(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119166350 A
(43)申请公布日 2024.12.20

(21)申请号 202411282560.0

(22)申请日 2024.09.13

(71)申请人 苏州元脑智能科技有限公司
地址 215100 江苏省苏州市吴中经济开发区郭巷街道官浦路1号9幢

(72)发明人 何也

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
专利代理师 宋萌

(51)Int.Cl.
G06F 9/50(2006.01)
G06F 12/02(2006.01)

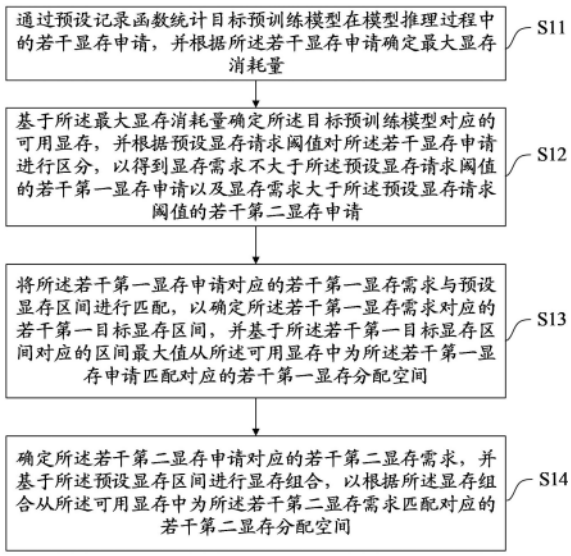
权利要求书3页 说明书11页 附图3页

(54)发明名称

一种显存管理方法、装置、设备及存储介质

(57)摘要

本发明公开了一种显存管理方法、装置、设备及存储介质,涉及内存管理技术领域,包括:通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请,以确定最大显存消耗量;基于最大显存消耗量确定目标预训练模型的可用显存,并确定显存需求不大于预设显存请求阈值的若干第一显存申请以及显存需求大于预设显存请求阈值的若干第二显存申请,并通过预设显存区间确定若干第一显存申请对应的若干第一显存分配空间以及若干第二显存申请对应的若干第二显存分配空间。由此,可以减少内存的频繁申请和释放,避免产生内存碎片,提升推理性能。



CN 119166350 A

1.一种显存管理方法,其特征在于,包括:

通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请,并根据所述若干显存申请确定最大显存消耗量;

基于所述最大显存消耗量确定所述目标预训练模型对应的可用显存,并根据预设显存请求阈值对所述若干显存申请进行区分,以得到显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请;

将所述若干第一显存申请对应的若干第一显存需求与预设显存区间进行匹配,以确定所述若干第一显存需求对应的若干第一目标显存区间,并基于所述若干第一目标显存区间对应的区间最大值从所述可用显存中为所述若干第一显存申请匹配对应的若干第一显存分配空间;

确定所述若干第二显存申请对应的若干第二显存需求,并基于所述预设显存区间进行显存组合,以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间。

2.根据权利要求1所述的显存管理方法,其特征在于,所述通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请,并根据所述若干显存申请确定最大显存消耗量,包括:

在目标预训练模型插入预设记录函数,并通过所述预设记录函数统计所述目标预训练模型在模型推理过程中的若干显存申请;

统计所述若干显存申请分别对应的若干显存需求,并基于所述若干显存需求确定最大显存消耗量。

3.根据权利要求2所述的显存管理方法,其特征在于,所述基于所述最大显存消耗量确定所述目标预训练模型对应的可用显存,并根据预设显存请求阈值对所述若干显存申请进行区分,以得到显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请,包括:

确定预设总显存与预设使用率系数的乘积,并计算所述乘积对应数值以及所述最大显存消耗量对应数值的差值,以将得到的所述差值作为所述目标预训练模型对应的可用显存;

对所述若干显存申请分别对应的若干显存需求以及预设显存请求阈值进行对比,以得到相应的对比结果;

根据所述对比结果从所述若干显存申请中筛选出显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请。

4.根据权利要求1所述的显存管理方法,其特征在于,所述将所述若干第一显存申请对应的若干第一显存需求与预设显存区间进行匹配,以确定所述若干第一显存需求对应的若干第一目标显存区间,并基于所述若干第一目标显存区间对应的区间最大值从所述可用显存中为所述若干第一显存申请匹配对应的若干第一显存分配空间,包括:

确定所述若干第一显存申请对应的若干第一显存需求;

对所述若干第一显存需求中的当前第一显存需求与预设显存区间进行匹配,以确定所述当前第一显存需求所匹配的第一目标显存区间;

确定所述第一目标显存区间对应的区间最大值,并将所述区间最大值作为所述当前第

一显存需求对应的目标第一显存需求；

从所述可用显存中为所述当前第一显存需求匹配与所述目标第一显存需求对应的第一显存分配空间；

将下一第一显存需求作为当前第一显存需求，并跳转至所述对所述若干第一显存需求中的当前第一显存需求与预设显存区间进行匹配，以确定所述当前第一显存需求所匹配的第一目标显存区间的步骤，直至所述若干第一显存需求均从所述可用显存中成功匹配相应的第一显存分配空间，以得到若干第一显存分配空间。

5.根据权利要求1所述的显存管理方法，其特征在于，所述确定所述若干第二显存申请对应的若干第二显存需求，并基于所述预设显存区间进行显存组合，以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间，包括：

确定所述若干第二显存申请对应的若干第二显存需求；

基于所述预设显存区间进行与所述若干第二显存需求中当前第二显存需求相应的显存组合，以得到满足所述当前第二显存需求的目标显存组合；

从所述可用显存中为所述当前第二显存需求匹配与所述目标显存组合对应的第二显存分配空间；

将下一第二显存需求作为当前第二显存需求，并跳转至所述基于所述预设显存区间进行与所述若干第二显存需求中当前第二显存需求相应的显存组合，以得到满足所述当前第二显存需求的目标显存组合的步骤，直至所述若干第二显存需求均从所述可用显存中成功匹配相应的第二显存分配空间，以得到若干第二显存分配空间。

6.根据权利要求1所述的显存管理方法，其特征在于，所述确定所述若干第二显存申请对应的若干第二显存需求，并基于所述预设显存区间进行显存组合，以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间之后，还包括：

计算所述若干第一显存分配空间以及所述若干第二显存分配空间对应的起始偏移量，以基于所述起始偏移量进行数据读取。

7.根据权利要求1至6任一项所述的显存管理方法，其特征在于，所述确定所述若干第二显存申请对应的若干第二显存需求，并基于所述预设显存区间进行显存组合，以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间之后，还包括：

若所述若干第一显存分配空间以及所述若干第二显存分配空间中存在已被释放的若干目标空间，则将所述若干目标空间标记为已释放状态，以得到若干已释放空间；

对所述若干已释放空间进行合并，以得到合并后已释放空间；

若接收到新的显存申请，则基于所述合并后已释放空间为所述新的现存申请进行显存匹配。

8.一种显存管理装置，其特征在于，包括：

显存消耗量确定模块，用于通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请，并根据所述若干显存申请确定最大显存消耗量；

显存申请区分模块，用于基于所述最大显存消耗量确定所述目标预训练模型对应的可用显存，并根据预设显存请求阈值对所述若干显存申请进行区分，以得到显存需求不大于

所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请;

第一显存分配模块,用于将所述若干第一显存申请对应的若干第一显存需求与预设显存区间进行匹配,以确定所述若干第一显存需求对应的若干第一目标显存区间,并基于所述若干第一目标显存区间对应的区间最大值从所述可用显存中为所述若干第一显存申请匹配对应的若干第一显存分配空间;

第二显存分配模块,用于确定所述若干第二显存申请对应的若干第二显存需求,并基于所述预设显存区间进行显存组合,以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间。

9.一种电子设备,其特征在于,包括:

存储器,用于保存计算机程序;

处理器,用于执行所述计算机程序,以实现如权利要求1至7任一项所述的显存管理方法。

10.一种计算机可读存储介质,其特征在于,用于保存计算机程序,其中,所述计算机程序被处理器执行时实现如权利要求1至7任一项所述的显存管理方法。

一种显存管理方法、装置、设备及存储介质

技术领域

[0001] 本发明涉及内存管理技术领域,特别涉及一种显存管理方法、装置、设备及存储介质。

背景技术

[0002] 随着科技的发展,预训练模型被广泛的运用在各行各业之中,例如vLLM、TensorRT-LLM等,其中vLLM以PagedAttention技术、连续批处理等而受到广泛使用。当前vLLM模型除了预先分配kv cache对其它算子的输入输出管理依赖于pytorch框架,它沿用pytorch的内存管理方案,pytorch的内存管理会根据所需的内存分配一块大的内存,然后将这块大的内存切块返回所需的部分,当后续需求新的内存时会从当前剩余的部分查找,如果有合适大小的内存就返回,否则就开辟新的内存块,如此反复。

[0003] 但现有技术分配大内存块不容易控制大小,分配过大,容易产生很多内存碎片,影响计算利用率,分配过小,则需要进行频繁的内存申请,对性能影响较大。

发明内容

[0004] 有鉴于此,本发明的目的在于提供一种显存管理方法、装置、设备及存储介质,可以尽可能减少内存的频繁申请和释放,避免产生内存碎片,提升推理性能。其具体方案如下:

[0005] 第一方面,本申请公开了一种显存管理方法,包括:

[0006] 通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请,并根据所述若干显存申请确定最大显存消耗量;

[0007] 基于所述最大显存消耗量确定所述目标预训练模型对应的可用显存,并根据预设显存请求阈值对所述若干显存申请进行区分,以得到显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请;

[0008] 将所述若干第一显存申请对应的若干第一显存需求与预设显存区间进行匹配,以确定所述若干第一显存需求对应的若干第一目标显存区间,并基于所述若干第一目标显存区间对应的区间最大值从所述可用显存中为所述若干第一显存申请匹配对应的若干第一显存分配空间;

[0009] 确定所述若干第二显存申请对应的若干第二显存需求,并基于所述预设显存区间进行显存组合,以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间。

[0010] 可选的,所述通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请,并根据所述若干显存申请确定最大显存消耗量,包括:

[0011] 在目标预训练模型插入预设记录函数,并通过所述预设记录函数统计所述目标预训练模型在模型推理过程中的若干显存申请;

[0012] 统计所述若干显存申请分别对应的若干显存需求,并基于所述若干显存需求确定

最大显存消耗量。

[0013] 可选的,所述基于所述最大显存消耗量确定所述目标预训练模型对应的可用显存,并根据预设显存请求阈值对所述若干显存申请进行区分,以得到显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请,包括:

[0014] 确定预设总显存与预设使用率系数的乘积,并计算所述乘积对应数值以及所述最大显存消耗量对应数值的差值,以将得到的所述差值作为所述目标预训练模型对应的可用显存;

[0015] 对所述若干显存申请分别对应的若干显存需求以及预设显存请求阈值进行对比,以得到相应的对比结果;

[0016] 根据所述对比结果从所述若干显存申请中筛选出显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请。

[0017] 可选的,所述将所述若干第一显存申请对应的若干第一显存需求与预设显存区间进行匹配,以确定所述若干第一显存需求对应的若干第一目标显存区间,并基于所述若干第一目标显存区间对应的区间最大值从所述可用显存中为所述若干第一显存申请匹配对应的若干第一显存分配空间,包括:

[0018] 确定所述若干第一显存申请对应的若干第一显存需求;

[0019] 对所述若干第一显存需求中的当前第一显存需求与预设显存区间进行匹配,以确定所述当前第一显存需求所匹配的第一目标显存区间;

[0020] 确定所述第一目标显存区间对应的区间最大值,并将所述区间最大值作为所述当前第一显存需求对应的目标第一显存需求;

[0021] 从所述可用显存中为所述当前第一显存需求匹配与所述目标第一显存需求对应的第一显存分配空间;

[0022] 将下一第一显存需求作为当前第一显存需求,并跳转至所述对所述若干第一显存需求中的当前第一显存需求与预设显存区间进行匹配,以确定所述当前第一显存需求所匹配的第一目标显存区间的步骤,直至所述若干第一显存需求均从所述可用显存中成功匹配相应的第一显存分配空间,以得到若干第一显存分配空间。

[0023] 可选的,所述确定所述若干第二显存申请对应的若干第二显存需求,并基于所述预设显存区间进行显存组合,以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间,包括:

[0024] 确定所述若干第二显存申请对应的若干第二显存需求;

[0025] 基于所述预设显存区间进行与所述若干第二显存需求中当前第二显存需求相应的显存组合,以得到满足所述当前第二显存需求的目标显存组合;

[0026] 从所述可用显存中为所述当前第二显存需求匹配与所述目标显存组合对应的第二显存分配空间;

[0027] 将下一第二显存需求作为当前第二显存需求,并跳转至所述基于所述预设显存区间进行与所述若干第二显存需求中当前第二显存需求相应的显存组合,以得到满足所述当前第二显存需求的目标显存组合的步骤,直至所述若干第二显存需求均从所述可用显存中

成功匹配相应的第二显存分配空间,以得到若干第二显存分配空间。

[0028] 可选的,所述确定所述若干第二显存申请对应的若干第二显存需求,并基于所述预设显存区间进行显存组合,以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间之后,还包括:

[0029] 计算所述若干第一显存分配空间以及所述若干第二显存分配空间对应的起始偏移量,以基于所述起始偏移量进行数据读取。

[0030] 可选的,所述确定所述若干第二显存申请对应的若干第二显存需求,并基于所述预设显存区间进行显存组合,以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间之后,还包括:

[0031] 若所述若干第一显存分配空间以及所述若干第二显存分配空间中存在已被释放的若干目标空间,则将所述若干目标空间标记为已释放状态,以得到若干已释放空间;

[0032] 对所述若干已释放空间进行合并,以得到合并后已释放空间;

[0033] 若接收到新的显存申请,则基于所述合并后已释放空间为所述新的现存申请进行显存匹配。

[0034] 第二方面,本申请公开了一种显存管理装置,包括:

[0035] 显存消耗量确定模块,用于通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请,并根据所述若干显存申请确定最大显存消耗量;

[0036] 显存申请区分模块,用于基于所述最大显存消耗量确定所述目标预训练模型对应的可用显存,并根据预设显存请求阈值对所述若干显存申请进行区分,以得到显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请;

[0037] 第一显存分配模块,用于将所述若干第一显存申请对应的若干第一显存需求与预设显存区间进行匹配,以确定所述若干第一显存需求对应的若干第一目标显存区间,并基于所述若干第一目标显存区间对应的区间最大值从所述可用显存中为所述若干第一显存申请匹配对应的若干第一显存分配空间;

[0038] 第二显存分配模块,用于确定所述若干第二显存申请对应的若干第二显存需求,并基于所述预设显存区间进行显存组合,以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间。

[0039] 第三方面,本申请公开了一种电子设备,包括:

[0040] 存储器,用于保存计算机程序;

[0041] 处理器,用于执行所述计算机程序,以实现如前述的显存管理方法。

[0042] 第四方面,本申请公开了一种计算机可读存储介质,用于保存计算机程序,其中,所述计算机程序被处理器执行时实现前述的显存管理方法。

[0043] 本申请中,首先通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请,并根据所述若干显存申请确定最大显存消耗量,然后基于所述最大显存消耗量确定所述目标预训练模型对应的可用显存,并根据预设显存请求阈值对所述若干显存申请进行区分,以得到显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请,最后将所述若干第一显存申请对应的若干第一显存需求与预设显存区间进行匹配,以确定所述若干第一显存需求对应的若干第

一目标显存区间,并基于所述若干第一目标显存区间对应的区间最大值从所述可用显存中为所述若干第一显存申请匹配对应的若干第一显存分配空间,同时确定所述若干第二显存申请对应的若干第二显存需求,并基于所述预设显存区间进行显存组合,以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间。

[0044] 由此可见,通过本申请的方法可以通过预设记录函数来统计目标预训练模型在模型推理过程中的若干显存申请,并根据确定的若干现存申请确定最大显存消耗量,然后可以根据最大显存消耗量确定相应的可用显存,并根据预设显存请求阈值区分出显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请。然后可以将若干第一显存申请对应的若干第一显存需求与预设显存区间进行匹配,以确定相应的若干第一目标显存区间,并从可用显存中为第一显存申请匹配与若干第一目标显存区间对应的区间最大值相应的第一显存分配空间;在另一种情况下,可以确定若干第二显存申请对应的若干第二显存需求,并基于预设显存区间进行显存组合,以根据得到的显存组合从可用显存中为所述若干第二显存需求匹配对应的第二显存分配空间。这样一来,可以根据统计到的信息,为后续显存使用统一分配空间,根据各个显存需求的特点,将统一分配的大块显存切成不同的显存模块以便于使用,将显存碎片化保持在低水平,避免产生内存碎片,进而提高模型推理性能。

附图说明

[0045] 为了更清楚地说明本发明实施例,下面将对实施例中所需要使用的附图做简单的介绍,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的前提下,还可以根据这些附图获得其他的附图。

[0046] 图1为本发明实施例提供的一种显存管理方法流程图;

[0047] 图2为本发明实施例提供的一种具体的显存管理方法流程图;

[0048] 图3为本发明实施例提供的一种显存管理装置结构示意图;

[0049] 图4为本发明实施例提供的一种电子设备结构图。

具体实施方式

[0050] 下面将结合本发明实施例中的附图,对本发明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发明一部分实施例,而不是全部实施例。基于本发明中的实施例,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,都属于本发明保护范围。

[0051] 本发明的说明书及上述附图中的术语“包括”和“具有”,以及与“包括”和“具有”相关的任何变形,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、产品或设备没有限定于已列出的步骤或单元,而是可包括没有列出的步骤或单元。

[0052] 现有技术中,分配大内存块不容易控制大小,分配过大,容易产生很多内存碎片,影响计算利用率,分配过小,则需要进行频繁的内存申请,对性能影响较大。

[0053] 为了克服上述技术问题,本申请公开了一种显存管理方法、装置、设备及存储介质,可以尽可能减少内存的频繁申请和释放,避免产生内存碎片,提升推理性能。

[0054] 为了使本技术领域的人员更好地理解本发明方案,下面结合附图和具体实施方式

对本发明作进一步的详细说明。

[0055] 参见图1所示,本发明实施例公开了一种显存管理方法,包括:

[0056] 步骤S11、通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请,并根据所述若干显存申请确定最大显存消耗量。

[0057] 本实施例中,需要通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请,并根据所述若干显存申请确定最大显存消耗量,具体的,需要在目标预训练模型插入预设记录函数,在本实施例中,目标预训练模型为vLLM模型,且vLLM模型包括graph和eager两种模式,graph模式基于cuda graph技术,以计算图的方式构建大模型,编译之后执行,该模式推理延迟低,但显存消耗大,eager模式类似于PyTorch逐算子执行,该模式推理延迟相对较高,但具有即时执行,易于调试,灵活性强的特点,且显存消耗小,对于新的硬件更易于接入。需要进行说明的是,vLLM模型在 eager模式在进行kv cache内存分配前会进行profile run模型推理,因此可以在PyTorch底层调用板卡runtime接口分配显存和释放显存的位置插入预设记录函数,然后通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请,进而确定实际发生的显存申请和释放情况,例如,在模型推理过程中发生4次32字节申请,10次1024字节申请,2次16字节释放。

[0058] 进一步需要进行说明的是在确定了若干显存申请之后,需要确定若干显存申请分别对应的若干显存需求,然后需要根据确定的若干显存需求确定最大显存消耗量,具体的,模型的推理过程是根据模型支持的最长输入来推理的,因此统计的显存使用量就是模型实际推理时所能达到的最大显存消耗量。因此可以根据确定的若干显存需求计算出推理过程峰值显存的使用量,并将推理过程峰值显存的使用量确定为最大显存消耗量。这样一来,可以记录函数统计的若干显存申请确定最大显存消耗量,进而保证为预训练模型分配的可用显存的准确度。

[0059] 步骤S12、基于所述最大显存消耗量确定所述目标预训练模型对应的可用显存,并根据预设显存请求阈值对所述若干显存申请进行区分,以得到显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请。

[0060] 本实施例中,需要根据确定的最大显存消耗量确定目标预训练模型对应的可用显存,具体的,需要确定预设总显存与预设使用率系数的乘积,然后将得到的乘积的对应数值减去最大显存消耗量的对应数值,以得到所述乘积对应数值以及所述最大显存消耗量对应数值的差值,然后将得到的差值作为目标预训练模型对应的可用显存的数值。

[0061] 进一步的,需要根据显存请求阈值对若干显存申请进行区分,由于当显存申请对应的显存需求比较小时,可以直接根据将显存需求所在区间的最大值为该显存申请匹配显存分配空间,而当显存申请对应的显存需求比较大时,则需要对显存进行组合,以为该申请匹配显存分配空间,避免造成显存浪费的情况,因此需要据显存请求阈值对若干显存申请进行区分,在本实施例中,将显存请求阈值设置为1MB,然后对若干显存申请分别对应的若干显存需求以及预设显存请求阈值进行对比,以根据对比结果从若干显存申请中筛选出显存需求不大于预设显存请求阈值的若干第一显存申请以及显存需求大于预设显存请求阈值的若干第二显存申请。这样一来,可以根据预先设定的显存请求阈值对显存申请进行区分,以便在显存分配的过程中进行利用率更高的显存分配。

[0062] 步骤S13、将所述若干第一显存申请对应的若干第一显存需求与预设显存区间进行匹配,以确定所述若干第一显存需求对应的若干第一目标显存区间,并基于所述若干第一目标显存区间对应的区间最大值从所述可用显存中为所述若干第一显存申请匹配对应的若干第一显存分配空间。

[0063] 本实施例中,需要根据确定的第一显存申请进行显存分配,具体的,需要确定若干第一显存申请对应的若干第一显存需求,并对若干第一显存需求中的当前第一显存需求与预设显存区间进行匹配,以确定当前第一显存需求所匹配的第一目标显存区间,由于第一显存申请对应的第一显存需求均小于1MB,且可能存在多个第一显存申请,因此可以依次对若干第一显存申请进行匹配,且预设显存区间为预先设定好的显存区间,例如1KB及以下为1档,1KB至64KB为一档,64KB到128KB为一档等。在确定了当前第一显存需求所匹配的第一目标显存区间之后,需要确定第一目标显存区间对应的区间最大值,并将区间最大值作为当前第一显存需求对应的目标第一显存需求,例如,当前第一显存需求所匹配的显存区间为1KB至64KB,则直接将区间最大值64KB作为当前第一显存需求对应的目标第一显存需求,然后再进行显存分配时,为第一显存需求分配64KB的可用空间,最后将下一第一显存需求作为当前第一显存需求,并跳转至所述对所述若干第一显存需求中的当前第一显存需求与预设显存区间进行匹配,以确定所述当前第一显存需求所匹配的第一目标显存区间的步骤,以进行下一次第一显存需求的显存分配,直至若干第一显存需求均从所述可用显存中成功匹配相应的第一显存分配空间,例如,有20个第一显存需求,且20个第一显存需求均小于1KB,则开辟20KB,分别为20个第一显存需求对应的第一显存申请分配1KB。这样一来,由于第一显存申请对应的显存需求较小,在分配过程中即使有一定的碎片,也不会造成过大的浪费,并且直接将显存区间对应的区间最大值作为当前第一显存需求对应的目标第一显存需求,可以有效节省分配时的运算量。

[0064] 步骤S14、确定所述若干第二显存申请对应的若干第二显存需求,并基于所述预设显存区间进行显存组合,以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间。

[0065] 本实施例中,需要根据确定的第二显存申请进行显存分配,具体的,需要确定若干第二显存申请对应的若干第二显存需求,并基于预设显存区间进行与若干第二显存需求中当前第二显存需求相应的显存组合,以得到满足当前第二显存需求的目标显存组合,由于第二显存申请对应的第一显存需求均不小于1MB,且可能存在多个第二显存申请,且当显存需求位于1MB以上这一档时,不能继续按层级上限进行分配,因为此时单个显存请求较大,仍然按照上限分配将造成严重的碎片化,显存利用率将变得很低,因此在1MB以上这个层级,将依次进行组合分配,需要从可用显存中为当前第二显存需求匹配与目标显存组合对应的第二显存分配空间,例如将一个64MB和一个30MB,54MB,组合到一起形成一个128MB的一次分配,并且这里仍然允许有一定的冗余空间,且考虑地址对齐能一定程度上提高推理性能,在大层级上分配时会有较多的组合可能,为了尽量减少碎片化,应当优先组合大的显存需求,例如统计到的最大单次显存需求为5GB,那么这个层级的上限是8GB,因此需要将1MB以上的需求按从大到小来和这个5GB组合填到这个8GB分配里,再从剩下的分配需求里根据此原则进行组合;最后需要将下一第二显存需求作为当前第二显存需求,并跳转至所述基于所述预设显存区间进行与所述若干第二显存需求中当前第二显存需求相应的显存

组合,以得到满足所述当前第二显存需求的目标显存组合的步骤,直至若干第二显存需求均从所述可用显存中成功匹配相应的第二显存分配空间,以得到若干第二显存分配空间。这样一来,满足了显存使用需求的同时,还尽可能减少了碎片化的影响。

[0066] 进一步需要进行说明的是,还需要计算若干第一显存分配空间以及若干第二显存分配空间对应的起始偏移量,以基于所述起始偏移量进行数据读取,例如,1KB有20次,那么在1KB档位就在起始地址到20KB这个区间,1KB-64KB档位则在20KB至724KB区间,依次类推,后续当pytorch框架需要显存时将会根据所需大小在这些块间选择,并记录下来。

[0067] 由此可见,通过本申请的方法可以通过预设记录函数来统计目标预训练模型在模型推理过程中的若干显存申请,并根据确定的若干现存申请确定最大显存消耗量,然后可以根据最大显存消耗量确定相应的可用显存,并根据预设显存请求阈值区分出显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请。然后可以将若干第一显存申请对应的若干第一显存需求与预设显存区间进行匹配,以确定相应的若干第一目标显存区间,并从可用显存中为第一显存申请匹配与若干第一目标显存区间对应的区间最大值相应的第一显存分配空间;在另一种情况下,可以确定若干第二显存申请对应的若干第二显存需求,并基于预设显存区间进行显存组合,以根据得到的显存组合从可用显存中为所述若干第二显存需求匹配对应的第二显存分配空间。这样一来,可以根据统计到的信息,为后续显存使用统一分配空间,根据各个显存需求的特点,将统一分配的大块显存切成不同的显存模块以便于使用,将显存碎片化保持在低水平,避免产生内存碎片,进而提高模型推理性能。

[0068] 基于前述实施例可知,在进行内存分配的同时,还可能存在内存释放的情况,为此,本实施例对如何进行内存释放以及内存的再利用进行了详细的说明。参见图2所示,本发明实施例公开了一种显存管理方法,包括:

[0069] 步骤S21、如果若干第一显存分配空间以及若干第二显存分配空间中存在已被释放的若干目标空间,则将所述若干目标空间标记为已释放状态,以得到若干已释放空间。

[0070] 本实施例中,如果发现若干第一显存分配空间以及若干第二显存分配空间中存在已被释放的若干目标空间,则需要对已被释放的若干目标空间进行标记,将其标记为已释放状态,需要进行说明的是,pytorch通常在一个作用于结束或在显示调用某些接口时进行内存的释放,但需要进行说明的是,因此在pytorch端释放时在底层内存不会真正的释放,而是给每个被占用的内存做一个标记,将其标记为已释放状态,进而得到若干已释放空间。

[0071] 步骤S22、对所述若干已释放空间进行合并,以得到合并后已释放空间。

[0072] 本实施例中,需要对得到的若干已释放内存进行合并,将其合并为一个大的内存,以得到合并后已释放空间,并且需要以每个被占用内存的起始地址和长度为键值构建一个map映射,如果检测到pytorch调用了接口来释放这块内存,则将起标记为已释放状态,并对当前所有的已释放状态的内存尽行合并。

[0073] 步骤S23、若接收到新的显存申请,则基于所述合并后已释放空间为所述新的现存申请进行显存匹配。

[0074] 本实施例中,如果接收到新的显存申请,则需要根据构建的map映射寻找相应的合并后已释放空间为新的现存申请进行显存匹配,并在最终退出模型推理时一次性释放所有的合并后已释放空间。

[0075] 需要进行说明的是,如果接收到新的显存申请,还需要确定新的显存申请对应的显存数量,因为可能在短时间内存在多个新的显存申请,因此可以构建一个显存申请队列,然后按照时间顺序将接收到的新的显存申请添加至显存申请队列中,然后根据队列中的顺序依次为新的显存申请进行显存匹配,在为新的显存申请进行显存匹配时,需要确定新的显存申请对应的显存需求,如果显存需求大于当前的合并后已释放空间,则当前的合并后已释放空间无法满足当前新的显存申请对应的显存需求,因此还需要设置一个待匹配显存申请队列,当存在上述当前的合并后已释放空间无法满足当前新的显存申请对应的显存需求的情况时,可以将当前的新的显存申请转移至待匹配显存申请队列中,然后为显存申请队列中的下一个新的显存申请进行显存匹配,并在匹配的同时依次将待匹配显存申请队列的显存申请重新添加至显存申请队列中。这样一来,可以保证合并后已释放空间优先满足条件符合的显存申请,避免造成显存浪费的情况。

[0076] 由此可见,本实施例中,如果若干第一显存分配空间以及若干第二显存分配空间中存在已被释放的若干目标空间,则将所述若干目标空间标记为已释放状态,以得到若干已释放空间,然后可以对所述若干已释放空间进行合并,以得到合并后已释放空间,最后,如果接收到新的显存申请,则基于所述合并后已释放空间为所述新的现存申请进行显存匹配。这样一来,可以减少调用接口进行显存开辟和释放的次数,提升了推理性能,并且充分考虑了显存复用的情况,进而降低了显存碎片化的问题。

[0077] 参见图3所示,本发明实施例公开了一种显存管理装置,包括:

[0078] 显存消耗量确定模块11,用于通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请,并根据所述若干显存申请确定最大显存消耗量;

[0079] 显存申请区分模块12,用于基于所述最大显存消耗量确定所述目标预训练模型对应的可用显存,并根据预设显存请求阈值对所述若干显存申请进行区分,以得到显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请;

[0080] 第一显存分配模块13,用于将所述若干第一显存申请对应的若干第一显存需求与预设显存区间进行匹配,以确定所述若干第一显存需求对应的若干第一目标显存区间,并基于所述若干第一目标显存区间对应的区间最大值从所述可用显存中为所述若干第一显存申请匹配对应的若干第一显存分配空间;

[0081] 第二显存分配模块14,用于确定所述若干第二显存申请对应的若干第二显存需求,并基于所述预设显存区间进行显存组合,以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间。

[0082] 在一些实施例中,所述显存消耗量确定模块11,具体可以包括:

[0083] 显存申请统计单元,用于在目标预训练模型插入预设记录函数,并通过所述预设记录函数统计所述目标预训练模型在模型推理过程中的若干显存申请;

[0084] 最大显存消耗量确定单元,用于统计所述若干显存申请分别对应的若干显存需求,并基于所述若干显存需求确定最大显存消耗量。

[0085] 在一些实施例中,所述显存申请区分模块12,具体可以包括:

[0086] 可用显存确定单元,用于确定预设总显存与预设使用率系数的乘积,并计算所述乘积对应数值以及所述最大显存消耗量对应数值的差值,以将得到的所述差值作为所述目

标预训练模型对应的可用显存；

[0087] 阈值对比单元，用于对所述若干显存申请分别对应的若干显存需求以及预设显存请求阈值进行对比，以得到相应的对比结果；

[0088] 显存申请区分单元，用于根据所述对比结果从所述若干显存申请中筛选出显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请。

[0089] 在一些实施例中，所述第一显存分配模块13，具体可以包括：

[0090] 第一显存需求确定单元，用于确定所述若干第一显存申请对应的若干第一显存需求；

[0091] 第一显存区间匹配单元，用于对所述若干第一显存需求中的当前第一显存需求与预设显存区间进行匹配，以确定所述当前第一显存需求所匹配的第一目标显存区间；

[0092] 目标显存需求确定单元，用于确定所述第一目标显存区间对应的区间最大值，并将所述区间最大值作为所述当前第一显存需求对应的目标第一显存需求；

[0093] 第一显存分配空间确定单元，用于从所述可用显存中为所述当前第一显存需求匹配与所述目标第一显存需求对应的第一显存分配空间；

[0094] 第一步骤跳转单元，用于将下一第一显存需求作为当前第一显存需求，并跳转至所述对所述若干第一显存需求中的当前第一显存需求与预设显存区间进行匹配，以确定所述当前第一显存需求所匹配的第一目标显存区间的步骤，直至所述若干第一显存需求均从所述可用显存中成功匹配相应的第一显存分配空间，以得到若干第一显存分配空间。

[0095] 在一些实施例中，所述第二显存分配模块14，具体可以包括：

[0096] 第二显存需求确定单元，用于确定所述若干第二显存申请对应的若干第二显存需求；

[0097] 显存组合确定单元，用于基于所述预设显存区间进行与所述若干第二显存需求中当前第二显存需求相应的显存组合，以得到满足所述当前第二显存需求的目标显存组合；

[0098] 第二显存分配空间确定单元，用于从所述可用显存中为所述当前第二显存需求匹配与所述目标显存组合对应的第二显存分配空间；

[0099] 第二步骤跳转单元，用于将下一第二显存需求作为当前第二显存需求，并跳转至所述基于所述预设显存区间进行与所述若干第二显存需求中当前第二显存需求相应的显存组合，以得到满足所述当前第二显存需求的目标显存组合的步骤，直至所述若干第二显存需求均从所述可用显存中成功匹配相应的第二显存分配空间，以得到若干第二显存分配空间。

[0100] 在一些实施例中，所述显存管理装置，还可以包括：

[0101] 数据读取单元，用于计算所述若干第一显存分配空间以及所述若干第二显存分配空间对应的起始偏移量，以基于所述起始偏移量进行数据读取。

[0102] 在一些实施例中，所述显存管理装置，还可以包括：

[0103] 控件状态标记单元，用于若所述若干第一显存分配空间以及所述若干第二显存分配空间中存在已被释放的若干目标空间，则将所述若干目标空间标记为已释放状态，以得到若干已释放空间；

[0104] 空间合并单元，用于对所述若干已释放空间进行合并，以得到合并后已释放空间；

[0105] 显存匹配单元,用于若接收到新的显存申请,则基于所述合并后已释放空间为所述新的现存申请进行显存匹配。

[0106] 进一步的,本申请实施例还公开了一种电子设备,图4是根据一示例性实施例示出的电子设备结构图,图中的内容不能认为是对本申请的使用范围的任何限制。该电子设备,具体可以包括:至少一个处理器21、至少一个存储器22、电源23、通信接口24、输入输出接口25和通信总线26。其中,所述存储器22用于存储计算机程序,所述计算机程序由所述处理器21加载并执行,以实现前述任一实施例公开的显存管理方法中的相关步骤。另外,本实施例中的电子设备具体可以为电子计算机。

[0107] 本实施例中,电源23用于为电子设备上的各硬件设备提供工作电压;通信接口24能够为电子设备创建与外界设备之间的数据传输通道,其所遵循的通信协议是能够适用于本申请技术方案的任意通信协议,在此不对其进行具体限定;输入输出接口25,用于获取外界输入数据或向外界输出数据,其具体的接口类型可以根据具体应用需要进行选取,在此不进行具体限定。

[0108] 另外,存储器22作为资源存储的载体,可以是只读存储器、随机存储器、磁盘或者光盘等,其上所存储的资源可以包括操作系统221、计算机程序222等,存储方式可以是短暂存储或者永久存储。

[0109] 其中,操作系统221用于管理与控制电子设备上的各硬件设备以及计算机程序222,其可以是Windows Server、Netware、Unix、Linux等。计算机程序222除了包括能够用于完成前述任一实施例公开的由电子设备执行的显存管理方法的计算机程序之外,还可以进一步包括能够用于完成其他特定工作的计算机程序。

[0110] 进一步的,本申请还公开了一种计算机可读存储介质,用于存储计算机程序;其中,所述计算机程序被处理器执行时实现前述公开的显存管理方法。关于该方法的具体步骤可以参考前述实施例中公开的相应内容,在此不再进行赘述。

[0111] 本说明书中各个实施例采用递进的方式描述,每个实施例重点说明的都是与其它实施例的不同之处,各个实施例之间相同或相似部分互相参见即可。对于实施例公开的装置而言,由于其与实施例公开的方法相对应,所以描述的比较简单,相关之处参见方法部分说明即可。

[0112] 专业人员还可以进一步意识到,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、计算机软件或者二者的结合来实现,为了清楚地说明硬件和软件的可互换性,在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。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。

[0113] 结合本文中所公开的实施例描述的方法或算法的步骤可以直接用硬件、处理器执行的软件模块,或者二者的结合来实施。软件模块可以置于随机存储器(RAM)、内存、只读存储器(ROM)、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

[0114] 最后,还需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作

之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0115] 以上对本申请所提供的技术方案进行了详细介绍,本文中应用了具体个例对本申请的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本申请的方法及其核心思想;同时,对于本领域的一般技术人员,依据本申请的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本申请的限制。

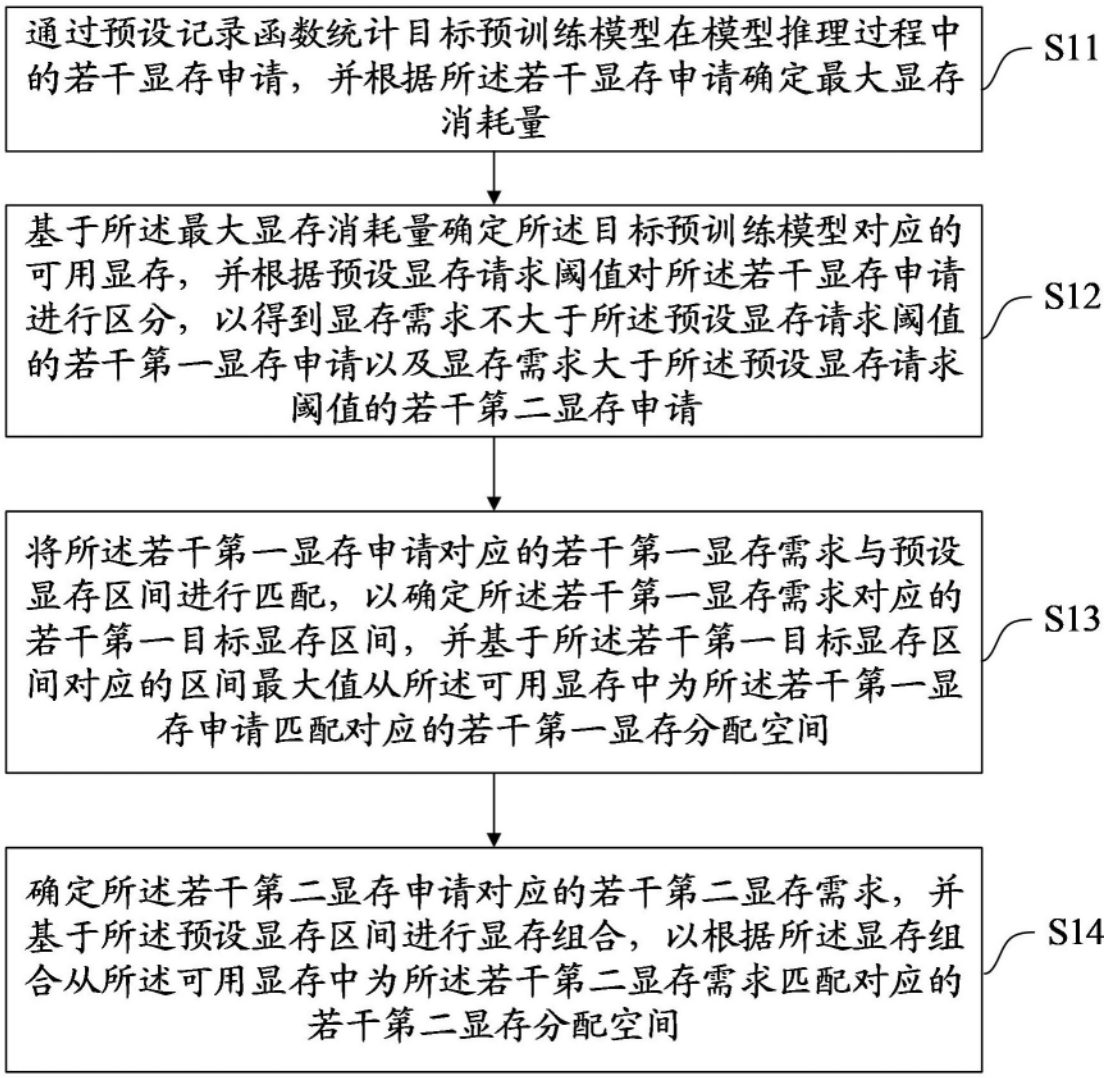

通过预设记录函数统计目标预训练模型在模型推理过程中的若干显存申请，并根据所述若干显存申请确定最大显存消耗量
S11
基于所述最大显存消耗量确定所述目标预训练模型对应的可用显存，并根据预设显存请求阈值对所述若干显存申请进行区分，以得到显存需求不大于所述预设显存请求阈值的若干第一显存申请以及显存需求大于所述预设显存请求阈值的若干第二显存申请
S12
将所述若干第一显存申请对应的若干第一显存需求与预设显存区间进行匹配，以确定所述若干第一显存需求对应的若干第一目标显存区间，并基于所述若干第一目标显存区间对应的区间最大值从所述可用显存中为所述若干第一显存申请匹配对应的若干第一显存分配空间
S13
确定所述若干第二显存申请对应的若干第二显存需求，并基于所述预设显存区间进行显存组合，以根据所述显存组合从所述可用显存中为所述若干第二显存需求匹配对应的若干第二显存分配空间
S14

图1

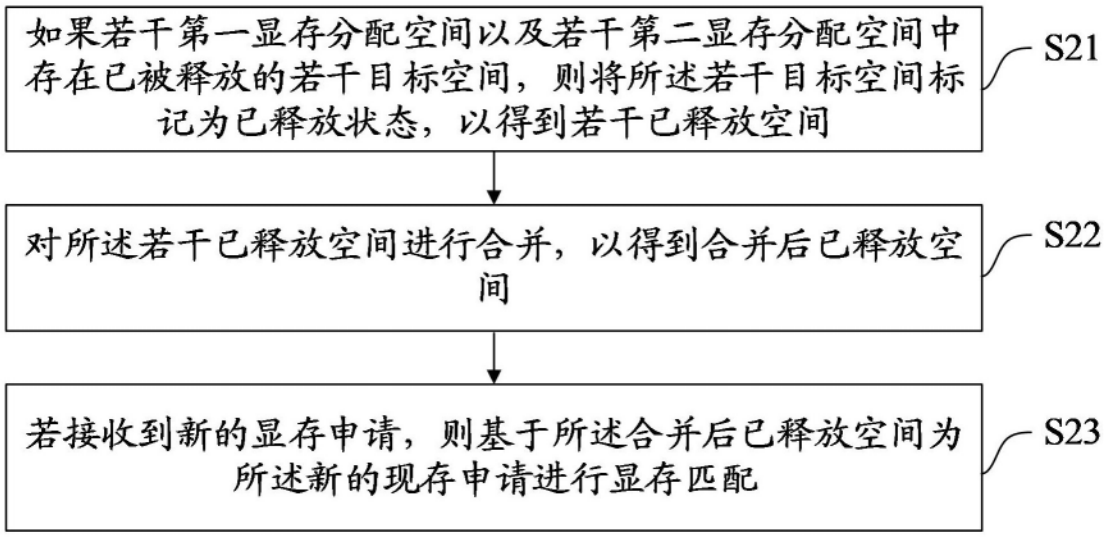

如果若干第一显存分配空间以及若干第二显存分配空间中存在已被释放的若干目标空间，则将所述若干目标空间标记为已释放状态，以得到若干已释放空间
S21
对所述若干已释放空间进行合并，以得到合并后已释放空间
S22
若接收到新的显存申请，则基于所述合并后已释放空间为所述新的现存申请进行显存匹配
S23

图2

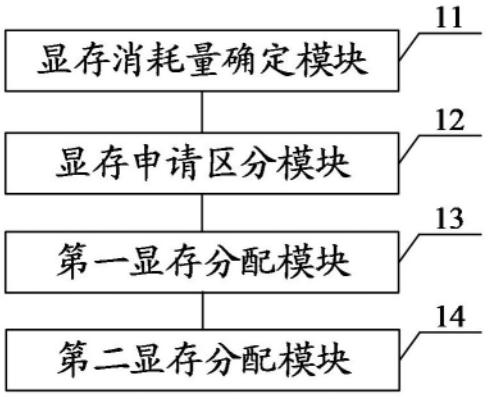

11
显存消耗量确定模块
12
显存申请区分模块
13
第一显存分配模块
14
第二显存分配模块

图3

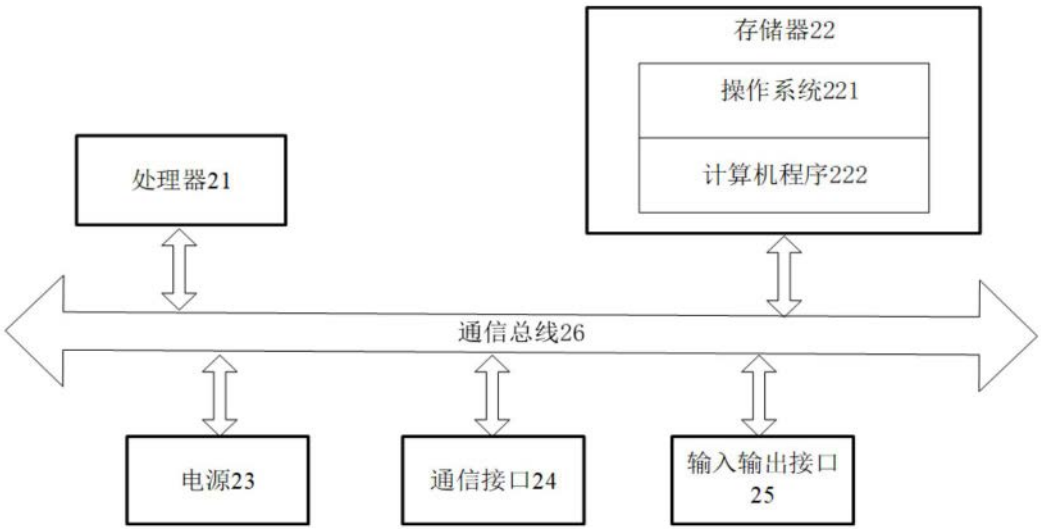


图4