之间可以双向通信，网络代理 305 和应用服务实例 B 之间可以双向通信。此外，网络代理 303 和网络代理 305 之间还可以双向通信。
20 信。

在一种实施方式中，应用服务实例 A 的所有流量都通过网络代理 303 被路由到合适的目的地，应用服务实例 B 的所有网络流量都通过网络代理 305 被路由到合适的目的地。需要说明的是，在此提及的网络流量包括但不限于超文本传输协议（Hyper Text Transfer Protocol，简称为 HTTP），表述性状态传递（Representational State Transfer，简称为 REST）
25 高性能等形式。

在一种实施方式中，可以通过为服务网格 300 中的代理（Envoy）编写自定义的过滤器（Filter）来实现扩展数据平面层的功能，服务网格代理配置可以是为了使服务网格正确地代理服务流量，实现服务互通和服务治理。网络代理 303 和网络代理 305 可以被配置成执行至少如下功能中的一种：服务发现（service discovery），健康检查（health checking），
30 路由（Routing），负载均衡（Load Balancing），认证和授权（authentication and authorization），以及可观测性（observability）。

如图 3 所示，该服务网格 300 还包括控制平面层。其中，控制平面层可以由一组在一个专用的命名空间中运行的服务，在机器/工作负载容器组（machine/Pod）302 中由托管

控制面组件 301 来托管这些服务。如图 3 所示, 托管控制面组件 301 与网络代理 303 和网络代理 305 进行双向通信。托管控制面组件 301 被配置成执行一些控制管理的功能。例如, 托管控制面组件 301 接收网络代理 303 和网络代理 305 传送的遥测数据, 可以进一步对这些遥测数据做聚合。这些服务, 托管控制面组件 301 还可以提供面向用户的应用程序接口 (Application Programming Interface, 简称为 API), 以便较容易地操纵网络行为, 以及向网络代理 303 和网络代理 305 提供配置数据等。

在上述运行环境下, 本申请提供了如图 4 所示的模型的数据处理系统。图 4 是根据本申请实施例的一种模型的数据处理装置的示意图。如图 4 所示, 该系统可以包括以下部分:

第一存储设备 402, 用于基于任意一个执行单元在深度学习模型中的预定执行顺序, 确定当前待执行的目标执行单元, 其中, 任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间。

在该实施例中, 可以将任意一个执行单元的权重数据加载到第一存储设备的第一存储空间中, 确定任意一个执行单元的预定执行顺序, 可以基于任意一个执行单元在深度学习模型中的预定执行顺序, 确定当前待执行的目标执行单元, 可以从第一存储空间中获取目标执行单元的权重数据。其中, 执行单元可以为深度学习模型中的最小执行单元 (OP)。可以将任意一个执行单元的权重数据待加载到第一存储空间, 该第一存储空间可以为显存空间, 比如, 可以为图形处理器的独占内存, 此处仅为举例说明, 不对第一存储空间做具体限制。权重数据可以为权重张量 (Tensor), 可以用于表征执行单元对应的权重大小。

可选地, 一个深度学习模型的推理过程, 可以包括给定输入、运行到得到输出的过程, 在推理开始时, 需要将深度学习模型包含的任意一个执行单元所依赖的权重数据加载至第一存储设备的第一存储空间中, 在推理执行时, 可以按照预定执行顺序, 确定执行单元中当前待执行的目标执行单元。

至少一第一目标存储设备 404, 用于将第一目标存储空间中存储的目标执行单元的权重数据, 依次经由存储设备集合中至少一存储设备返回存储至第一存储空间, 其中, 存储设备集合包括至少一第一目标存储设备, 第一目标存储设备用于预先按照预定执行顺序, 依次将每个执行单元的权重数据存储至第一目标存储空间;。

在该实施例中, 存储设备集合中包括至少一第一目标存储设备 404, 可以预先按照预定执行顺序, 依次将每个执行单元的权重数据存储至第一目标存储设备的第一目标存储空间中, 可以在第一目标存储空间存储的目标执行单元的权重数据, 依次经由存储设备集合中至少一第一存储设备返回存储至第一存储设备的第一存储空间中。其中, 存储设备集合可以包括第一存储设备和至少一第一目标存储设备。至少一第一目标存储设备的存储空间的存储性能可以低于第一存储空间的存储性能, 比如, 可以为存储性能低于显存空间的磁盘空间、内存空间等, 此处仅为举例说明, 不对至少一第一目标存储设备的类型做具体

限制。第一目标存储空间可以为磁盘空间，磁盘空间可以为性能低，成本也低的存储空间。

在该实施例中，可以将存储设备集合分成由低到高多个级别（level）的存储设备，从而得到至少一第一目标存储设备，级别越低的存储设备的存储空间越靠近显存空间，且级别越低性能越高，单位费用（成本）越高。其中，第一存储设备的第一存储空间可以为低
5 级别的显存空间。至少一第一目标存储设备的第一目标存储空间可以为高级别的存储空间，比如，可以为内存空间 level1、磁盘空间 level2 等。

需要说明的是，至少一第一目标存储设备的数量可以根据实际情况进行选择，可以为一个，也可以为两个，此处不对存储设备集合中包含的至少一第一目标存储设备和至少一第一目标存储设备的第一目标存储空间的数量做具体限制。

10 可选地，第一目标存储设备的数量可以与执行单元的数量相同，比如，当有两个执行单元时，可以从存储设备集合中确定两个第一目标存储设备。

在该实施例中，可以在存储设备集合中确定至少一个第一目标存储设备，可以按照每个执行单元在深度学习模型中的预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储设备中的第一目标存储空间中。

15 可选地，在深度学习模型的推理开始时，可以按照每个执行单元在深度学习模型中的预定执行顺序，以此将每个执行单元所依赖的权重数据存储至第一目标存储空间中，作为第一目标存储空间中的存储数据（冷存储）。

举例而言，可以在离线情况下，通过配套工具，将每个执行单元的权重数据，按照每个执行单元在深度学习模型中的预定执行顺序，一起加载（load）到第一目标存储空间（高
20 Level）中，作为冷存储。或者，可以在在线状态下，当深度学习模型开始第一次推理时，将每个执行单元的权重数据，按照每个执行单元在深度学习模型中的预定执行顺序，一起加载（load）到第一目标存储空间中，实现冷存储目的。需要说明的是，上述存储方式仅为举例说明，此处不对权重数据的存储方式做具体限制。

25 在该实施例中，可以在执行单元执行前，将第一目标存储空间中存储的目标执行单元的权重数据，依次经由存储设备集合中包含的至少一第一目标存储设备的第一目标存储设备返回至第一存储空间中。

可选地，在每个执行单元执行前，可以将执行的目标执行单元的权重数据从第一目标存储空间依次经由存储设备集合中包含的至少一第一目标存储设备的第一目标存储空间，返回存储至第一存储空间中，从而达到将权重数据从高 Level 的存储空间复制到低 Level
30 的存储空间中，从而达到将权重数据复制到第一存储空间（显存空间）中的目的。

举例而言，假设至少一第一目标存储设备的第一目标存储空间包括磁盘空间 Level2 和内存空间 Level1，执行单元包括 OP0、OP1 和 OP2，当 OP1 为目标执行单元时，可以将磁

盘空间 Level2 中存储的 OP1 对应的权重数据从磁盘空间 Level2 复制到内存空间 Level1，再从内存空间 Level1 复制到显存空间 Level0 中；当 OP2 为目标执行单元时，可以将磁盘空间 Level2 中存储的 OP2 对应的权重数据从磁盘空间 Level2 复制到内存空间 Level1，再从内存空间 Level1 复制到显存空间 Level0 中。

- 5 执行设备 406，用于基于第一存储空间中存储的目标执行单元的权重数据，执行目标执行单元。

在该实施例中，按照预定执行顺序，可以基于第一存储空间中存储的目标执行单元的权重数据，通过执行设备 406 执行目标执行单元。

- 10 举例而言，在深度学习模型推理执行时，可以基于预定执行顺序，确定目标执行单元，可以基于第一存储空间中存储的目标执行单元对应的权重数据，通过执行设备执行目标执行单元。

- 15 通过本申请上述系统，通过第一存储设备基于任意一个执行单元在深度学习模型中的预定执行顺序，确定当前待执行的目标执行单元，其中，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；通过至少一第一目标存储设备，将第一目标存储空间中存储的目标执行单元的权重数据，依次经由存储设备集合中至少一存储设备返回存储至第一存储空间，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间；通过执行设备，基于第一存储空间中存储的目标执行单元的权重数据，执行目标执行单元。也就是说，本申请实施例获取存储设备集合，将每个执行单元中的权重数据依次存储至存储设备集合中的第一目标存储空间中，第一目标存储空间中的权重数据可以依次经由存储设备集合中至少一存储设备返回至第一存储空间中，最终达到将权重数据存储在第一存储设备的目的，在目标执行单元执行前，可以获取第一存储空间中存储的目标执行单元对应的权重数据，基于权重数据执行目标执行单元，从而实现了降低模型执行过程中数据处理所占用的存储空间的技术效果，解决了模型执行过程中数据处理所占用的存储空间多的技术问题。
- 20
- 25

下面对该实施例的上述系统进行进一步的介绍。

作为一种可选的模型的数据处理系统，存储设备集合中包含的存储设备的存储性能均低于第一存储设备的存储性能，和/或，第一存储空间小于存储设备集合中包含的存储设备的存储空间。

- 30 由于一张图形处理器的显存空间（可以为独占内存）通常只有不到 20G，但是成本极高；而内存空间的成本却十分低，相同的存储空间，成本却不到显存空间的 1/100。因此，在该实施例中，存储设备集合中包含的存储设备的存储性能均低于第一存储设备的存储性能；第一存储空间（显存空间）小于存储设备集合中包含的存储设备的存储空间。

作为一种可选的模型的数据处理系统，存储设备集合包括除第一目标存储设备之外的第二目标存储设备，第二目标存储设备的存储性能高于第一目标存储设备的存储性能，且低于第一存储设备的存储性能，其中，第一目标存储设备用于将从第一目标存储空间中调取到的目标执行单元的权重数据，经由第二目标存储设备返回存储至第一存储空间。

5 在该实施例中，存储设备集合可以包括第一目标存储设备和除第一目标存储设备之外的第二目标存储设备，其中，第二目标存储设备可以为存储性能低于第一存储设备但高于第一目标存储设备的存储设备，可以为中央处理器的通用内存，比如，可以为存储空间为内存空间的存储设备等。可以从第一目标存储设备中调取到目标执行单元的权重数据，可以将提取到的权重数据复制到第二目标存储设备中，再将第二目标存储设备中存储的权重
10 数据返回存储至第一存储空间中。

在该实施例中，可以将存储设备集合分成由低到高多个级别（level），级别越低性能越高，单位费用（成本）越高，第二目标存储设备的级别可以低于第一目标存储设备的级别但高于第一存储设备的级别，因此，第二目标存储设备的存储性能高于第一目标存储设备的存储性能且低于第一存储设备的存储性能。

15 可选地，可以根据实际场景中对性能和成本的取舍，确定存储设备集合中存储空间的数量。

作为一种可选的模型的数据处理系统，第一目标存储设备用于将从第一目标存储空间中调取到的目标执行单元的权重数据，依次经由排序后的多个第二目标存储设备返回存储至第一存储空间，其中，多个第二目标存储设备按照存储性能依次增高的顺序进行排序。

20 在该实施例中，可以按照多个第二目标存储设备的存储性能，以存储性能依次增高的顺序，对多个第二目标存储设备进行排序，可以将第一目标存储设备中调取到的目标执行单元的权重数据，依次经由排序后的多个第二目标存储设备返回存储至第一存储空间中。

可选地，可以将存储设备集合分为多个第二目标存储设备，响应于操作界面上的第一排序指令，可以按照多个第二目标存储设备的存储性能，将多个第二目标存储设备从弱到
25 强进行排序，存储性能越高，对应的级别就越低。可以将第一目标存储设备调取到的目标执行单元的权重数据，依次经由排序后的多个第二目标存储设备返回存储至第一存储空间中。

举例而言，可以将存储设备集合分为三个第二目标存储设备，响应于操作界面上的第一排序指令，可以按照三个第二目标存储设备的存储性能，将三个第二目标存储设备从弱到
30 强进行排序，将排序后的第二目标存储设备依次命名为 OP4、OP3 和 OP2，其中，后面的数字可以用于表征第二目标存储设备对应的级别。由于第二目标存储设备的存储性能高于第一目标存储设备的存储性能，且低于第一存储设备的存储性能，所以可以假设第一存储设备为 OP0，第一目标存储设备为 OP5。可以从 OP5 中调取目标执行单元的权重数据，

并将得到的权重数据返回存储至 OP4 中, OP4 可以将存储的权重数据再次返回存储至 OP3 中, OP3 可以将存储的权重数据再次返回存储至 OP2 中, OP2 最后可以将存储的权重数据返回存储至第一存储设备的第一存储空间中。需要说明的是, 上述返回存储可以通过调用、复制等方式完成数据的返回存储, 此处不对数据返回存储的方式做具体限制。

- 5 在相关技术中, 在执行目标执行数据时, 需要将所有执行数据的权重数据返回存储至第一存储空间中, 该上述方案中, 权重数据的显存空间的占用情况为所有权重数据占用的显存空间之和, 从而导致在模型执行过程中数据处理会占用较大的存储空间, 而在本申请实施例中, 对存储设备集合的存储空间进行分级, 得到第一存储设备、至少一第一目标存储设备和多个第二目标存储设备, 将每个执行单元的权重数据存储至至少一第一目标存储设备的第一目标存储空间中, 在目标执行单元执行前, 可以将第一目标存储空间中目标执行单元的权重数据, 依次经由存储设备集合中包含的多个第二目标存储设备返回至第一存储空间中并存储, 从而使得模型在执行过程中的显存空间占用情况为所有目标执行数据的权重数据中的最大权重数据, 极大程度的减少了权重数据占用的内存空间, 进而实现了降低模型执行过程中数据处理所占用的存储空间的技术效果, 解决了模型执行过程中数据处理所占用的存储空间多的技术问题。
- 10
- 15

作为一种可选的模型的数据处理系统, 至少一第一目标存储设备用于将目标执行单元的权重数据依次经由存储设备集合中包含的至少一个存储设备复制到第一存储设备。

在该实施例中, 可以将至少一第一目标存储设备中存储的目标执行单元的权重数据依次经由存储设备集合中包含的至少一个存储设备复制到第一存储设备中。

- 20 作为一种可选的实施方式, 该系统还可以包括: 处理器, 用于从第一存储设备关联的多个存储设备的标识中选取至少一目标标识, 且将至少一目标标识对应的存储设备构成存储设备集合, 其中, 多个存储设备的存储性能低于第一存储设备的存储性能。

- 在该实施例中, 可以获取操作界面上的存储设备获取指令, 可以由处理器响应于获取到的存储设备获取指令, 可以在操作界面上显示第一存储设备关联的多个存储设备的标识, 25 可以从多个存储设备的标识中选取至少一目标标识, 且可以将选取的至少一目标标识对应的存储设备组成存储设备集合。其中, 存储设备的标识可以用于表征存储空间级别, 比如, 可以为级别 1, 级别 2, 级别 3 等, 此处仅为举例, 不对标识的显示形式做具体限制。多个存储设备的存储性能低于第一存储设备的存储性能。

- 可选地, 可以由用户基于实际需要, 对操作界面上的存储设备的标识进行选取, 响应作 30 用于操作界面上的存储设备选取指令, 可以在操作界面上从多个存储设备的标识中选取至少一目标标识, 且可以将至少一目标标识对应的存储设备构成存储设备集合。

举例而言, 可以在操作界面上显示与第一存储设备关联的多个存储设备的标识, 分别为级别 1、级别 2、级别 3, 其中, 级别越高, 性能越低, 对应的成本越低。用户可以根据

深度学习模型的实际使用场景对性能和成本进行取舍，在操作界面中对需要的存储设备进行选取，在操作界面上从多个存储设备的标识中选取至少一目标标识，可以将至少一目标标识对应的存储设备构成存储设备集合。

5 在该实施例中，可以由深度学习模型的实际使用场景，对深度学习模型的性能和成本进行取舍，从而选择合适的存储设备进行数据的存储，进一步提高本申请的灵活性和实用性。

作为一种可选的模型的数据处理系统，第一存储设备用于对多个执行单元按照预定执行顺序进行排序，且在排序后的多个执行单元中，将位于首位的执行单元确定为目标执行单元。

10 在该实施例中，可以对深度学习模型中包含的多个执行单元按照预定执行顺序进行排序，在排序后的多个执行单元中，可以将位于首位的执行单元确定为目标执行单元。

作为一种可选的模型的数据处理系统，第一存储设备用于在执行设备基于第一存储空间中存储的目标执行单元的权重数据，执行目标执行单元之后，在排序后的多个执行单元中，将目标执行单元的下一执行单元，确定为当前待执行的目标执行单元，以使得第一目标存储设备执行将第一目标存储空间中存储的目标执行单元的权重数据，依次经由存储设备集合中至少一个存储设备返回存储至第一存储空间中的步骤。

20 在该实施例中，可以按照预定执行顺序，对多个执行单元进行排序，在排序后的多个执行单元中，可以将目标执行单元的下一个执行单元确定为当前待执行的目标执行单元，可以将第一目标存储设备中存储的目标执行单元的权重数据，依次经由存储设备集合中包含的至少一个存储设备返回存储至第一存储空间。

举例而言，假设预定执行顺序为 OP1、OP3、OP2，基于预定执行顺序对执行单元进行排序，在排序后的 OP1、OP3、OP2 中，可以将首位的执行单元 OP1 确定为目标执行单元，将 OP1 的下一个执行单元 OP3 确定为当前待执行的目标执行单元。当目标执行单元为 OP1 时，可以将第一目标存储设备中存储的 OP1 的权重数据，依次经由存储设备集合中包含的至少一个存储设备返回存储至第一存储空间中。当 OP1 执行完后，目标执行单元变为 OP3，可以将第一目标存储设备中存储的 OP3 的权重数据，依次经由存储设备集合中包含的至少一个存储设备返回存储至第一存储空间中，以此类推，每个执行单元按照预定执行顺序，按顺序执行。

30 作为一种可选的模型的数据处理系统，深度学习模型包括执行单元集，其中，第一目标存储设备用于在执行单元集的权重数据中选取出多个执行单元的权重数据，且按照选出的每个执行单元在深度学习模型中的预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间。

在该实施例中，用户可以根据实际需要，自定义设置执行单元集中需要获取权重数据

的执行单元，可以在执行单元集的权重数据中选取多个执行单元的权重数据，可以按照选取出的每个执行单元在深度学习模型中的预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间。

5 举例而言，用户可以根据实际需要，自定义设置执行单元集中需要获取权重数据的执行单元，响应作用于操作界面上的由用户输入的数据选取指令，在执行单元集的权重数据中只选取与数据选取指令对应的多个执行单元的权重数据，可以按照选取出的每个执行单元在深度学习模型中的预定执行顺序，将执行单元的权重数据以此存储在第一目标存储空间中。

10 由于将权重数据从第一目标存储设备返回存储至第一存储设备的过程会产生一定的时间损失，为了减少时间损失，提高深度学习模型的数据处理效率，本申请实施例还提出一种用户自定义策略。也即，可以由用户根据实际使用情况，决定需要进行分级换入换出的权重数据，可以基于用户设定的复制策略只处理需要复制的权重数据，从而达到优化深度学习模型的数据处理速度的目的。

15 举例而言，数据选取指令可以为只对模型中前 N (TopN) 个最大的权重数据进行换入换出，响应作用于操作界面上的数据选取指令，在执行单元集的权重数据中选取 TopN 个最大的权重数据，确定 TopN 个最大的权重数据对应的执行单元，可以按照选取出的每个执行单元在深度学习模型中的预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间中。

20 作为一种可选的模型的数据处理系统，第一存储设备用于释放执行后的目标执行单元在第一存储设备中所占用的存储空间，和/或，将目标执行单元在第一存储设备中所占用的存储空间标记为失效状态。

在该实施例中，在目标执行单元执行完之后，可以释放执行后的目标执行单元在第一存储设备中所占用的存储空间，和/或，可以将目标执行单元在第一存储设备中所占用的存储空间标记为失效状态。

25 举例而言，在每个目标执行单元执行完之后，可以释放目标执行单元对应的权重数据所占用的显存空间，或者可以将目标执行单元对应的权重数据所占用的显存空间标记为无效的 (invalid)。

30 在本申请实施例中，对存储设备进行分级，得到第一存储设备、至少一第一目标存储设备和多个第二目标存储设备，可以将每个执行单元的权重数据存储至至少一第一目标存储设备的第一目标存储空间中，在目标执行单元执行前，可以将第一目标存储设备中目标执行单元的权重数据，依次经由多个第二目标存储设备返回至第一存储设备中并存储，从而实现了降低模型执行过程中数据处理所占用的存储空间的技术效果，解决了模型执行过程中数据处理所占用的存储空间多的技术问题。

本申请实施例还提供了一种模型的数据处理方法，该方法可以应用于图形处理器中。图 5 是根据本申请实施例的一种模型的数据处理方法的流程图，如图 5 所示，该方法可以包括以下步骤。

步骤 S502，确定待执行的深度学习模型。

- 5 在本申请上述步骤 S502 提供的技术方案中，可以获取目标场景下的待执行的深度学习模型，其中，目标场景可以为多语种机器翻译场景，或者多语音识别等场景，此处不对目标场景做具体限制。

举例而言，在多语种机器翻译场景或者多语言识别等场景中，在线服务的深度学习模型的数量可以达到 100 个以上，可以获取在该目标场景下待执行的深度学习模型。

- 10 步骤 S504，基于深度学习模型中包含的任意一个执行单元的预定执行顺序，确定当前待执行的目标执行单元，其中，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间。

- 在本申请上述步骤 S504 提供的技术方案中，可以基于深度学习模型中包含的任意一个执行单元的预定执行顺序，确定当前待执行的目标执行单元。其中，执行单元可以为深度学习模型中的最小执行单元（OP）。可以将任意一个执行单元的权重数据待加载到第一存储空间。第一存储空间可以为显存空间，比如，可以为图形处理器的独占内存，此处仅为举例说明，不对第一存储空间做具体限制。权重数据可以为权重张量（Tensor），可以用于表征执行单元对应的权重大小。
- 15

- 可选地，一个深度学习模型的推理过程，可以包括给定输入、运行到得到输出的过程，在推理开始时，需要将深度学习模型包含的任意一个执行单元所依赖的权重数据加载至第一存储设备的第一存储空间中，在推理执行时，可以按照预定执行顺序，确定执行单元中当前待执行的目标执行单元。
- 20

- 步骤 S506，获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回的目标执行单元的权重数据，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间。
- 25

步骤 S508，将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据用于执行目标执行单元。

- 在本申请上述步骤 S508 提供的技术方案中，可以获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回的目标执行单元的权重数据，将目标执行单元的权重数据存储至第一存储空间中，可以基于第一存储空间中存储的目标执行单元的权重数据，执行目标执行单元。
- 30

在多语种机器翻译或多语言识别等场景中，显存空间的限制导致实际处理时可同时推理的深度学习模型数量只有几个，远小于总的模型数量；可同时处理的翻译请求数量和同时推理的深度学习模型数量正相关；中央处理器或图形处理器计算资源的利用率在一定范围内和单位时间的翻译请求数量正相关。因此，在相关技术中，显存限制会导致可推理模型数量限制，从而导致翻译请求数量不足，使得中央处理器或图形处理器的计算资源不能全部利用而导致浪费。而在本申请实施例中，确定待执行的深度学习模型；基于深度学习模型中包含的任意一个执行单元的预定执行顺序，确定当前待执行的目标执行单元，其中，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回的目标执行单元的权重数据，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间；将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据用于执行目标执行单元，从而实现了降低模型执行过程中数据处理所占用的存储空间的技术效果，解决了模型执行过程中数据处理所占用的存储空间多的技术问题。

本申请实施例还提供了另一种模型的数据处理方法，该方法可以应用于人机交互场景中。需要说明的是，该实施例的模型的数据处理方法可以由本申请实施例的模型的数据处理系统执行。

图 6 是根据本申请实施例的另一种模型的数据处理方法的流程图，如图 6 所示，该方法可以包括以下步骤。

步骤 S602，响应作用于操作界面上的模型执行指令，调用当前待执行的目标执行单元，其中，目标执行单元为基于任意一个执行单元在深度学习模型中的预定执行顺序而确定，任意一个执行单元的权重数据加载到第一存储设备的第一存储空间。

在本申请上述步骤 S602 提供的技术方案中，可以获取操作界面上的模型执行指令，响应作用于操作界面上的模型执行指令，可以调用当前待执行的目标执行单元。

可选地，可以基于深度学习模型中的预定执行顺序，确定深度学习模型中的目标执行单元，可以响应作用于操作界面上的模型执行指令，调用当前待执行的目标执行单元。

步骤 S604，响应作用于操作界面上的对象执行指令，基于加载到第一存储空间中的目标执行单元的权重数据，执行目标执行单元，其中，每个执行单元的权重数据为按照每个执行单元在深度学习模型中的预定执行顺序，存储至存储设备中至少一第一目标存储设备的第一目标存储空间，且目标执行单元的权重数据依次经由存储设备集合中包含的至少一个存储设备返回存储至第一存储空间。

在本申请上述步骤 S604 提供的技术方案中，可以按照每个执行单元在深度学习模型

中的预定执行顺序，将每个执行单元的权重数据，存储至存储设备中至少一第一目标存储设备的第一目标存储空间中，其中，存储设备可以与第一存储设备相关联，且存储设备中包含的存储空间的存储性能均低于第一存储设备的存储性能。可以将目标执行单元的权重数据依次经由存储设备集合中包含的至少一个存储设备返回存储至第一存储空间中。响应

5 作用于操作界面上的对象执行指令，可以基于加载到第一存储设备中的目标执行单元的权重数据，执行目标执行单元。

通过本申请上述步骤 S602 至步骤 S604，响应作用于操作界面上的模型执行指令，调用当前待执行的目标执行单元，其中，目标执行单元为基于任意一个执行单元在深度学习模型中的预定执行顺序而确定，任意一个执行单元的权重数据加载到第一存储设备的第一

10 存储空间；响应作用于操作界面上的对象执行指令，基于加载到第一存储空间中的目标执行单元的权重数据，执行目标执行单元，其中，每个执行单元的权重数据为按照每个执行单元的预定执行顺序，存储至存储设备集合中至少一第一目标存储设备的第一目标存储空间，目标执行单元的权重数据依次经由存储设备集合中包含的至少一个存储设备返回存储至第一存储空间，从而实现了降低模型执行过程中数据处理所占用的存储空间的技术效果，

15 解决了模型执行过程中数据处理所占用的存储空间多的技术问题。

本申请实施例还提供了另一种模型的数据处理方法，该方法可以应用于软件服务侧（Software-as-a-Service，简称为 SaaS）。需要说明的是，该实施例的模型的数据处理方法可以由本申请实施例的模型的数据处理系统执行。

图 7 是根据本申请实施例的另一种模型的数据处理方法的流程图，如图 7 所示，该方法可以包括以下步骤。

20

步骤 S702，通过调用第一接口获取当前待执行的目标执行单元，其中，第一接口包括第一参数，第一参数的参数值为目标执行单元，目标执行单元为基于深度学习模型中包含的任意一个执行单元的预定执行顺序而确定，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间。

25 在本申请上述步骤 S702 提供的技术方案中，第一接口可以是服务器与用户端之间进行数据交互的接口，用户端可以通过调用第一接口获取当前待执行的目标执行单元，目标执行单元作为第一接口的一个第一参数，实现获取到目标执行单元的目的。

步骤 S704，获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回存储的目标执行单元的权重数据，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依

30 次将每个执行单元的权重数据存储至第一目标存储空间。

步骤 S706，将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据，用于执行目标执行单元，得到执行结果。

步骤 S708, 通过调用第二接口输出执行结果, 其中, 第二接口包括第二参数, 第二参数的参数值为执行结果。

在本申请上述步骤 S708 提供的技术方案中, 第二接口可以是服务器与用户端之间进行数据交互的接口, 服务器可以执行结果下发至客户端, 使得客户端可以输出执行结果至第二接口中, 作为第二接口的一个参数, 实现将执行结果下发至用户端的目的。

图 8 是根据本申请实施例的一种计算机设备对私有网络的访问的示意图, 如图 8 所示, 可以通过调用第一接口获取当前待执行的目标执行单元, 计算机设备执行: 步骤 S802, 通过调用第一接口获取当前待执行的目标执行单元; 步骤 S804, 获取从至少一第一目标存储设备的第一目标存储空间开始, 依次经由存储设备集合中包含的至少一个存储设备返回存储的目标执行单元的权重数据; 步骤 S806, 将目标执行单元的权重数据存储至第一存储空间; 步骤 S808, 通过调用第二接口输出执行结果。

可选地, 平台可以通过调用第二接口输出执行结果, 其中, 第二接口可以用于执行结果下发至客户端, 使得客户端发送执行结果的目的。

本申请实施例通过调用第一接口获取当前待执行的目标执行单元, 其中, 第一接口包括第一参数, 第一参数的参数值为目标执行单元, 目标执行单元为基于深度学习模型中包含的任意一个执行单元的预定执行顺序而确定, 任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间; 获取从至少一第一目标存储设备的第一目标存储空间开始, 依次经由存储设备集合中包含的至少一个存储设备返回存储的目标执行单元的权重数据, 其中, 存储设备集合包括至少一第一目标存储设备, 第一目标存储设备用于预先按照预定执行顺序, 依次将每个执行单元的权重数据存储至第一目标存储空间; 将目标执行单元的权重数据存储至第一存储空间, 其中, 第一存储空间中存储的目标执行单元的权重数据, 用于执行目标执行单元, 得到执行结果; 通过调用第二接口输出执行结果, 其中, 第二接口包括第二参数, 第二参数的参数值为执行结果, 从而实现了降低模型执行过程中数据处理所占用的存储空间的技术效果, 解决了模型执行过程中数据处理所占用的存储空间多的技术问题。

实施例 2

随着场景越来越复杂, 使用的深度学习模型也会越来越多, 在多语种机器翻译或多语言识别等场景中, 在线服务的深度学习模型数量可达到 100 个以上。

目前, 单机的内存单机的中央处理器的通用内存 (内存) 以及单张显卡的图形处理器的独占内存 (显存) 是有限的, 无法同时将所有深度学习模型在单机或者单卡上部署。虽然, 可以通过将一些深度学习模型分组部署到不同集群中, 以在一定程度上缓解内存问题, 但这样也存在着比如负载不均导致的计算资源浪费等问题。因此, 在计算资源有限的情况下, 仍然会出现内存或显存不够导致部署失败的情况。

在相关技术中，深度学习模型服务场景越来越复杂，随着模型数量的增长，显存资源称为深度学习模型发展的瓶颈，从而导致计算资源（比如，中央处理器或图形处理器）的浪费。

5 举例而言，在多语种翻译服务场景中，存在几百个不同的深度学习模型。显存的限制导致实际单机可同时推理的深度学习模型数量只有几个，远小于总的模型数量；单机可同时处理的翻译请求数量和可单机同时推理的深度学习模型数量正相关；中央处理器或图形处理器计算资源的利用率在一定范围内和单位时间的翻译请求数量正相关。因此，显存限制会导致可推理模型数量限制，从而导致翻译请求数量不足，使得中央处理器或图形处理器的计算资源不能全部利用而导致浪费。

10 为解决上述在模型执行过程中数据处理所占的存储空间过多，导致计算资源被浪费的问题，本申请提出一种深度学习模型服务节约显存的方法，通过一些适当的时间损失来换取显存的降低，从而提升资源利用率，降低成本。

下面对该实施例的上述方法进行进一步的介绍。

15 显存资源不足的核心问题在于，显存资源成本较高，一张图形处理器通常只有不到 20G 显存，成本通常在几万人民币，而内存成本却非常便宜，相同的存储设备成本不到图形处理器的成本的 1/100。

在计算机体系架构上通常是分级存储和以时间换空间。也即，模型在运行时，显存占用主要分为权重部分和运行中间的最小数据存储单元（Tensor）部分。其中，权重部分可以为模型最小执行单元（OP）的权重数据，可以属于静态不变的部分，也即，权重部分不需要随着请求的改变而改变。模型最小执行单元与运行中间最小数据存储单元共同占用显存空间，但是，运行中间最小数据存储单元的权重会随着请求的改变而改变。

可选地，一般在自注意力机制的深度学习（Transroformer）模型中，权重部分的权重占比为权重部分除以权重部分和运行中间最小数据存储单元二者之和，可以通过以下公式进行计算：

25
$$\text{权重占比} = \text{权重部分} / (\text{权重部分} + \text{运行中间 tensor 的权重})$$

因此，建立深度学习模型，当设置较小的序列长度（sequence_len，简称为 seq_len）和较小的批处理时，中间最小数据存储部分较小，则中间张量的权重就较小，从而执行单元的权重部分就可以得到较大的权重占比，比如，可以约占总显存占用的 80%。

30 本申请主要是针对静态不变的权重部分做的优化，针对这种静态不变的显存占用，提出了换入换出（分级存储）的优化方案。

图 9 是根据本申请实施例的一种模型的最小执行单元连接的示意图，如图 9 所示，最小执行单元 0（OP0）与最小执行单元 1（OP1）相连，最小执行单元 1（OP1）和最小执

行单元 2 (OP2) 分别与最小执行单元 3 (OP3) 相连。

以图 9 所示的模型为例, 图 10 是根据相关技术的一种占用显存空间的示意图, 如图 10 所示, 模型前 4 个最小执行单元所依赖的权重数据 (权重张量) 在显存空间上的占用情况 (分别对应图 10 中的 0, 1, 2, 3), 其中, OP0 依赖的权重张量为 0, OP1 依赖的权重张量为 1, OP2 依赖的权重张量为 2, OP3 依赖的权重张量为 3。图 10 中黑色部分为当前执行到的最小执行单元所需要的张量。

如图 9 和图 10 所示, OP0~OP3 对应依赖的 4 个权重存储单元, 可以为权重张量。对于一个深度学习模型, 给定输入, 运行, 得到输出的过程 (推理过程) 开始时, 需要将所有最小执行单元所依赖的权重张量一起加载到显存空间中; 在推理执行时, 按每个最小执行单元的 10 执行顺序, 分别使用对应依赖的权重张量, 循环执行, 模型执行过程中的权重显存占用为所有权重张量占用显存之和, 因此, 相关技术中存在模型执行过程中数据处理所占用的存储空间多的技术问题。

在该实施例中, 以图 9 所示的模型为例, 可以将存储空间分成由高到低的多个级别, 比如, 可以为显存空间 level0、内存空间 level1 和磁盘空间 level2 或者可以为显存空间 level0 15 和内存空间 level1 等, 需要说明的是, 可以为显存空间 level0 和其他任意一种或多种 level, 此处不对级别的数量做具体限制。level 越低越靠近显存, 性能越高, 单位费用越高。其中, 可以基于实际场景确定级别的数量, 从而达到可以兼顾模型性能和单位费用的目的。

图 11 是根据本申请实施例的一种权重张量分级换入换出的示意图, 如图 11 所示, 推理开始时, 所有的最小执行单元所依赖的权重数据一起加载到高级别 (磁盘空间 level2) 20 上, 作为冷存储。其中, 可以离线通过配套工具生成冷存储, 或可以在在线开始第一次推理时生成这个冷存储; 冷存储对应图 11 中磁盘空间 level2 的部分, 这部分性能低但成本也低。推理执行时, 每个最小执行单元在执行前, 可以将这个最小执行单元依赖的权重数据从高 level 的空间复制到低 level 的空间, 最终复制到显存空间(level0)中。

可选地, 复制到显存空间后, 每个最小执行单元按照执行顺序进行顺序执行, 循环执行。 25 每个最小执行单元执行完之后, 最小执行单元对应的权重部分所占的内存可以释放, 或也可以标记为无效的, 或者也可以在后面的最小执行单元运行前直接覆盖即可。

在相关技术中, 在模型执行的过程中, 显存空间的占用为所有权重张量的和; 而在该实施例中, 模型执行过程中显存空间的占用为所有权重张量中占用显存空间的最大值, 上述方法只占用部分显存空间, 因此, 显存占用远小于相关技术中的显存占用, 当模型总的 30 权重越大时效果越明显, 从而达到了在模型执行过程中减少数据处理所占用的存储空间的目的。

图 12 (a) 是根据相关技术的一种执行时间的示意图, 如图 12 (a) 所示, 总的执行时间为所有最小执行单元的执行时间之和; 图 12 (b) 是根据本申请的一种执行时间的示意图

图, 如图 12 (b) 所示, 本申请的时间损失主要为在权重张量从高 level 存储空间复制到低 level 存储空间的复制时间, 为了减少这种多出来的复制对性能的影响, 该实施例还提供一种自定义哪些权重张量做分级换入换出的机制, 可由用户根据实际需要制定相关策略来决定哪些权重张量需要做复制, 从而可以灵活的决定的显存的响应时间 (RunTime, 简称为 RT)。

5 举例而言, 图 12 (c) 是根据本申请的一种灵活决定做分级存储的执行时间的示意图, 如图 12 (c) 所示, 虚线框可以表示权重是可灵活做分级换入换出的数据存储单元, 可以标记模型中权重最大的前 N 为需要做换入换出, N 的取值可经过实际测试确定, 和机器总显存大小以及服务可允许的运行时间的范围相关。

在本申请实施例中, 将模型中每个最小执行单元的权重张量, 分级存放在硬盘-中央处
10 理器内存-图形处理器显存等不同性能和价格的存储设备上, 通过适当的策略利用推理时间换取显存的空间, 以获得更好的性价比, 从而解决了模型执行过程中数据处理所占用的存储空间多的技术问题, 实现了减少模型在执行过程中数据处理所占用的存储空间的技术效果。

需要说明的是, 对于前述的各方法实施例, 为了简单描述, 故将其都表述为一系列的动作组合, 但是本领域技术人员应该知悉, 本申请并不受所描述的动作顺序的限制, 因为
15 依据本申请, 某些步骤可以采用其他顺序或者同时进行。其次, 本领域技术人员也应该知悉, 说明书中所描述的实施例均属于优选实施例, 所涉及的动作和模块并不一定是本申请所必须的。

通过以上的实施方式的描述, 本领域的技术人员可以清楚地了解到根据上述实施例的方法可借助软件加必需的通用硬件平台的方式来实现, 当然也可以通过硬件, 但很多情况下前者是更佳的实施方式。基于这样的理解, 本申请的技术方案本质上或者说对现有技术
20 做出贡献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 (如 ROM/RAM、磁碟、光盘) 中, 包括若干指令用使得一台终端设备 (可以是手机, 计算机, 服务器, 或者网络设备等) 执行本申请各个实施例的方法。

25 实施例 3

根据本申请实施例, 还提供了一种用于实施上述图 5 所示的模型的数据处理方法的数据处理装置。

图 13 是根据本申请实施例的一种模型的数据处理方法的示意图, 如图 13 所示, 该模型的数据处理装置 1300 可以包括: 第一确定单元 1302、第二确定单元 1304、第一获取单元 1306 和第一存储单元 1308。
30

第一确定单元 1302, 用于确定待执行的深度学习模型。

第二确定单元 1304, 用于基于深度学习模型中包含的任意一个执行单元的预定执行顺

序，确定当前待执行的目标执行单元，其中，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间。

第一获取单元 1306，用于获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回的目标执行单元的权重数据，其中，
5 存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间。

第一存储单元 1308，用于将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据用于执行目标执行单元。

此处需要说明的是，上述第一确定单元 1302、第二确定单元 1304、第一获取单元 1306
10 和第一存储单元 1308 对应于实施例 1 中的步骤 S502 至步骤 S508，四个单元与对应的步骤所实现的实例和应用场景相同，但不限于上述实施例 1 所公开的内容。需要说明的是，上述单元可以是存储在存储器（例如，存储器 104）中并由一个或多个处理器（例如，处理器 102a，102b.....，102n）处理的硬件组件或软件组件，上述单元也可以作为装置的一部分可以运行在实施例 1 提供的计算机终端 A 中。

15 根据本申请实施例，还提供了一种用于实施上述图 6 所示的模型的数据处理方法的数据处理装置。

图 14 是根据本申请实施例的另一种模型的数据处理装置的示意图，如图 14 所示，该模型的数据处理装置 1400 可以包括：第一调用单元 1402 和执行单元 1404。

第一调用单元 1402，用于响应作用于操作界面上的模型执行指令，调用当前待执行的目标执行单元，其中，目标执行单元为基于深度学习模型中包含的任意一个执行单元在深度学习模型中的预定执行顺序而确定，任意一个执行单元的权重数据加载到第一存储设备的第一存储空间。
20

执行单元 1404，用于响应作用于操作界面上的对象执行指令，基于加载到第一存储空间中的目标执行单元的权重数据，执行目标执行单元，其中，每个执行单元的权重数据为
25 按照每个执行单元的预定执行顺序，存储至存储设备集合中至少一第一目标存储设备的第一目标存储空间，目标执行单元的权重数据依次经由存储设备集合中包含的至少一个存储设备返回存储至第一存储空间。

此处需要说明的是，上述第一调用单元 1402 和执行单元 1404 对应于实施例 1 中的步骤 S602 至步骤 S604，两个单元与对应的步骤所实现的实例和应用场景相同，但不限于上述实施例 1 所公开的内容。需要说明的是，上述单元可以是存储在存储器（例如，存储器 104）中并由一个或多个处理器（例如，处理器 102a，102b.....，102n）处理的硬件组件或软件组件，上述单元也可以作为装置的一部分可以运行在实施例 1 提供的计算机终端 A 中。
30

根据本申请实施例，还提供了一种用于实施上述图 7 所示的模型的数据处理方法的模型的数据处理装置。

图 15 是根据本申请实施例的另一种模型的数据处理装置的示意图，如图 15 所示，该模型的数据处理装置 1500 可以包括：第二获取单元 1502、第三获取单元 1504、第二存储单元 1506 和输出单元 1508。

第二获取单元 1502，用于通过调用第一接口获取当前待执行的目标执行单元，其中，第一接口包括第一参数，第一参数的参数值为目标执行单元，目标执行单元为基于深度学习模型中包含的任意一个执行单元的预定执行顺序而确定，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间。

10 第三获取单元 1504，用于获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回存储的目标执行单元的权重数据，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间。

15 第二存储单元 1506，用于将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据，用于执行目标执行单元，得到执行结果。

输出单元 1508，用于通过调用第二接口输出执行结果，其中，第二接口包括第二参数，第二参数的参数值为执行结果。

此处需要说明的是，上述第二获取单元 1502、第三获取单元 1504、第二存储单元 1506 和输出单元 1508 对应于实施例 1 中的步骤 S702 至步骤 S708，四个单元与对应的步骤所实现的实例和应用场景相同，但不限于上述实施例 1 所公开的内容。需要说明的是，上述单元可以是存储在存储器（例如，存储器 104）中并由一个或多个处理器（例如，处理器 102a，102b……，102n）处理的硬件组件或软件组件，上述单元也可以作为装置的一部分可以运行在实施例 1 提供的计算机终端 A 中。

25 在该实施例的模型的数据处理装置中，获取存储设备集合，将每个执行单元中的权重数据依次存储至存储设备集合中的第一目标存储空间中，第一目标存储空间中的权重数据可以依次经由存储设备集合中至少一存储设备返回至第一存储空间中，最终达到将权重数据存储在第一存储设备的目的，在目标执行单元执行前，可以获取第一存储空间中存储的目标执行单元对应的权重数据，基于权重数据执行目标执行单元，从而实现了降低模型执行过程中数据处理所占用的存储空间的技术效果，解决了模型执行过程中数据处理所占用的存储空间多的技术问题。

实施例 4

本申请的实施例可以提供一种计算机终端，该计算机终端可以是计算机终端群中的任

意一个计算机终端设备。可选地，在本实施例中，上述计算机终端也可以替换为移动终端等终端设备。

可选地，在本实施例中，上述计算机终端可以位于计算机网络的多个网络设备中的至少一个网络设备。

5 在本实施例中，上述计算机终端可以执行应用程序的模型的数据处理方法中以下步骤的程序代码：确定待执行的深度学习模型；基于深度学习模型中包含的任意一个执行单元的预定执行顺序，确定当前待执行的目标执行单元，其中，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回的目标执行单元的权重数据，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间；将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据用于执行目标执行单元。

15 可选地，图 16 是根据本申请实施例的一种计算机终端的结构框图。如图 16 所示，该计算机终端 A 可以包括：一个或多个（图中仅示出一个）处理器 1602、存储器 1604、以及传输装置 1606。

20 其中，存储器可用于存储软件程序以及模块，如本申请实施例中的模型的数据处理方法和装置对应的程序指令/模块，处理器通过运行存储在存储器内的软件程序以及模块，从而执行各种功能应用以及预测，即实现上述的模型的数据处理方法。存储器可包括高速随机存储器，还可以包括非易失性存储器，如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中，存储器可进一步包括相对于处理器远程设置的存储器，这些远程存储器可以通过网络连接至计算机终端 A。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

25 处理器可以通过传输装置调用存储器存储的信息及应用程序，以执行下述步骤：确定待执行的深度学习模型；基于深度学习模型中包含的任意一个执行单元的预定执行顺序，确定当前待执行的目标执行单元，其中，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回的目标执行单元的权重数据，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间；将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据用于执行目标执行单元。

30 作为一种可选的示例，处理器可以通过传输装置调用存储器存储的信息及应用程序，

以执行下述步骤：响应作用于操作界面上的模型执行指令，调用当前待执行的目标执行单元，其中，目标执行单元为基于深度学习模型中包含的任意一个执行单元在深度学习模型中的预定执行顺序而确定，任意一个执行单元的权重数据加载到第一存储设备的第一存储空间；响应作用于操作界面上的对象执行指令，基于加载到第一存储空间中的目标执行单元

5 的权重数据，执行目标执行单元，其中，每个执行单元的权重数据为按照每个执行单元的预定执行顺序，存储至存储设备集合中至少一第一目标存储设备的第一目标存储空间，目标执行单元的权重数据依次经由存储设备集合中包含的至少一个存储设备返回存储至第一存储空间。

作为一种可选的示例，处理器可以通过传输装置调用存储器存储的信息及应用程序，

10 以执行下述步骤：通过调用第一接口获取当前待执行的目标执行单元，其中，第一接口包括第一参数，第一参数的参数值为目标执行单元，目标执行单元为基于深度学习模型中包含的任意一个执行单元的预定执行顺序而确定，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回存储的目标执行单元的权重数据，

15 其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间；将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据，用于执行目标执行单元，得到执行结果；通过调用第二接口输出执行结果，其中，第二接口包括第二参数，第二参数的参数值为执行结果。

20 本申请实施例获取存储设备集合，将每个执行单元中的权重数据依次存储至存储设备集合中的第一目标存储空间中，第一目标存储空间中的权重数据可以依次经由存储设备集合中至少一存储设备返回至第一存储空间中，最终达到将权重数据存储在第一存储设备的目的，在目标执行单元执行前，可以获取第一存储空间中存储的目标执行单元对应的权重数据，基于权重数据执行目标执行单元，从而实现了降低模型执行过程中数据处理所占用的

25 的存储空间的技术效果，解决了模型执行过程中数据处理所占用的存储空间多的技术问题。

本领域普通技术人员可以理解，图 16 示的结构仅为示意，计算机终端 A 也可以是智能手机（如、平板电脑、掌上电脑以及移动互联网设备（Mobile Internet Devices, MID）、PAD 等终端设备。图 16 并不对上述计算机终端 A 的结构造成限定。例如，计算机终端 A 还可包括比图 16 所示更多或者更少的组件（如网络接口、显示装置等），或者具有与图 16

30 所示不同的配置。

本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令终端设备相关的硬件来完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：闪存盘、只读存储器（Read-Only Memory, ROM）、随机存取器（Random Access Memory, RAM）、磁盘或光盘等。

实施例 5

本申请的实施例还提供了一种计算机可读存储介质。可选地，在本实施例中，上述计算机可读存储介质可以用于保存上述实施例 1 所提供的模型的数据处理方法所执行的程序代码。

5 可选地，在本实施例中，上述计算机可读存储介质可以位于计算机网络中计算机终端群中的任意一个计算机终端中，或者位于移动终端群中的任意一个移动终端中。

 可选地，在本实施例中，上述计算机可读存储介质被设置为存储用于执行以下步骤的程序代码：确定待执行的深度学习模型；基于深度学习模型中包含的任意一个执行单元的预定执行顺序，确定当前待执行的目标执行单元，其中，任意一个执行单元的权重数据待
10 加载到第一存储设备的第一存储空间；获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回的目标执行单元的权重数据，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间；将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数
15 据用于执行目标执行单元。

 作为一种可选的示例，计算机可读存储介质被设置为存储用于执行以下步骤的程序代码：响应作用于操作界面上的模型执行指令，调用当前待执行的目标执行单元，其中，目标执行单元为基于深度学习模型中包含的任意一个执行单元在深度学习模型中的预定执行顺序而确定，任意一个执行单元的权重数据加载到第一存储设备的第一存储空间；响应
20 作用于操作界面上的对象执行指令，基于加载到第一存储空间中的目标执行单元的权重数据，执行目标执行单元，其中，每个执行单元的权重数据为按照每个执行单元的预定执行顺序，存储至存储设备集合中至少一第一目标存储设备的第一目标存储空间，目标执行单元的权重数据依次经由存储设备集合中包含的至少一个存储设备返回存储至第一存储空间。

25 作为一种可选的示例，计算机可读存储介质被设置为存储用于执行以下步骤的程序代码：通过调用第一接口获取当前待执行的目标执行单元，其中，第一接口包括第一参数，第一参数的参数值为目标执行单元，目标执行单元为基于深度学习模型中包含的任意一个执行单元的预定执行顺序而确定，任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储
30 设备集合中包含的至少一个存储设备返回存储的目标执行单元的权重数据，其中，存储设备集合包括至少一第一目标存储设备，第一目标存储设备用于预先按照预定执行顺序，依次将每个执行单元的权重数据存储至第一目标存储空间；将目标执行单元的权重数据存储至第一存储空间，其中，第一存储空间中存储的目标执行单元的权重数据，用于执行目标执行单元，得到执行结果；通过调用第二接口输出执行结果，其中，第二接口包括第二参

数，第二参数的参数值为执行结果。

上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。

在本申请的上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

- 5 在本申请所提供的几个实施例中，应该理解到，所揭露的技术内容，可通过其它的方式实现。其中，以上所描述的装置实施例仅仅是示意性的，例如单元的划分，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，单元或模块的间接耦合或通信连接，
10 可以是电性或其它的形式。

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。

- 15 另外，在本申请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。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。

- 20 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全部或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（可为个人计算机、服务器或者网络设备等）执行本申请各个实施例方法的全部或部分步骤。而前述的存储介质包括：U 盘、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

- 25 以上仅是本申请的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本申请的保护范围。

权 利 要 求 书

1. 一种模型数据处理系统，其特征在于，包括：

第一存储设备，用于基于任意一个执行单元在深度学习模型中的预定执行顺序，
确定当前待执行的目标执行单元，其中，所述任意一个执行单元的权重数据待加载
5 到所述第一存储设备的第一存储空间；

至少一第一目标存储设备，用于将第一目标存储空间中存储的所述目标执行单
元的权重数据，依次经由存储设备集合中至少一存储设备返回存储至所述第一存储
空间，其中，所述存储设备集合包括所述至少一第一目标存储设备，所述第一目标
存储设备用于预先按照所述预定执行顺序，依次将每个所述执行单元的权重数据存
10 储至所述第一目标存储空间；

执行设备，用于基于所述第一存储空间中存储的所述目标执行单元的权重数据，
执行所述目标执行单元。
2. 根据权利要求 1 所述的系统，其特征在于，所述存储设备集合中包含的存储设备的存
储性能均低于所述第一存储设备的存储性能，和/或，所述第一存储空间小于所述存储
15 设备集合中包含的存储设备的存储空间。
3. 根据权利要求 1 所述的系统，其特征在于，所述存储设备集合包括除所述第一目标存
储设备之外的第二目标存储设备，所述第二目标存储设备的存储性能高于所述第一目
标存储设备的存储性能，且低于所述第一存储设备的存储性能，其中，所述第一目标
存储设备用于将从所述第一目标存储空间中调取到的所述目标执行单元的权重数据，
20 经由所述第二目标存储设备返回存储至所述第一存储空间。
4. 根据权利要求 3 所述的系统，其特征在于，所述第一目标存储设备用于将从所述第一
目标存储空间中调取到的所述目标执行单元的权重数据，依次经由排序后的多个所述
第二目标存储设备返回存储至所述第一存储空间，其中，多个所述第二目标存储设备
按照存储性能依次增高的顺序进行排序。
- 25 5. 根据权利要求 1 所述的系统，其特征在于，所述至少一第一目标存储设备用于将所述
目标执行单元的权重数据依次经由所述存储设备集合中包含的至少一个存储设备复
制到所述第一存储设备。
6. 根据权利要求 1 所述的系统，其特征在于，所述系统还包括：

处理器，用于从所述第一存储设备关联的多个存储设备的标识中选取至少一目
30 标标识，且将所述至少一目标标识对应的存储设备构成所述存储设备集合，其中，
所述多个存储设备的存储性能低于所述第一存储设备的存储性能。
7. 根据权利要求 1 所述的系统，其特征在于，所述第一存储设备用于对多个所述执行单

- 元按照所述预定执行顺序进行排序，且在排序后的多个所述执行单元中，将位于首位的所述执行单元确定为所述目标执行单元。
8. 根据权利要求 7 所述的系统，其特征在于，所述第一存储设备用于在所述执行设备基于所述第一存储空间中存储的所述目标执行单元的权重数据，执行所述目标执行单元
- 5 之后，在排序后的多个所述执行单元中，将所述目标执行单元的下一执行单元，确定为所述当前待执行的目标执行单元，以使得所述第一目标存储设备执行将所述第一目标存储空间中存储的所述目标执行单元的权重数据，依次经由所述存储设备集合中至少一个存储设备返回存储至所述第一存储空间
- 10 的步骤。
9. 根据权利要求 7 所述的系统，其特征在于，所述深度学习模型包括执行单元集，其中，所述第一目标存储设备用于在所述执行单元集的权重数据中选取多个所述执行单元的权重数据，且按照选取出的每个所述执行单元在所述深度学习模型中的预定执行顺序，依次将每个所述执行单元的权重数据存储至所述第一目标存储空间。
10. 根据权利要求 1 至 9 中任意一项所述的系统，其特征在于，所述第一存储设备用于释放执行后的所述目标执行单元在所述第一存储设备中所占用的存储空间，和/或，将所述目标执行单元在所述第一存储设备中所占用的存储空间标记为失效状态。
- 15 11. 一种模型数据处理方法，其特征在于，应用于图形处理器，包括：
- 确定待执行的深度学习模型；
- 基于所述深度学习模型中包含的任意一个执行单元的预定执行顺序，确定当前待执行的目标执行单元，其中，所述任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；
- 20 获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回的所述目标执行单元的权重数据，其中，所述存储设备集合包括至少一第一目标存储设备，所述第一目标存储设备用于预先按照所述预定执行顺序，依次将每个所述执行单元的权重数据存储至第一目标存储空间；
- 25 将所述目标执行单元的权重数据存储至所述第一存储空间，其中，所述第一存储空间中存储的所述目标执行单元的权重数据用于执行所述目标执行单元。
12. 一种模型数据处理方法，其特征在于，包括：
- 响应作用于操作界面上的模型执行指令，调用当前待执行的目标执行单元，其中，所述目标执行单元为基于任意一个执行单元在深度学习模型中的预定执行顺序而确定，所述任意一个执行单元的权重数据加载到第一存储设备的第一存储空间；
- 30 响应作用于所述操作界面上的对象执行指令，基于加载到所述第一存储空间中的所述目标执行单元的权重数据，执行所述目标执行单元，其中，每个所述执行单

元的权重数据为按照每个所述执行单元的所述预定执行顺序，存储至存储设备集合中至少一第一目标存储设备的第一目标存储空间，所述目标执行单元的权重数据依次经由所述存储设备集合中包含的至少一个存储设备返回存储至所述第一存储空间。

13. 一种模型数据处理方法，其特征在于，包括：

5 通过调用第一接口获取当前待执行的目标执行单元，其中，所述第一接口包括第一参数，所述第一参数的参数值为所述目标执行单元，所述目标执行单元为基于深度学习模型中包含的任意一个执行单元的预定执行顺序而确定，所述任意一个执行单元的权重数据待加载到第一存储设备的第一存储空间；

10 获取从至少一第一目标存储设备的第一目标存储空间开始，依次经由存储设备集合中包含的至少一个存储设备返回存储的所述目标执行单元的权重数据，其中，所述存储设备集合包括至少一第一目标存储设备，所述第一目标存储设备用于预先按照所述预定执行顺序，依次将每个所述执行单元的权重数据存储至所述第一目标存储空间；

15 将所述目标执行单元的权重数据存储至所述第一存储空间，其中，所述第一存储空间中存储的所述目标执行单元的权重数据，用于执行所述目标执行单元，得到执行结果；

 通过调用第二接口输出所述执行结果，其中，所述第二接口包括第二参数，所述第二参数的参数值为所述执行结果。

20 14. 一种计算机可读存储介质，其特征在于，所述计算机可读存储介质包括存储的程序，其中，在所述程序被处理器运行时控制所述计算机可读存储介质所在设备执行权利要求 11 至 13 中任意一项所述的方法。