

显存管理方法、装置、设备及系统

本申请要求 2020 年 11 月 03 日递交的申请号为 202011219652.6、发明名称为“显存管理方法、装置、设备及系统”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。

5

技术领域

本申请涉及机器学习技术领域，具体涉及显存管理方法、装置及系统，机器学习系统，以及电子设备。

10 背景技术

随着深度学习算法的不断发展、及图形处理单元（Graphics Processing Unit，GPU）算力的加持，深度学习已经成为企业产品数据流中至关重要的一环。为了支持大规模深度学习应用，企业通常会构建共享的 GPU 集群，用来支持横跨多个领域产品的发展，如计算机视觉、自然语言处理、语音识别、推荐和广告服务等。

15 为了提升 GPU 资源利用率、及整个 GPU 集群的生产量（throughput），深度学习系统允许在一个 GPU 上同时运行多个深度学习任务，这样可以用相同的资源量完成更多的深度学习训练任务。目前，一种典型的 GPU 显存资源复用方式是，由深度学习框架内的一个统一的显存分配器进行显存管理，当该分配器接收到任何一个任务的显存资源申请时，只要运行该任务的 GPU 有空闲显存资源，就为该任务分配相应的显存空间，而不考虑该 GPU 上同时运行的其它任务对显存资源的需求，这种处理方式可以加速任务的小批
20 量训练速度。

然而，在实现本发明过程中，发明人发现上述技术方案均至少存在如下问题：1）上述资源复用方式没有提供任何性能隔离保障，会带来难以控制的多任务间的互相影响。具体而言，当 GPU 被分配给一个“资源保障”任务单独使用时，深度学习系统可以保障
25 它的任务训练性能。而由于 GPU 上缺乏了性能隔离的机制，如果这样的 GPU 上面还有其他一起执行的任务，那么潜在的显存资源竞争可能导致“资源保障”任务严重的性能下降。2）随着训练的进行，“资源保障”任务的 GPU 显存需求可能突然增长，而如果此时 GPU 显存被其它任务占用，则该“资源保障”任务将失败，这更是不能接受的。综上所述，如何对机器学习系统的共享显存资源进行管理，以在确保高优先级任务的显存
30 资源使用的前提下，提升 GPU 集群利用率，成为本领域技术人员急需解决的问题。

发明内容

本申请提供显存管理方法，以解决现有技术存在的无法保障高优先级任务性能的问题。本申请另外提供显存管理装置和系统，机器学习系统，以及电子设备。

本申请提供一种显存管理方法，包括：

确定通过图形处理单元运行的多个机器学习任务的优先级；

若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量，则释放低优先级任务占用的至少一部分显存资源；

- 5 为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务。

可选的，还包括：

释放所述多个机器学习任务占用的空闲显存资源。

可选的，所述释放所述多个机器学习任务占用的空闲显存资源，包括：

- 10 确定所述机器学习任务的显存资源使用状况信息；

若所述信息满足显存资源释放条件，则释放所述空闲显存资源。

可选的，所述使用状况信息包括：所述任务实际使用显存资源的上限值；

所述释放条件包括：所述任务的显存资源分配量大于所述上限值的时长达到时长阈值。

- 15 可选的，还包括：

释放其它高优先级任务的空闲显存资源。

可选的，所述释放低优先级任务占用的至少一部分显存资源，包括：

若低优先级任务的空闲显存资源大于或者等于显存资源需求量，则释放低优先级任务占用的空闲显存资源。

- 20 可选的，所述释放低优先级任务占用的至少一部分显存资源，包括：

若低优先级任务的空闲显存资源小于显存资源需求量，则为低优先级任务分配内存资源，并释放低优先级任务使用的至少一部分显存资源，以至少根据内存空间的张量数据，继续运行低优先级任务。

可选的，还包括：

- 25 若可分配显存资源增长到所述低优先级任务的显存资源需求量，则为所述低优先级任务分配显存资源，以根据显存空间的张量数据，继续运行所述低优先级任务。

可选的，所述低优先级任务包括：迭代学习任务；

所述释放低优先级任务使用的显存资源，包括：

在低优先级任务完成当前迭代学习后，释放所述使用的至少一部分显存资源。

- 30 可选的，在所述释放低优先级任务使用的至少一部分显存资源前，还包括：

为高优先级任务分配内存资源，以根据内存空间的张量数据和显存空间的张量数据，运行高优先级任务。

可选的，在所述为高优先级任务分配显存资源后，还包括：

释放高优先级任务的内存资源。

可选的,所述机器学习任务包括分布式深度学习任务。

本申请还提供一种显存管理方法,包括:

通过图形处理单元运行机器学习任务;

确定所述机器学习任务的显存资源使用状况信息;

- 5 若所述信息满足显存资源释放条件,则释放所述任务占用的空闲显存资源,以便将所述空闲显存资源分配给通过所述图形处理单元并行运行的其它机器学习任务。

本申请还提供一种显存管理装置,包括:

优先级确定单元,用于确定通过图形处理单元运行的多个机器学习任务的优先级;

- 10 显存释放单元,用于若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量,则释放低优先级任务占用的至少一部分显存资源;

显存分配单元,用于为高优先级任务分配显存资源,以至少根据显存空间的张量数据,运行高优先级任务。

本申请还提供一种电子设备,包括:

处理器和存储器;

- 15 存储器,用于存储实现显存管理方法的程序,该设备通电并通过所述处理器运行该方法的程序后,执行下述步骤:确定通过图形处理单元运行的多个机器学习任务的优先级;若要为高优先级任务分配显存空间、且可分配显存空间小于高优先级任务的显存空间需求量,则释放低优先级任务的至少一部分显存空间;根据低优先级任务释放的显存空间,为高优先级任务分配显存空间,以至少根据显存中的张量数据,运行高优先级任务。
- 20

本申请还提供一种显存管理装置,包括:

任务运行单元,用于通过图形处理单元运行机器学习任务。

信息确定单元,用于确定所述机器学习任务的显存资源使用状况信息。

- 25 显存释放单元,用于若所述信息满足显存资源释放条件,则释放所述任务占用的空闲显存资源,以便将所述空闲显存资源分配给通过所述图形处理单元并行运行的其它机器学习任务。

本申请还提供一种电子设备,包括:

处理器和存储器;

- 30 存储器,用于存储实现显存管理方法的程序,该设备通电并通过所述处理器运行该方法的程序后,执行下述步骤:通过图形处理单元运行机器学习任务;确定所述机器学习任务的显存资源使用状况信息;若所述信息满足显存资源释放条件,则释放所述任务占用的空闲显存资源,以便将所述空闲显存资源分配给通过所述图形处理单元并行运行的其它机器学习任务。

本申请还提供一种显存管理系统,包括:

存储资源协调器，用于确定通过图形处理单元运行的多个机器学习任务的优先级；若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源量，则向低优先级任务的第一存储资源分配器发送显存资源释放指令，向高优先级任务的第二存储资源分配器发送显存资源分配指令；

5 第一存储资源分配器，用于根据所述释放指令，释放低优先级任务占用的至少一部分显存资源；

第二存储资源分配器，用于根据所述分配指令，为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务。

可选的，所述分配器，还用于向所述协调器发送所述任务的显存资源使用状况信息；
10 所述协调器，还用于若所述信息满足显存资源释放条件，则向所述分配器发送显存资源释放指令。

可选的，所述分配器，具体用于根据预设周期，向所述协调器发送所述信息。

本申请还提供一种机器学习系统，包括：

客户端，用于向服务端发送机器学习任务的优先级信息；

15 服务端，用于确定通过图形处理单元运行的多个机器学习任务的优先级；若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量，则释放低优先级任务占用的至少一部分显存资源；为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务。

本申请还提供一种计算机可读存储介质，所述计算机可读存储介质中存储有指令，
20 当其在计算机上运行时，使得计算机执行上述各种方法。

本申请还提供一种包括指令的计算机程序产品，当其在计算机上运行时，使得计算机执行上述各种方法。

与现有技术相比，本申请具有以下优点：

本申请实施例提供的显存管理方法，通过确定通过图形处理单元运行的多个机器学习任务的优先级；若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量，则释放低优先级任务占用的至少一部分显存资源；为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务；这种处理方式，使得在可分配显存资源不足时，将低优先级任务占用的显存资源分配给高优先级任务使用，由此实现对在一个 GPU 上并行的多个机器学习任务占用的 GPU 显存资源进行
25 动态伸缩优化，这样就可以在保障高优先级任务性能的前提下，把 GPU 显存资源分配给其它任务使用；因此，可以有效提升整体集群的资源利用率，同时保障高优先级任务性能。
30

本申请实施例提供的显存管理方法，通过图形处理单元运行机器学习任务；确定所述机器学习任务的显存资源使用状况信息；若所述信息满足显存资源释放条件，则释放

所述任务占用的空闲显存资源，以便将所述空闲显存资源分配给通过所述图形处理单元并行运行的其它机器学习任务；这种处理方式，使得及时释放任务占用的空闲显存资源；因此，可以有效提升整体集群的资源利用率。

本申请实施例提供的显存管理系统，通过存储资源协调器确定通过图形处理单元运行
5 行的多个机器学习任务的优先级；若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量，则向低优先级任务的第一存储资源分配器发送显存资源释放指令，向高优先级任务的第二存储资源分配器发送显存资源分配指令；第一存储资源分配器根据所述释放指令，释放低优先级任务占用的至少一部分显存资源；第二存储资源分配器根据所述分配指令，为高优先级任务分配显存资源，以至少根据显
10 存空间的张量数据，运行高优先级任务；这种处理方式，使得可通过 GPU 设备本地的存储资源协调器，在可分配显存资源不足时，将低优先级任务占用的显存资源分配给高优先级任务使用，由此实现对在一个 GPU 上并行的多个机器学习任务占用的 GPU 显存资源进行动态伸缩优化，这样就可以在保障高优先级任务性能的前提下，把 GPU 显存资源分配给其它任务使用；因此，可以有效提升整体集群的资源利用率，同时保障高优先级
15 任务性能。

本申请实施例提供的机器学习系统，通过客户端向服务端发送机器学习任务的优先级信息；服务端确定通过图形处理单元运行的多个机器学习任务的优先级；若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量，则释放低优先级任务占用的至少一部分显存资源；为高优先级任务分配显存资源，以至少根据
20 显存空间的张量数据，运行高优先级任务；这种处理方式，使得在可分配显存资源不足时，将低优先级任务占用的显存资源分配给高优先级任务使用，由此实现对在一个 GPU 上并行的多个机器学习任务占用的 GPU 显存资源进行动态伸缩优化，这样就可以在保障高优先级任务性能的前提下，把 GPU 显存资源分配给其它任务使用；因此，可以有效提升整体集群的资源利用率，同时保障高优先级任务性能。

25

附图说明

- 图 1 本申请提供的一种显存管理方法的实施例的流程示意图；
- 图 2 本申请提供的一种显存管理方法的实施例的应用场景示意图；
- 图 3 本申请提供的一种显存管理方法的实施例的显存资源动态伸缩示意图；
- 30 图 4 本申请提供的一种显存管理方法的实施例的显存资源变化示意图；
- 图 5 本申请提供的一种显存管理装置的实施例的结构示意图；
- 图 6 本申请提供的一种显存管理系统的实施例的结构示意图。

具体实施方式

在下面的描述中阐述了很多具体细节以便于充分理解本申请。但是本申请能够以很多不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本申请内涵的情况下做类似推广，因此本申请不受下面公开的具体实施的限制。

在本申请中，提供了显存管理方法、装置和系统，机器学习系统，以及电子设备。

5 在下面的实施例中逐一对各种方案进行详细说明。

第一实施例

请参考图 1，其为本申请的显存管理方法的实施例的流程示意图。本实施例提供的方法可包括如下步骤：

步骤 S101：确定通过图形处理单元运行的多个机器学习任务的优先级。

10 本申请提供的显存管理方法可应用在机器学习系统中，用于在多个机器学习任务共享 GPU 显存资源时，对 GPU 显存资源进行分配和使用。所述机器学习系统可以是基于深度学习计算框架构建的深度学习系统，如 TensorFlow、PyTorch 等深度学习计算框架。所述机器学习任务，又称为机器学习模型训练任务，可从训练数据中学习机器学习模型。所述模型，可以是基于深度神经网络的模型，相应的，机器学习任务是深度学习任务。
15 例如，所述模型为从训练数据中学习得到的命名实体识别模型、语音识别模型、商品推荐模型等。所述模型，也可以是决策树等非神经网络的机器学习模型。

为了更加直观地说明本申请提供的显存管理方法，下面首先对该方法的应用场景进行说明。如图 2 所示，机器学习系统可以是分布式机器学习系统，包括一个或者多个基于 GPU 的计算节点，又称为 GPU 设备。其中，每个 GPU 设备可包括一个或者多个 GPU。
20 在一个 GPU 上可并行运行多个机器学习任务，这些机器学习任务共享该 GPU 的显存资源。此外，GPU 设备还包括中央处理器（CPU）及内存，其中 CPU 又可称为 GPU 的宿主机。在图 2 中，节点 1 包括 GPU1 和 GPU2，在 GPU1 上同时运行任务 D 和任务 E，在 GPU2 上同时运行任务 A、任务 B 和任务 C。

本申请提供的显存管理方法，可根据在一个 GPU 上并行的多个任务的性能保障优先级（简称优先级），任务的优先级越高，其性能越需要被保障，动态调整各个任务占用的显存资源。所述任务的优先级可根据应用需求确定，如只设置两个优先级：高优先级和低优先级。例如，命名实体识别模型的学习任务 1 为有服务水平保证的“性能保障任务”，语音识别模型的学习任务 2 为没有服务水平保证的“投机执行任务”，则可将任务 1 设置为高优先级，将任务 2 设置为低优先级。

30 具体实施时，也可以设置多个优先级。表 1 示出了一个示例中机器学习任务的优先级设置信息。

任务标识	任务优先级
任务 1（命名实体识别模型）	一级（最高优先级）
任务 2（语音识别模型）	二级（次高优先级）

任务 3（商品推荐模型）	二级（次高优先级）
任务 4（语言模型）	三级（最低优先级）
...	

表 1、机器学习任务表

由表 1 可见，本实施例中设置了三个任务优先级，其中一级为高优先级任务，二级为次高优先级，三级为最低优先级。在这种情况下，如果并行任务中有一级任务，则首先要保障一级任务的性能；如果不包括一级任务，则要保障二级任务的性能。

5 步骤 S103：若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量，则释放低优先级任务占用的至少一部分显存空间。

本申请提供的显存管理方法，在为高优先级任务分配显存资源时，如果 GPU 上可分配显存资源小于高优先级任务的显存资源需求量，则要释放低优先级任务占用的部分或者全部显存空间，将释放的显存空间分配给高优先级任务，以便尽可能在显存中存储高
10 优先级任务的张量数据，这样可以确保高优先级任务的运行性能。

所述显存资源需求量，可以是正在运行的深度学习任务要增加的显存资源，也可以是将要运行的深度学习任务初始需要的显存资源。所述显存资源需求量，可由深度学习任务确定。由于显存资源需求量的确定方式属于较为成熟的现有技术，因此此处不再赘述。

15 在本实施例中，如果一个 GPU 的机器学习任务表中的某个深度学习任务要运行，则先确定该任务需要的显存资源。在确定一个任务需要的显存资源后，如果该 GPU 的可分配显存资源小于该任务的显存资源需求量，则要确定该任务的优先级、及该 GPU 中正在运行的其它任务的优先级，如果将要运行的任务的优先级高于正在运行的其它任务的优先级，则为了保障高优先级任务的性能，需要释放低优先级任务占用的显存资源。

20 具体实施时，可以根据将要运行的任务的显存资源需求量，确定需要释放的显存资源量。例如，高优先级任务需要 1G 显存资源，则可释放 1.2G 低优先级任务占用的显存资源。

具体实施时，可以释放一个或者多个低优先级任务占用的显存资源。例如，高优先级任务需要 5G 显存资源，而释放一个低优先级任务占用的显存资源并不够用，则可多
25 个低优先级任务占用的显存资源，以释放大于 5G 的显存资源。

在一个示例中，所述方法还可包括如下步骤：释放所述多个机器学习任务占用的空闲显存资源。在本发明实现过程中，发明人发现大部分深度学习任务无法在所有时间完全利用所有分配给它的 GPU 显存资源，通常存在空闲显存资源。发明人经过研究发现，深度学习任务的空闲显存资源可能由以下几点原因产生。

30 1) 面向产品的深度学习训练任务通常会包含很多的计算部分，其中有一些并不容易并行化，因而难以完全占满 GPU 显存资源，如图神经网络中的图采样、广告模型中的特

征提取、计算机视觉里的数据增强等。

2) 深度学习系统面临海量训练数据, 随着数据的增多, 在超大规模的分布式训练中, 很多时间会花费在模型的网络数据同步阶段, 如图 2 中任务 E 同时在节点 1 和节点 n-1 上运行, 相应的在数据同步时间段内 GPU 显存资源也是空闲的。

5 3) 分布式深度学习训练通常使用同步随机梯度下降 (SGD) 的方法, 这要求训练任务所需要的资源需要同时被满足, 这样训练任务才会开始。因此, 从集群调度器的角度, 在资源不足的时候, 调度器需要为分布式任务预留部分的可用资源, 直到所有需要的资源都被满足, 这个预留的过程也导致了 GPU 显存资源处于空闲等待的状态。

10 4) 在实际应用中, 有一些张量数据仅在特定的深度学习训练阶段中使用, 如数据处理、模型评估等阶段, 而过了这个阶段就不会再用到这些张量数据, 会从显存中清除这些数据, 由此产生空闲显存资源。

然而, 在现有的深度学习框架内将不会释放上述空闲显存资源, 而是为该任务始终保有这部分显存资源, 该任务占用的整体显存资源并没有减少, 只是其中一部分为空闲显存, 该任务可在后续运行过程中, 可能会利用这部分可用显存, 如存储其它张量数据, 也可能不会利用这部分空闲显存资源, 因此导致很多时候 GPU 资源仍然处于相对低的利用率状态。

本实施例提供的方法, 基于上述显存资源使用状况, 将任务未利用完的空闲显存资源释放, 这样就可以及时分配给其它任务使用, 实现在共享 GPU 资源场景内优化多个任务的处理, 避免其它任务排队等待, 从而提升 GPU 资源利用率, 进而可提升共享 GPU 15 集群的吞吐量。

具体实施时, 所述释放所述多个机器学习任务占用的空闲显存资源, 可包括如下子步骤: 1) 确定所述机器学习任务的显存资源使用状况信息; 2) 若所述信息满足显存资源释放条件, 则释放所述空闲显存资源。所述使用状况信息包括: 所述任务实际使用显存资源的上限值; 所述释放条件包括: 所述任务的显存资源分配量大于所述上限值的时 25 长达到时长阈值。例如, 在任务运行的过程中, 每 10 秒确定一次实际使用显存资源的上限值 (波峰值), 当连续 30 秒 (3 次确定所述使用状况信息后) 都呈现任务占用的显存资源比真实需要的波峰值大的情况, 就会释放所述空闲显存资源。

在一个示例中, 所述释放低优先级任务占用的至少一部分显存空间, 可包括如下子步骤: 若低优先级任务的空闲显存资源大于或者等于显存资源需求量, 则释放低优先级 30 任务占用的空闲显存资源。在这种情况下, 使得这两类任务都可基于显存空间存储的数据运行, 这样不仅可以保障高优先级任务的运行性能, 还可避免影响低优先级任务的性能。

在另一个示例中, 所述释放低优先级任务占用的至少一部分显存空间, 还可包括如下子步骤: 若低优先级任务的空闲显存资源小于显存资源需求量, 则为低优先级任务分

配内存资源，并释放低优先级任务使用的至少一部分显存资源，以至少根据内存空间的张量数据，继续运行低优先级任务。

所述低优先级任务使用的显存资源不属于空闲显存资源，低优先级任务的张量数据仍然存储在这部分显存资源中，低优先级任务需要该资源才能保障运行性能。然而，由于释放了一部分或者全部所述低优先级任务使用的显存资源，而暂时将低优先级任务的将一部分或者全部张量数据从显存空间切换到内存空间存储，因此低优先级任务至少会根据内存空间的张量数据继续运行。如果只是释放了所述低优先级任务使用的一部分显存资源，则低优先级任务可同时基于内存空间的一部分张量数据、及显存空间的另一部分张量数据，继续运行低优先级任务。如果释放了所述低优先级任务使用的全部显存资源，则可基于内存空间的全部张量数据，继续运行低优先级任务。由此可见，无论哪种情况都不会导致任务失败。

采用这种处理方式，使得如果在释放低优先级任务的空闲显存资源后，仍不能满足高优先级任务的显存资源需求时，就需要进一步释放低优先级任务的张量数据占用的显存资源，将低优先级任务的全部或者部分张量数据转存至 GPU 宿主机的内存空间存储，至少基于内存空间的张量数据运行低优先级任务，这样就可以避免低优先级任务与高优先级任务竞争显存资源，确保高优先级任务的性能，其代价是牺牲了低优先级任务的部分性能。

如图 3 所示，子图 a 显示了一个深度学习任务在初始阶段占用的显存资源，包括一个张量的数据，其中虚线高度表示占用显存资源的水位。子图 b 显示了随着该任务所需要的张量数据相应增多，显存使用上限也在增长，这些张量数据可缓存在该任务的内存分配器（显存资源分配器）的池子里，即缓存在为该任务分配的显存空间中，可使得该任务的下一个小批量的张量数据继续复用，该任务增大后的显存资源可容纳三个张量的数据。随着任务的持续运行，有一些张量数据仅在特定的深度学习训练阶段中使用，而过了这个阶段就不会再用到这些张量数据，会从显存中清除这些数据，由此产生空闲显存资源。子图 c 显示了清除一个张量后，出现空闲显存资源，显存资源水位没有变化。

本申请实施例提供的方法，按照任务实际所需要的显存资源来动态调整显存使用上限。在本实施例中，主动检测当前正在使用的显存资源，释放空闲显存资源，从而调整其显存使用上限值到一个恰当的数值。子图 d 显示了上述可用显存一段时间始终没有使用后，将该部分可用显存（空闲显存资源）回收（释放）后的显存资源水位。

在本实施例中，深度学习系统对外暴露了接口，允许在任务运行时，上调或者缩减任务的 GPU 显存使用上限，甚至可缩小到比该任务的真实需求更低的数值。子图 e 可以是来了一个高优先级任务时，可分配显存不够用，将低优先级的一个张量的数据转存至宿主机的内存空间的情况，这样就可以把释放出来的显存资源分配给高优先级任务使用，以保障高优先级任务的性能。采用这种处理方式，即便没有足够的显存资源，低优先级

任务仍然可以继续运行，并不会失败。

在一个示例中，所述方法还可包括如下步骤：若可分配显存资源增长到所述低优先级任务的显存资源需求量，则为所述低优先级任务分配显存资源，以根据显存空间的张量数据，继续运行所述低优先级任务。采用这种处理方式，使得当 GPU 显存资源不再紧
5 缺，可上调低优先级任务的显存上限，低优先级任务的张量也可以在 GPU 上被重新分配。子图 f 显示当 GPU 显存资源不再紧缺时，重新将低优先级任务的张量数据从内存空间存储至显存空间的情况，以尽可能提升低优先级任务的性能。

在一个示例中，所述低优先级任务包括：迭代学习任务；所述释放低优先级任务使用的显存资源，可采用如下方式实现：在低优先级任务完成当前迭代学习后，释放所述
10 使用的显存资源。

在深度学习场景中，要对训练集执行梯度下降算法，通常会将整个数据集分为若干的小的训练集，每一次对一个小的子集进行训练。这样，一方面可避免全部数据集一次性参与训练带来的计算量巨大的问题；另一方面，子集的梯度方向和整个数据集差别不会太大，可保证训练的正确性。这种训练方式又称为小批量训练方式。将这种学习任务
15 称为迭代学习任务。在每次进行小批量训练时，要在存储空间中存储该次训练涉及的各种数据，存储这些数据的容器称为张量（Tensor）。张量是深度学习框架中存储数据的数据单元，作为数据容器可存储训练过程中的数据，一个任务在一次训练中可包括多个张量。张量数据可以由一组原始值组成的一个任意数量的多维数组，包括一次小批量训练的训练数据，还可包括训练过程中产生的模型参数，以及网络中各种中间节点
20 的数据等。

在本实施例中，通过在显存和内存间搬移张量数据，实现在保障高优先级任务的运行性能的前提下，提升 GPU 显存资源的利用率。采用在显存和内存间搬移张量数据的处理方式，使得可以在显存资源不足的时候将内存资源来当显存资源使用，但这通常会引入巨大的数据拷贝开销。

25 本发明的发明人通过研究深度学习任务独特的特性，即上述迭代学习的特性，发现在深度学习任务中，训练以小批量的方式进行，张量数据将在一个小批量内被创建和销毁，而相同的张量则将被在小批量之间反复创建，因此选择在小批量的边界上动态调整任务的显存使用上限，此时张量数据已经被释放，这样就可以避免显式地在显存和内存间进行数据拷贝，从而避免了巨大的拷贝开销。

30 具体实施时，在所述释放低优先级任务使用的显存资源前，所述方法还可包括如下步骤：为高优先级任务分配一部分内存资源和一部分显存资源，以根据内存空间的张量数据和显存空间的张量数据，运行高优先级任务。

具体实施时，在所述为高优先级任务分配显存资源后，所述方法还可包括如下步骤：释放高优先级任务的内存资源，这样就可以完全基于显存空间的张量数据，运行高优先

级任务。

如图 4 中的 a 所示, 高优先级任务在 T0 时刻突然激增显存需求, 显存需求量为 a, 而此时可分配显存不够用, 于是先在内存空间中占用 a, 暂时基于内存中的张量数据运行高优先级任务。在 T1 时刻, 低优先级任务的一次小批量训练完成后, 将低优先级任务的张量数据搬移至内存, 并将低优先级任务释放的显存资源 a 分配给高优先级任务使用, 同时释放高优先级任务占用的内存资源 a, 使得高优先级任务的显存资源增长了 a, 并在显存上继续执行高优先级任务。采用这种处理方式, 仅在极短时间内降低了高优先级任务的性能, 但不会导致失败, 同时可避免巨大的拷贝开销。

如图 4 中的 b 所示, 在 T0 到 T1 时间段内释放低优先级任务使用的显存资源 b, 并在内存空间中占用 b, 基于内存空间的张量数据运行低优先级任务, 损失了部分运行性能。此后, 在其它任务释放显存资源后, T1 时刻低优先级任务的显存资源增长了 b, 继续基于显存空间的张量数据运行低优先级任务。

在一个示例中, 不仅可释放低优先级任务占用的显存空间, 还可释放其它高优先级任务占用的空闲显存空间。采用这种处理方式, 可以释放更多的显存资源, 有效提升低优先级任务性能。

在一个示例中, 所述方法还可以包括步骤 S105: 根据低优先级任务释放的显存空间, 为高优先级任务分配显存空间, 以至少根据显存中的张量数据, 运行高优先级任务。

从上述实施例可见, 本申请实施例提供的显存管理方法, 通过确定通过图形处理单元运行的多个机器学习任务的优先级; 若要为高优先级任务分配显存空间、且可分配显存空间小于高优先级任务的显存空间需求量, 则释放低优先级任务的至少一部分显存空间; 根据低优先级任务释放的显存空间, 为高优先级任务分配显存空间, 以至少根据显存中的张量数据, 运行高优先级任务; 这种处理方式, 使得在可分配显存资源不足时, 将低优先级任务占用的显存资源分配给高优先级任务使用, 实现动态伸缩优化在一个 GPU 上并行的多个机器学习任务占用的 GPU 显存资源, 这样就可以在保障高优先级任务性能的前提下, 把 GPU 显存资源分配给其它任务; 因此, 可以有效提升整体集群的资源利用率, 同时保障高优先级任务性能。

第二实施例

在上述的实施例中, 提供了一种显存管理方法, 与之相对应的, 本申请还提供一种显存管理装置。该装置是与上述方法的实施例相对应。由于装置实施例基本相似于方法实施例, 所以描述得比较简单, 相关之处参见方法实施例的部分说明即可。下述描述的装置实施例仅仅是示意性的。

请参考图 5, 其为本申请的显存管理装置的实施例的结构示意图。本申请另外提供一种显存管理装置, 包括:

优先级确定单元 501, 用于确定通过图形处理单元运行的多个机器学习任务的优先

级；

显存释放单元 502，用于若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量，则释放低优先级任务占用的至少一部分显存资源；

5 显存分配单元 503，用于为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务。

第三实施例

在上述的实施例中，提供了一种显存管理方法，与之相对应的，本申请还提供一种电子设备。该装置是与上述方法的实施例相对应。由于设备实施例基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。下述描述的设备
10 实施例仅仅是示意性的。

本实施例的一种电子设备，该电子设备包括：处理器和存储器；存储器，用于存储实现显存管理方法的程序，该设备通电并通过所述处理器运行该方法的程序后，执行下述步骤：确定通过图形处理单元运行的多个机器学习任务的优先级；若要为高优先级任务分配显存空间、且可分配显存空间小于高优先级任务的显存空间需求量，则释放低
15 优先级任务的至少一部分显存空间；根据低优先级任务释放的显存空间，为高优先级任务分配显存空间，以至少根据显存中的张量数据，运行高优先级任务。

第四实施例

与上述的显存管理方法相对应，本申请还提供一种显存管理方法。本实施例与第一实施例内容相同的部分不再赘述，请参见实施例一中的相应部分。

20 在本实施例中，所述方法可包括如下步骤：

步骤 1：通过图形处理单元运行机器学习任务。

步骤 2：确定所述机器学习任务的显存资源使用状况信息。

步骤 3：若所述信息满足显存资源释放条件，则释放所述任务占用的空闲显存资源，以便将所述空闲显存资源分配给通过所述图形处理单元并行运行的其它机器学习任务。

25 所述机器学习任务，包括但不限于：深度学习任务。所述空闲显存资源，可以包括以下资源的至少一项：深度学习任务中无法并行处理的模块产生的空闲显存资源，等待深度学习任务需要的多个资源被满足时产生的空闲显存资源。

所述机器学习任务，可以是分布式深度学习任务。所述空闲显存资源，可以分布式深度学习任务对应的多个图像处理单元同步数据时产生的空闲显存资源。

30 具体而言，深度学习任务的空闲显存资源可能由以下几点原因产生。

1) 深度学习任务中无法并行处理的模块产生的空闲显存资源。面向产品的深度学习训练任务通常会包含很多的计算部分，其中有一些并不容易并行化，因而难以完全占满 GPU 显存资源，如图神经网络中的图采样、广告模型中的特征提取、计算机视觉里的数据增强等。

2) 分布式深度学习任务对应的多个图像处理单元同步数据时产生的空闲显存资源。深度学习系统面临海量训练数据,随着数据的增多,在超大规模的分布式训练中,很多时间会花费在模型的网络数据同步阶段,如图2中任务E同时在节点1和节点n-1上运行,相应的在数据同步时间段内GPU显存资源也是空闲的。

5 3) 等待深度学习任务需要的多个资源被满足时产生的空闲显存资源。分布式深度学习训练通常使用同步随机梯度下降(SGD)的方法,这要求训练任务所需要的资源需要同时被满足,这样训练任务才会开始。因此,从集群调度器的角度,在资源不足的时候,调度器需要为分布式任务预留部分的可用资源,直到所有需要的资源都被满足,这个预留的过程也导致了GPU显存资源处于空闲等待的状态。

10 4) 在实际应用中,有一些张量数据仅在特定的深度学习训练阶段中使用,如数据处理、模型评估等阶段,而过了这个阶段就不会再用到这些张量数据,会从显存中清除这些数据,由此产生空闲显存资源。

所述使用状况信息,包括但不限于:所述任务实际使用显存资源的上限值;所述释放条件,包括但不限于:所述任务的显存资源分配量大于所述上限值的时长达到时长阈值。所述时长阈值,可根据应用需求确定,如设置为30秒。

例如,在任务运行的过程中,每10秒确定一次实际使用显存资源的上限值(波峰值),当连续30秒(3次确定所述使用状况信息后)都呈现任务占用的显存资源比真实需要的波峰值大的情况,就会释放所述空闲显存资源。

从上述实施例可见,本申请实施例提供的显存管理方法,通过图形处理单元运行机器学习任务;确定所述机器学习任务的显存资源使用状况信息;若所述信息满足显存资源释放条件,则释放所述任务占用的空闲显存资源,以便将所述空闲显存资源分配给通过所述图形处理单元并行运行的其它机器学习任务;这种处理方式,使得及时释放任务占用的空闲显存资源;因此,可以有效提升整体集群的资源利用率。

第五实施例

25 在上述的实施例中,提供了一种显存管理方法,与之相对应的,本申请还提供一种显存管理装置。该装置是与上述方法的实施例相对应。由于装置实施例基本相似于方法实施例,所以描述得比较简单,相关之处参见方法实施例的部分说明即可。下述描述的装置实施例仅仅是示意性的。

在本实施例中,所述装置包括:

30 任务运行单元,用于通过图形处理单元运行机器学习任务。

信息确定单元,用于确定所述机器学习任务的显存资源使用状况信息。

显存释放单元,用于若所述信息满足显存资源释放条件,则释放所述任务占用的空闲显存资源,以便将所述空闲显存资源分配给通过所述图形处理单元并行运行的其它机器学习任务。

第六实施例

在上述的实施例中，提供了一种显存管理方法，与之相对应的，本申请还提供一种电子设备。该装置是与上述方法的实施例相对应。由于设备实施例基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。下述描述的设备

5 实施例仅仅是示意性的。

本实施例的一种电子设备，该电子设备包括：处理器和存储器；存储器，用于存储实现显存管理方法的程序，该设备通电并通过所述处理器运行该方法的程序后，执行下述步骤：通过图形处理单元运行机器学习任务；确定所述机器学习任务的显存资源使用状况信息；若所述信息满足显存资源释放条件，则释放所述任务占用的空闲显存资源，

10 以便将所述空闲显存资源分配给通过所述图形处理单元并行运行的其它机器学习任务。

第七实施例

与上述的显存管理方法相对应，本申请还提供一种显存管理系统。本实施例与第一实施例内容相同的部分不再赘述，请参见实施例一中的相应部分。

请参考图 6，其为本申请的显存管理装置的实施例的结构示意图。本申请提供一种

15 种显存管理系统包括：存储资源协调器，低优先级任务（如任务 A）的存储资源分配器（第一存储资源分配器），高优先级任务（如任务 B）的存储资源分配器（第二存储资源分配器）。

本实施例提供的系统，通过所述存储资源协调器与所述机器学习计算框架协同设计，实现针对机器学习任务的自适应的 GPU 显存资源动态伸缩优化。如图 6 所示，所述存储

20 资源协调器支持自适应 GPU 显存资源调节，可部署在 GPU 计算节点中，用于对该节点中一个或者多个 GPU 的显存资源进行调度管理，为在一个 GPU 上的多个任务动态调节 GPU 显存资源的使用量。所述机器学习计算框架与机器学习任务相对应，每个任务通过机器学习计算框架运行。

所述机器学习计算框架，可以是端到端的机器学习平台，各种机器学习框架可拥有

25 各自的生态系统，其中包含各种工具、库等资源，可助力开发者能够轻松地构建和部署由机器学习提供支持的应用。在本实施例中，机器学习计算框架为深度学习计算框架，包括但不限于 TensorFlow、PyTorch、MXNet、Caffe 等。

其中，存储资源协调器用于确定通过图形处理单元运行的多个机器学习任务的优先级；若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源量，则向低优先级任务的存储资源分配器发送显存资源释放指令，向高优先级任务的

30 存储资源分配器（第二存储资源分配器）发送显存资源分配指令；低优先级任务的存储资源分配器用于根据所述释放指令，释放低优先级任务占用的至少一部分显存资源；高优先级任务的存储资源分配器用于根据所述分配指令，为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务。

在一个示例中，所述分配器还用于向所述协调器发送所述任务的显存资源使用状况信息。具体实施时，所述分配器可根据预设周期，向所述协调器发送所述信息。相应的，所述协调器还用于若所述信息满足显存资源释放条件，则向所述分配器发送显存资源释放指令。

5 从上述实施例可见，本申请实施例提供的显存管理系统，通过存储资源协调器确定通过图形处理单元运行的多个机器学习任务的优先级；若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量，则向低优先级任务的第一存储资源分配器发送显存资源释放指令，向高优先级任务的第二存储资源分配器发送显存资源分配指令；第一存储资源分配器根据所述释放指令，释放低优先级任务占用的至少一部分显存资源；第二存储资源分配器根据所述分配指令，为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务；这种处理方式，使得可通过 GPU 设备本地的存储资源协调器，在可分配显存资源不足时，将低优先级任务占用的显存资源分配给高优先级任务使用，由此实现对在一个 GPU 上并行的多个机器学习任务占用的 GPU 显存资源进行动态伸缩优化，这样就可以在保障高优先级任务性能的前提下，把 GPU 显存资源分配给其它任务使用；因此，可以有效提升整体集群的资源利用率，
10 同时保障高优先级任务性能。
15

第八实施例

与上述的显存管理方法相对应，本申请还提供一种机器学习系统。本实施例与第一实施例内容相同的部分不再赘述，请参见实施例一中的相应部分。本申请提供的一种机器学习系统包括：客户端和服务端。
20

所述客户端用于向服务端发送机器学习任务的优先级信息；服务端用于确定通过图形处理单元运行的多个机器学习任务的优先级；若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量，则释放低优先级任务占用的至少一部分显存资源；为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务。
25

所述客户端包括但不限于移动通讯设备，即：通常所说的手机或者智能手机，还包括个人电脑、PAD、iPad 等终端设备。所述服务端，可在 GPU 集群上运行机器学习任务。

在本实施例中，可通过客户端向用户提供任务服务装置，用户可通过客户端的任务服务装置，确定要运行的机器学习任务，并设置任务优先级，如选择“二级”优先级，不同优先级可能需要支付不同的服务费用。在确定要运行的机器学习任务，并设置任务优先级后，可通过客户端向服务端提交任务运行请求。所述服务端响应该请求，可存储该任务的优先级信息，并将该任务存储至任务表中。
30

在本实施例中，所述服务端可包括一个或多个 GPU 计算节点，每个节点可通过机器学习框架运行机器学习任务，部署在 GPU 计算节点中的存储资源协调器可从服务端获取

要通过该计算节点运行的多个任务的优先级信息。表 2 示出了本实施例中机器学习任务信息。

任务标识	用户标识	任务优先级
任务 1（命名实体识别模型）	用户 A	一级（性能保障任务）
任务 2（命名实体识别模型）	用户 B	二级（投机执行任务）
任务 3（商品推荐模型）	用户 C	二级（投机执行任务）
任务 4（语言模型）	用户 C	一级（性能保障任务）
...		

表 2、机器学习任务表

由表 2 可见，用户 A 的命名实体识别模型的学习任务 1 为“性能保障任务”，用户 B 的命名实体识别模型的学习任务 2 为“投机执行任务”，“性能保障任务”的优先级高于“投机执行任务”的优先级。

在确定通过图形处理单元运行的多个机器学习任务的优先级后，就可以执行如下处理：若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量，则释放低优先级任务占用的至少一部分显存资源；为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务。

在一个示例中，所述服务端还可用于确定机器学习任务的性能信息；根据所述性能信息，调整所述任务的优先级。

例如，任务 A 的优先级原来为二级，但通过所述系统并没有使得该任务的性能达到用户要求的“二级”服务水平要求，那么可将该任务的优先级调整为一级，使得其真实性能达到用户要求的“二级”服务水平要求。

具体实施时，所述服务端可记录机器学习任务在运行过程中的显存资源变化信息，根据这些变化信息，调整该任务的优先级信息。例如，如果高优先级任务 30%的时间基于内存空间的张量数据运行，该任务的性能信息没有达到服务水平要求，则可调高该任务的优先级。

在另一个示例中，所述服务端还可用于确定机器学习任务的性能信息；根据所述性能信息，确定可服务的机器学习任务。

例如，可通过所述系统达到任务 A 和任务 B 的服务水平要求，但无法达到任务 C 的服务水平要求，则可为任务 A 和任务 B 提供显存资源管理服务。

从上述实施例可见，本申请实施例提供的机器学习系统，通过客户端向服务端发送机器学习任务的优先级信息；服务端确定通过图形处理单元运行的多个机器学习任务的优先级；若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显

存资源需求量，则释放低优先级任务占用的至少一部分显存资源；为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务；这种处理方式，使得在可分配显存资源不足时，将低优先级任务占用的显存资源分配给高优先级任务使用，由此实现对在一个 GPU 上并行的多个机器学习任务占用的 GPU 显存资源进行动态伸缩优化，这样就可以在保障高优先级任务性能的前提下，把 GPU 显存资源分配给其它任务使用；因此，可以有效提升整体集群的资源利用率，同时保障高优先级任务性能。

本申请虽然以较佳实施例公开如上，但其并不是用来限定本申请，任何本领域技术人员在不脱离本申请的精神和范围内，都可以做出可能的变动和修改，因此本申请的保护范围应当以本申请权利要求所界定的范围为准。

10 在一个典型的配置中，计算设备包括一个或多个处理器 (CPU)、 输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器 (RAM) 和/或非易失性内存等形式，如只读存储器 (ROM) 或闪存(flash RAM)。内存是计算机可读介质的示例。

15 1、计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、 程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、 其他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、
20 只读光盘只读存储器 (CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括非暂存电脑可读媒体 (transitory media)，如调制的数据信号和载波。

2、本领域技术人员应明白，本申请的实施例可提供为方法、系统或计算机程序产品。
25 因此，本申请可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

权利要求书

- 1. 一种显存管理方法，其特征在于，包括：
确定通过图形处理单元运行的多个机器学习任务的优先级；
若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求
5 量，则释放低优先级任务占用的至少一部分显存资源；
为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务。
- 2. 根据权利要求1所述的方法，其特征在于，还包括：
释放所述多个机器学习任务占用的空闲显存资源。
- 10 3. 根据权利要求2所述的方法，其特征在于，所述释放所述多个机器学习任务占用的空闲显存资源，包括：
确定所述机器学习任务的显存资源使用状况信息；
若所述信息满足显存资源释放条件，则释放所述空闲显存资源。
- 15 4. 根据权利要求3所述的方法，其特征在于，
所述使用状况信息包括：所述任务实际使用显存资源的上限值；
所述释放条件包括：所述任务的显存资源分配量大于所述上限值的时长达到时长阈值。
- 5. 根据权利要求1所述的方法，其特征在于，还包括：
释放其它高优先级任务的空闲显存资源。
- 20 6. 根据权利要求1所述的方法，其特征在于，
所述释放低优先级任务占用的至少一部分显存资源，包括：
若低优先级任务的空闲显存资源大于或者等于显存资源需求量，则释放低优先级任务占用的空闲显存资源。
- 25 7. 根据权利要求1或6所述的方法，其特征在于，
所述释放低优先级任务占用的至少一部分显存资源，包括：
若低优先级任务的空闲显存资源小于显存资源需求量，则为低优先级任务分配内存资源，并释放低优先级任务使用的至少一部分显存资源，以至少根据内存空间的张量数据，继续运行低优先级任务。
- 30 8. 根据权利要求7所述的方法，其特征在于，还包括：
若可分配显存资源增长到所述低优先级任务的显存资源需求量，则为所述低优先级任务分配显存资源，以根据显存空间的张量数据，继续运行所述低优先级任务。
- 9. 根据权利要求7所述的方法，其特征在于，
所述低优先级任务包括：迭代学习任务；
所述释放低优先级任务使用的显存资源，包括：
35 在低优先级任务完成当前迭代学习后，释放所述使用的至少一部分显存资源。

10. 根据权利要求 9 所述的方法，其特征在于，
在所述释放低优先级任务使用的至少一部分显存资源前，还包括：
为高优先级任务分配内存资源，以根据内存空间的张量数据和显存空间的张量数据，运行高优先级任务。
- 5 11. 根据权利要求 10 所述的方法，其特征在于，
在所述为高优先级任务分配显存资源后，还包括：
释放高优先级任务的内存资源。
12. 根据权利要求 1 所述的方法，其特征在于，
所述机器学习任务包括分布式深度学习任务。
- 10 13. 一种显存管理方法，其特征在于，包括：
通过图形处理单元运行机器学习任务；
确定所述机器学习任务的显存资源使用状况信息；
若所述信息满足显存资源释放条件，则释放所述任务占用的空闲显存资源，以便将
所述空闲显存资源分配给通过所述图形处理单元并行运行的其它机器学习任务。
- 15 14. 一种显存管理装置，其特征在于，包括：
优先级确定单元，用于确定通过图形处理单元运行的多个机器学习任务的优先级；
显存释放单元，用于若要为高优先级任务分配显存资源、且可分配显存资源小于高
优先级任务的显存资源需求量，则释放低优先级任务占用的至少一部分显存资源；
显存分配单元，用于为高优先级任务分配显存资源，以至少根据显存空间的张量数
20 据，运行高优先级任务。
15. 一种电子设备，其特征在于，包括：
处理器和存储器；
存储器，用于存储实现根据权利要求 1 至 12 中任一项所述的显存管理方法的程序，
该设备通电并通过所述处理器运行该方法的程序。
- 25 16. 一种显存管理装置，其特征在于，包括：
任务运行单元，用于通过图形处理单元运行机器学习任务；
信息确定单元，用于确定所述机器学习任务的显存资源使用状况信息；
显存释放单元，用于若所述信息满足显存资源释放条件，则释放所述任务占用的空
闲显存资源，以便将所述空闲显存资源分配给通过所述图形处理单元并行运行的其它机
30 器学习任务。
17. 一种电子设备，其特征在于，包括：
处理器和存储器；
存储器，用于存储实现根据权利要求 13 所述的显存管理方法的程序，该设备通电
并通过所述处理器运行该方法的程序。
- 35 18. 一种显存管理系统，其特征在于，包括：

存储资源协调器，用于确定通过图形处理单元运行的多个机器学习任务的优先级；若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源量，则向低优先级任务的第一存储资源分配器发送显存资源释放指令，向高优先级任务的第二存储资源分配器发送显存资源分配指令；

5 第一存储资源分配器，用于根据所述释放指令，释放低优先级任务占用的至少一部分显存资源；

第二存储资源分配器，用于根据所述分配指令，为高优先级任务分配显存资源，以至少根据显存空间的张量数据，运行高优先级任务。

19. 根据权利要求 18 所述的系统，其特征在于，

10 所述分配器，还用于向所述协调器发送所述任务的显存资源使用状况信息；
所述协调器，还用于若所述信息满足显存资源释放条件，则向所述分配器发送显存资源释放指令。

20. 根据权利要求 19 所述的系统，其特征在于，
所述分配器，具体用于根据预设周期，向所述协调器发送所述信息。

15 21. 一种机器学习系统，其特征在于，包括：

客户端，用于向服务端发送机器学习任务的优先级信息；

服务端，用于确定通过图形处理单元运行的多个机器学习任务的优先级；若要为高优先级任务分配显存资源、且可分配显存资源小于高优先级任务的显存资源需求量，则释放低优先级任务占用的至少一部分显存资源；为高优先级任务分配显存资源，以至少

20 根据显存空间的张量数据，运行高优先级任务。