(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116681113 A
(43)申请公布日 2023.09.01

(21)申请号 202310714591.8

(22)申请日 2023.06.15

(71)申请人 京东科技信息技术有限公司
地址 100176 北京市大兴区北京经济技术开发区科创十一街18号院2号楼6层601

(72)发明人 张振羽 陈蒙 宋阳

(74)专利代理机构 北京律智知识产权代理有限公司 11438
专利代理师 陈媛婧

(51)Int.Cl.
G06N 3/0495(2023.01)
G06N 3/048(2023.01)
G06F 17/16(2006.01)
G06F 18/241(2023.01)

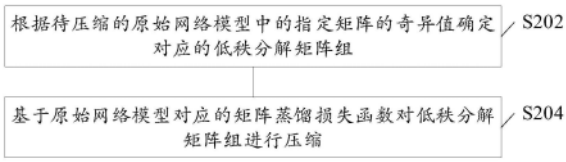
权利要求书2页 说明书11页 附图5页

(54)发明名称

网络模型压缩方法、装置、电子设备和可读介质

(57)摘要

本公开提供一种网络模型压缩方法、装置、电子设备和可读介质，其中，网络模型压缩方法包括：根据待压缩的原始网络模型的指定矩阵的奇异值确定对应的低秩分解矩阵组；基于原始网络模型对应的矩阵蒸馏损失函数对低秩分解矩阵组进行压缩。通过本公开实施例，降低了网络模型的体量，在保证网络模型的计算准确性的前提下，提升了网络模型的计算速率。



CN 116681113 A

1.一种网络模型压缩方法,其特征在于,包括:

根据待压缩的原始网络模型的指定矩阵的奇异值确定对应的低秩分解矩阵组;

基于所述原始网络模型对应的矩阵蒸馏损失函数对所述低秩分解矩阵组进行压缩。

2.如权利要求1所述的网络模型压缩方法,其特征在于,根据待压缩的原始网络模型的指定矩阵的奇异值确定对应的低秩分解矩阵组包括:

通过奇异值分解算法确定所述原始网络模型中的指定矩阵的奇异值;

根据所述奇异值的个数和预设的超参数确定所述低秩分解矩阵组的第一维度参数;

根据所述原始网络模型确定奇异值矩阵的维度;

通过正态分布随机初始化第一低维矩阵和第二低维矩阵,并构建所述奇异值矩阵,所述低秩分解矩阵组包括所述第一低维矩阵、所述第二低维矩阵和所述奇异值矩阵。

3.如权利要求2所述的网络模型压缩方法,其特征在于,根据所述原始网络模型确定奇异值矩阵的维度包括:

确定所述原始网络模型中的指定矩阵的第二维度参数和第三维度参数;

根据所述第一维度参数和所述第二维度参数确定所述第一低维矩阵的维度;

根据所述第二维度参数和所述第三维度参数确定所述第二低维矩阵的维度;

根据所述第二维度参数确定奇异值矩阵的维度。

4.如权利要求3所述的网络模型压缩方法,其特征在于,根据所述奇异值的个数和预设的超参数确定所述低秩分解矩阵组的第一维度参数包括:

根据所述奇异值的个数和所述超参数确定所述第一维度参数的初始值;

根据所述原始网络模型中的指定矩阵的维度确定所述第一维度参数的目标值;

针对所述第一维度参数的初始值创建迭代参数;

根据所述迭代参数对所述第一维度参数的初始值进行迭代计算,至达到所述第一维度参数的目标值为止。

5.如权利要求4所述的网络模型压缩方法,其特征在于,根据所述原始网络模型中的指定矩阵的维度确定所述第一维度参数的目标值包括:

确定所述原始网络模型中的指定矩阵的维度的最小值;

根据所述最小值与预设比例值确定所述第一维度参数的目标值。

6.如权利要求3所述的网络模型压缩方法,其特征在于,基于所述原始网络模型对应的矩阵蒸馏损失函数对所述低秩分解矩阵组进行压缩包括:

将样本输入至所述原始网络模型对应的矩阵,并确定所述原始网络模型的输出结果,记作第一输出结果;

将所述样本输入至待压缩的低秩分解矩阵组,并确定所述低秩分解矩阵组的输出结果,记作第二输出结果;

根据所述第一输出结果和所述第二输出结果确定输出损失;

确定所述奇异值矩阵对应的正则损失;

确定所述原始网络模型对所述样本的处理损失;

根据所述输出损失、所述正则损失和所述处理损失对所述低秩分解矩阵组进行压缩。

7.如权利要求6所述的网络模型压缩方法,其特征在于,确定所述奇异值矩阵对应的正则损失包括:

基于L0优化算法计算所述奇异值矩阵对应的L0范式的正则损失。

8.如权利要求6或7所述的网络模型压缩方法，其特征在于，确定所述原始网络模型对所述样本的处理损失包括：

确定所述样本的处理任务为分类任务；

确定所述原始网络模型对所述分类任务处理后的分类损失，所述处理损失包括所述分类损失。

9.一种网络模型压缩装置，其特征在于，包括：

确定模块，设置为根据待压缩的原始网络模型的指定矩阵的奇异值确定对应的低秩分解矩阵组；

压缩模块，设置为基于所述原始网络模型对应的矩阵蒸馏损失函数对所述低秩分解矩阵组进行压缩。

10.一种电子设备，其特征在于，包括：

存储器；以及

耦合到所述存储器的处理器，所述处理器被配置为基于存储在所述存储器中的指令，执行如权利要求1-8任一项所述的网络模型压缩方法。

11.一种计算机可读存储介质，其上存储有程序，该程序被处理器执行时实现如权利要求1-8任一项所述的网络模型压缩方法。

网络模型压缩方法、装置、电子设备和可读介质

技术领域

[0001] 本公开涉及网络技术领域，具体而言，涉及一种网络模型压缩方法、装置、电子设备和可读介质。

背景技术

[0002] 目前，深度学习模型被广泛应用于各类线上任务中，例如文本分类、图片检测和语音识别等等。随着预训练技术的行起，在NLP(自然语言处理)、CV(计算机视觉)、ASR(语音识别)等领域，模型规模日益呈指数级增长，大模型虽然具有更高的准确率且能适配多种下游任务，但是其参数量巨大且计算速度缓慢，需要大量的GPU计算资源，运算效率低下，且数据维护成本高昂，对于模型的高效推理造成了严重的挑战，阻碍了大模型的在线部署和工业化落地。

[0003] 需要说明的是，在上述背景技术部分公开的信息仅用于加强对本公开的背景的理解，因此可以包括不构成对本领域普通技术人员已知的现有技术的信息。

发明内容

[0004] 本公开的目的在于提供一种网络模型压缩方法、装置、电子设备和可读介质，用于至少在一定程度上克服由于相关技术的限制和缺陷而导致的网络模型体量大的问题。

[0005] 根据本公开实施例的第一方面，提供一种网络模型压缩方法，包括：根据待压缩的原始网络模型的指定矩阵的奇异值确定对应的低秩分解矩阵组；基于所述原始网络模型对应的矩阵蒸馏损失函数对所述低秩分解矩阵组进行压缩。

[0006] 在本公开的一种示例性实施例中，根据待压缩的原始网络模型的指定矩阵的奇异值确定对应的低秩分解矩阵组包括：

[0007] 通过奇异值分解算法确定所述原始网络模型中的指定矩阵的奇异值；

[0008] 根据所述奇异值的个数和预设的超参数确定所述低秩分解矩阵组的第一维度参数；

[0009] 根据所述原始网络模型确定奇异值矩阵的维度；

[0010] 通过正态分布随机初始化第一低维矩阵和第二低维矩阵，并构建所述奇异值矩阵，所述低秩分解矩阵组包括所述第一低维矩阵、所述第二低维矩阵和所述奇异值矩阵。

[0011] 在本公开的一种示例性实施例中，根据所述原始网络模型确定奇异值矩阵的维度包括：

[0012] 确定所述原始网络模型中的指定矩阵的第二维度参数和第三维度参数；

[0013] 根据所述第一维度参数和所述第二维度参数确定所述第一低维矩阵的维度；

[0014] 根据所述第二维度参数和所述第三维度参数确定所述第二低维矩阵的维度；

[0015] 根据所述第二维度参数确定奇异值矩阵的维度。

[0016] 在本公开的一种示例性实施例中，根据所述奇异值的个数和预设的超参数确定所述低秩分解矩阵组的第一维度参数包括：

[0017] 根据所述奇异值的个数和所述超参数确定所述第一维度参数的初始值；

[0018] 根据所述原始网络模型中的指定矩阵的维度确定所述第一维度参数的目标值；

[0019] 针对所述第一维度参数的初始值创建迭代参数；

[0020] 根据所述迭代参数对所述第一维度参数的初始值进行迭代计算，至达到所述第一维度参数的目标值为止。

[0021] 在本公开的一种示例性实施例中，根据所述原始网络模型中的指定矩阵的维度确定所述第一维度参数的目标值包括：

[0022] 确定所述原始网络模型中的指定矩阵的维度的最小值；

[0023] 根据所述最小值与预设比例值确定所述第一维度参数的目标值。

[0024] 在本公开的一种示例性实施例中，基于所述原始网络模型对应的矩阵蒸馏损失函数对所述低秩分解矩阵组进行压缩包括：

[0025] 将样本输入至所述原始网络模型对应的矩阵，并确定所述原始网络模型的输出结果，记作第一输出结果；

[0026] 将所述样本输入至待压缩的低秩分解矩阵组，并确定所述低秩分解矩阵组的输出结果，记作第二输出结果；

[0027] 根据所述第一输出结果和所述第二输出结果确定输出损失；

[0028] 确定所述奇异值矩阵对应的正则损失；

[0029] 确定所述原始网络模型对所述样本的处理损失；

[0030] 根据所述输出损失、所述正则损失和所述处理损失对所述低秩分解矩阵组进行压缩。

[0031] 在本公开的一种示例性实施例中，确定所述奇异值矩阵对应的正则损失包括：

[0032] 基于L0优化算法计算所述奇异值矩阵对应的L0范式的正则损失。

[0033] 在本公开的一种示例性实施例中，确定所述原始网络模型对所述样本的处理损失包括：

[0034] 确定所述样本的处理任务为分类任务；

[0035] 确定所述原始网络模型对所述分类任务处理后的分类损失，所述处理损失包括所述分类损失。

[0036] 根据本公开实施例的第二方面，提供一种网络模型压缩装置，包括：

[0037] 确定模块，设置为根据待压缩的原始网络模型的指定矩阵的奇异值确定对应的低秩分解矩阵组；

[0038] 压缩模块，设置为基于所述原始网络模型对应的矩阵蒸馏损失函数对所述低秩分解矩阵组进行压缩。

[0039] 根据本公开的第三方面，提供一种电子设备，包括：存储器；以及耦合到所述存储器的处理器，所述处理器被配置为基于存储在所述存储器中的指令，执行如上述任意一项所述的方法。

[0040] 根据本公开的第四方面，提供一种计算机可读存储介质，其上存储有程序，该程序被处理器执行时实现如上述任意一项所述的网络模型压缩方法。

[0041] 本公开实施例，通过根据待压缩的原始网络模型的指定矩阵的奇异值确定对应的低秩分解矩阵组，并基于原始网络模型对应的矩阵蒸馏损失函数对低秩分解矩阵组进行压

缩，降低了网络模型的体量，在保证网络模型的计算准确性的前提下，提升了网络模型的计算速率。

[0042] 应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本公开。

附图说明

[0043] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本公开的实施例，并与说明书一起用于解释本公开的原理。显而易见地，下面描述中的附图仅仅是本公开的一些实施例，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。

[0044] 图1示出了可以应用本发明实施例的网络模型压缩方案的示例性系统架构的示意图；

[0045] 图2是本公开示例性实施例中一种网络模型压缩方法的流程图；

[0046] 图3是本公开示例性实施例中另一种网络模型压缩方法的流程图；

[0047] 图4是本公开示例性实施例中另一种网络模型压缩方法的流程图；

[0048] 图5是本公开示例性实施例中另一种网络模型压缩方法的流程图；

[0049] 图6是本公开示例性实施例中另一种网络模型压缩方法的流程图；

[0050] 图7是本公开示例性实施例中另一种网络模型压缩方法的流程图；

[0051] 图8是本公开示例性实施例中另一种网络模型压缩方法的流程图；

[0052] 图9是本公开示例性实施例中另一种网络模型压缩方法的流程图；

[0053] 图10是本公开示例性实施例中一种网络模型压缩方案的架构示意图；

[0054] 图11是本公开示例性实施例中一种网络模型压缩装置的方框图；

[0055] 图12是本公开示例性实施例中一种电子设备的方框图。

具体实施方式

[0056] 现在将参考附图更全面地描述示例实施方式。然而，示例实施方式能够以多种形式实施，且不应被理解为限于在此阐述的范例；相反，提供这些实施方式使得本公开将更加全面和完整，并将示例实施方式的构思全面地传达给本领域的技术人员。所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施方式中。在下面的描述中，提供许多具体细节从而给出对本公开的实施方式的充分理解。然而，本领域技术人员将意识到，可以实践本公开的技术方案而省略所述特定细节中的一个或更多，或者可以采用其它的方法、组元、装置、步骤等。在其它情况下，不详细示出或描述公知技术方案以避免喧宾夺主而使得本公开的各方面变得模糊。

[0057] 此外，附图仅为本公开的示意性图解，图中相同的附图标记表示相同或类似的部分，因而将省略对它们的重复描述。附图中所示的一些方框图是功能实体，不一定必须与物理或逻辑上独立的实体相对应。可以采用软件形式来实现这些功能实体，或在一个或多个硬件模块或集成电路中实现这些功能实体，或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0058] 图1示出了可以应用本发明实施例的网络模型压缩方案的示例性系统架构的示意

图。

[0059] 如图1所示，系统架构100可以包括终端设备101、102、103中的一种或多种，网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104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。

[0060] 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、网络和服务器。比如服务器105可以是多个服务器组成的服务器集群等。

[0061] 用户可以使用终端设备101、102、103通过网络104与服务器105交互，以接收或发送消息等。终端设备101、102、103可以是具有显示屏的各种电子设备，包括但不限于智能手机、平板电脑、便携式计算机和台式计算机等等。

[0062] 在一些实施例中，本发明实施例所提供的网络模型压缩方法一般由服务器105执行，相应地，网络模型压缩装置一般设置于终端设备103(也可以是终端设备101或102)中。在另一些实施例中，某些终端可以具有与服务器设备相似的功能从而执行本公开。

[0063] 下面结合附图对本公开示例实施方式进行详细说明。

[0064] 图2是本公开示例性实施例中网络模型压缩方法的流程图。

[0065] 参考图2，网络模型压缩方法可以包括：

[0066] 步骤S202，根据待压缩的原始网络模型的指定矩阵的奇异值确定对应的低秩分解矩阵组。

[0067] 步骤S204，基于所述原始网络模型对应的矩阵蒸馏损失函数对所述低秩分解矩阵组进行压缩。

[0068] 本公开实施例，通过根据待压缩的原始网络模型的指定矩阵的奇异值确定对应的低秩分解矩阵组，并基于原始网络模型对应的矩阵蒸馏损失函数对低秩分解矩阵组进行压缩，降低了网络模型的体量，在保证网络模型的计算准确性的前提下，提升了网络模型的计算速率。

[0069] 在本公开的一种示例性实施例中，指定矩阵可以为为权重矩阵。

[0070] 下面，对网络模型压缩方法的各步骤进行详细说明。

[0071] 在本公开的一种示例性实施例中，如图3所示，根据待压缩的原始网络模型的指定矩阵的奇异值确定对应的低秩分解矩阵组包括：

[0072] 步骤S302，通过奇异值分解算法确定所述原始网络模型中的指定矩阵的奇异值。

[0073] 步骤S304，根据所述奇异值的个数和预设的超参数确定所述低秩分解矩阵组的第一维度参数。

[0074] 步骤S306，根据所述原始网络模型确定奇异值矩阵的维度。

[0075] 步骤S308，通过正态分布随机初始化第一低维矩阵和第二低维矩阵，并构建所述奇异值矩阵，所述低秩分解矩阵组包括所述第一低维矩阵、所述第二低维矩阵和所述奇异值矩阵。

[0076] 在本公开的一种示例性实施例中，如图4所示，根据所述原始网络模型确定奇异值矩阵的维度包括：

[0077] 步骤S402，确定所述原始网络模型中的指定矩阵的第二维度参数和第三维度参数。

[0078] 步骤S404,根据所述第一维度参数和所述第二维度参数确定所述第一低维矩阵的维度。

[0079] 步骤S406,根据所述第二维度参数和所述第三维度参数确定所述第二低维矩阵的维度。

[0080] 步骤S408,根据所述第二维度参数确定奇异值矩阵的维度。

[0081] 在本公开的一种示例性实施例中,通过构建奇异值矩阵来降低神经网络模型的计算量,并且通过第一低维矩阵和第二低维矩阵来降低整个计算架构的维度,在替换原始网络模型后,有效地降低了模型体量和数据量。

[0082] 在本公开的一种示例性实施例中,第一低维矩阵和第二低维矩阵都是相对指定矩阵的维度而言的。

[0083] 在本公开的一种示例性实施例中,如图5所示,根据所述奇异值的个数和预设的超参数确定所述低秩分解矩阵组的第一维度参数包括:

[0084] 步骤S502,根据所述奇异值的个数和所述超参数确定所述第一维度参数的初始值。

[0085] 步骤S505,根据所述原始网络模型中的指定矩阵的维度确定所述第一维度参数的目标值。

[0086] 步骤S506,针对所述第一维度参数的初始值创建迭代参数。

[0087] 步骤S508,根据所述迭代参数对所述第一维度参数的初始值进行迭代计算,至达到所述第一维度参数的目标值为止。

[0088] 在本公开的一种示例性实施例中,如图6所示,根据所述原始网络模型中的指定矩阵的维度确定所述第一维度参数的目标值包括:

[0089] 步骤S602,确定所述原始网络模型中的指定矩阵的维度的最小值。

[0090] 步骤S604,根据所述最小值与预设比例值确定所述第一维度参数的目标值。

[0091] 在本公开的一种示例性实施例中,通过确定所述原始网络模型中的指定矩阵的维度的最小值,以及根据所述最小值与预设比例值确定所述第一维度参数的目标值,能够最小化奇异值矩阵,从而降低低秩分解矩阵组的维度,进而降低了原始网络模型的维度。

[0092] 在本公开的一种示例性实施例中,如图7所示,基于所述原始网络模型对应的矩阵蒸馏损失函数对所述低秩分解矩阵组进行压缩包括:

[0093] 步骤S702,将样本输入至所述原始网络模型对应的矩阵,并确定所述原始网络模型的输出结果,记作第一输出结果。

[0094] 步骤S704,将所述样本输入至待压缩的低秩分解矩阵组,并确定所述低秩分解矩阵组的输出结果,记作第二输出结果。

[0095] 步骤S707,根据所述第一输出结果和所述第二输出结果确定输出损失。

[0096] 步骤S708,确定所述奇异值矩阵对应的正则损失。

[0097] 步骤S710,确定所述原始网络模型对所述样本的处理损失。

[0098] 步骤S712,根据所述输出损失、所述正则损失和所述处理损失对所述低秩分解矩阵组进行压缩。

[0099] 在本公开的一种示例性实施例中,通过所述输出损失、所述正则损失和所述处理损失确定总的损失函数,并通过对总的损失函数的迭代至收敛,通过迭代结果对所述低秩

分解矩阵组的参数进行优化，以实现低秩分解矩阵组的压缩。

[0100] 在本公开的一种示例性实施例中，如图8所示，确定所述奇异值矩阵对应的正则损失包括：

[0101] 步骤S802，基于L0优化算法计算所述奇异值矩阵对应的L0范式的正则损失。

[0102] 在本公开的一种示例性实施例中，如图9所示，确定所述原始网络模型对所述样本的处理损失包括：

[0103] 步骤S902，确定所述样本的处理任务为分类任务。

[0104] 步骤S904，确定所述原始网络模型对所述分类任务处理后的分类损失，所述处理损失包括所述分类损失。

[0105] 图10是本公开示例性实施例中一种网络模型压缩方案的架构示意图。

[0106] 如图10所示，本公开示例性实施例的网络模型压缩方案主要包括低秩分解矩阵组1006和矩阵蒸馏损失1008两部分，针对给定的原始模型的指定权重矩阵W 1004，可以通过一组低维矩阵对权重矩阵W 1004进行低参数化的表示。

[0107] 其中，低质分解矩阵组主要包括两个低维矩阵U(维度是d1×r)和V(维度是r×d2)，与一个奇异值矩阵Σ，其最低维度取值为r，主要用于降低模型参数W 1004(维度是d1×d2)和提高计算速度。

[0108] 另外，矩阵蒸馏损失主要用于将原始矩阵所代表得深度网络信息转移到低秩分解矩阵组1006，并不断优化矩阵组的最低维度r，在保证模型准确率的前提下，尽可能地减少r。

[0109] 在本公开示例性实施例的网络模型压缩方案中，确定低秩分解矩阵组1006的步骤包括：

[0110] (1)通过SVD(奇异值分解算法)求出原始模型中矩阵W 1004的奇异值，假设为$\{\lambda_1,\lambda_2,...,\lambda_K\}$，其中$\lambda_k$表征W 1004的第一个奇异值，矩阵W 1004最多包含K个奇异值(K≤min(d1,d2))。

[0111] (2)设置初始化$r_0=\alpha*K$，α为一个超参数，其初始值可以设置为2，使用正太分布随机初始化低质矩阵U和V的每个元素，同时构造对角矩阵Σ=diagonal($\{\lambda_1,\lambda_2,...,\lambda_K\}$)。

[0112] (3)为参数r设计一个线性参数scheduler，其初始值为r_0，最终目标值为r_s<<min(d1,d2)，随着迭代进行，r逐渐减小到给定最终目标值。经过初始化设置可以让待压缩模型(低秩分解矩阵组1006参数)很好的学习适应原始模型和数据中的知识。

[0113] 在本公开的一种实施例中，最终目标值一般根据实际需求人工选定，通常设置为min(d1,d2)的1/30到1/20左右。

[0114] 在本公开的一种实施例中，最终目标值的取值范围为[8,64]。

[0115] 在本公开示例性实施例的网络模型压缩方案中，针对某个具体得下游任务(例如分类任务)，对于一个给定的输入$X\in R^{d1}$，在第t步骤迭代时，r的期望值为r_t，具体可以通过以下步骤对低秩分解矩阵组1006进行压缩：

[0116] (1)对原模型的每个W 1004，计算原模型的输出$f(W)=\sigma(W^TX)$，其中σ是神经网络权重连接对应得激活函数。

[0117] (2)对需要压缩的低秩分解矩阵组1006，低秩分解矩阵组1006可以表征为g(UΣV)，计算$g(U\Sigma V)=\sigma((U\Sigma V)^TX)$，从Σ中选择最大的$r_t$个参数$(top_{r_t}(\Sigma))$将其遮掩(mask)

为1,同时把其他元素遮掩设置为0(这样就可以使得U与V对应的行与列的计算结果全部为0)。

[0118] (3)对第t步骤迭代中的低秩分解矩阵组1006(秩为r的可压缩矩阵)$W_t^r = U|\sum|_0 V(\in R^{d1\times d2})$进行二阶泰勒展开,推导过程如下公式(1)和公式(2)所示:

[0119] $$g(U\sum V) = g(W_{t-1}^r) + \frac{1}{2}g'(W_{t-1}^r) * (W_t^r - W_{t-1}^r) + g''(W_{t-1}^r) * (W_t^r - W_{t-1}^r)^2 + o((W_t^r - W_{t-1}^r)^3),\ 公式(1);$$

[0120] $$g(U\sum V) \approx \sigma((W_{t-1}^r)^T X) + \sigma'((W_{t-1}^r)X)X + \sigma''\left((W_{t-1}^r)X\right)X^T X,\ 公式(2)。$$

[0121] 在上式中,对于一些线性激活函数,上述二阶导(g″)为0,只包含一阶导的信息(g′),对于其他非线下激活函数,则二阶导数提供额外的拟合信息。

[0122] 其中,U表征左奇异矩阵,V表征右奇异矩阵,Σ表征奇异值矩阵(只有对角线有数值λ_i,其他全为0),W_t^r表征第t步时的可压缩矩阵,W_0^r表征待压缩矩阵W 1004的原本权重,σ表征激活函数,σ′表征激活函数的一阶求导,σ″表征激活函数的二阶求导,X表征输入样本矩阵1002,X^T表征输入样本的转置矩阵,g表征表示参数W 1004的神经网络函数。

[0123] (4)基于这个矩阵分解前后得输出值,计算原模型和分解后模型的输出损失如下公式(3)所示:

[0124] $$\iota_r = (g(W_0^r) - f(W))^2,\ 公式(3)。$$

[0125] (5)引入L0优化算法,计算低秩分解矩阵组1006中对角矩阵Σ参数的L0范式的正则损失$\iota_{0\text{-norm}}$,此外,还需要引入下游任务的分类蒸馏损失ι_{pred},从而得到总体的计算损失ι的表达式如下公式(4)所示:

[0126] $\iota = \iota_r + \iota_{0\text{-norm}} + \iota_{pred}$,公式(4)。

[0127] (6)通过如上损失函数进行低秩矩阵组U、V、Σ参数的优化,直到模型收敛,就可以学习到最大限度保留原始模型知识的低秩矩阵参数。

[0128] 在按照上述步骤完成神经网络压缩后,基于压缩后的神经网络模型的推理计算过程如下:

[0129] 从Σ矩阵中选择最大的r_s个参数,然后将其他参数置0,同时从U和V中删除这些0元素对应的行和列的数据,将分解矩阵组压缩到$U \in R^{d1\times r_s}, V \in R^{r_s\times d2}$,且$\Sigma = \text{diagonal}(\{\lambda_1, \lambda_2, \ldots, \lambda_{r_s}\})$。此时神经网络的计算复杂度和存储体积降低到$\frac{r_s(d1+d2)}{d1*d2}$,由于$r_s << \min(d1,d2)$,所以这个比例非常小,模型获得大幅度的提速。

[0130] 基于此,本公开的实施例限定的神经网络压缩方案可以针对任意神经网络矩阵进行优化压缩,通用性非常高,对于工业化落地与部署等十分有用。

[0131] 对应于上述方法实施例,本公开还提供一种网络模型压缩装置,可以用于执行上述方法实施例。

[0132] 图11是本公开示例性实施例中一种网络模型压缩装置的方框图。

[0133] 参考图11,网络模型压缩装置1100可以包括:

[0134] 确定模块1102,设置为根据待压缩的原始网络模型的指定矩阵的奇异值确定对应

的低秩分解矩阵组。

[0135] 压缩模块1104，设置为基于所述原始网络模型对应的矩阵蒸馏损失函数对所述低秩分解矩阵组进行压缩。

[0136] 在本公开的一种示例性实施例中，确定模块1102还设置为：

[0137] 通过奇异值分解算法确定所述原始网络模型中的指定矩阵的奇异值；

[0138] 根据所述奇异值的个数和预设的超参数确定所述低秩分解矩阵组的第一维度参数；

[0139] 根据所述原始网络模型确定奇异值矩阵的维度；

[0140] 通过正态分布随机初始化第一低维矩阵和第二低维矩阵，并构建所述奇异值矩阵，所述低秩分解矩阵组包括所述第一低维矩阵、所述第二低维矩阵和所述奇异值矩阵。

[0141] 在本公开的一种示例性实施例中，确定模块1102还设置为：

[0142] 确定所述原始网络模型中的指定矩阵的第二维度参数和第三维度参数；

[0143] 根据所述第一维度参数和所述第二维度参数确定所述第一低维矩阵的维度；

[0144] 根据所述第二维度参数和所述第三维度参数确定所述第二低维矩阵的维度；

[0145] 根据所述第二维度参数确定奇异值矩阵的维度。

[0146] 在本公开的一种示例性实施例中，确定模块1102还设置为：

[0147] 根据所述奇异值的个数和所述超参数确定所述第一维度参数的初始值；

[0148] 根据所述原始网络模型中的指定矩阵的维度确定所述第一维度参数的目标值；

[0149] 针对所述第一维度参数的初始值创建迭代参数；

[0150] 根据所述迭代参数对所述第一维度参数的初始值进行迭代计算，至达到所述第一维度参数的目标值为止。

[0151] 在本公开的一种示例性实施例中，确定模块1102还设置为：

[0152] 确定所述原始网络模型中的指定矩阵的维度的最小值；

[0153] 根据所述最小值与预设比例值确定所述第一维度参数的目标值。

[0154] 在本公开的一种示例性实施例中，压缩模块1104还设置为：

[0155] 将样本输入至所述原始网络模型对应的矩阵，并确定所述原始网络模型的输出结果，记作第一输出结果；

[0156] 将所述样本输入至待压缩的低秩分解矩阵组，并确定所述低秩分解矩阵组的输出结果，记作第二输出结果；

[0157] 根据所述第一输出结果和所述第二输出结果确定输出损失；

[0158] 确定所述奇异值矩阵对应的正则损失；

[0159] 确定所述原始网络模型对所述样本的处理损失；

[0160] 根据所述输出损失、所述正则损失和所述处理损失对所述低秩分解矩阵组进行压缩。

[0161] 在本公开的一种示例性实施例中，压缩模块1104还设置为：

[0162] 基于L0优化算法计算所述奇异值矩阵对应的L0范式的正则损失。

[0163] 在本公开的一种示例性实施例中，确定所述原始网络模型对所述样本的处理损失包括：

[0164] 确定所述样本的处理任务为分类任务；

[0165] 确定所述原始网络模型对所述分类任务处理后的分类损失，所述处理损失包括所述分类损失。

[0166] 由于网络模型压缩装置1100的各功能已在其对应的方法实施例中予以详细说明，本公开于此不再赘述。

[0167] 应当注意，尽管在上文详细描述中提及了用于动作执行的设备的若干模块或者单元，但是划分并非强制性的。实际上，根据本公开的实施方式，上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之，上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

[0168] 在本公开的示例性实施例中，还提供了一种能够实现上述方法的电子设备。

[0169] 所属技术领域的技术人员能够理解，本发明的各个方面可以实现为系统、方法或程序产品。因此，本发明的各个方面可以具体实现为以下形式，即：完全的硬件实施方式、完全的软件实施方式(包括固件、微代码等)，或硬件和软件方面结合的实施方式，这里可以统称为“电路”、“模块”或“系统”。

[0170] 下面参照图12来描述根据本发明的实施方式的电子设备1200。图12显示的电子设备1200仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。

[0171] 如图12所示，电子设备1200以通用计算设备的形式表现。电子设备1200的组件可以包括但不限于：上述至少一个处理单元1210、上述至少一个存储单元1220、连接不同系统组件(包括存储单元1220和处理单元1210)的总线1230。

[0172] 其中，所述存储单元存储有程序代码，所述程序代码可以被所述处理单元1210执行，使得所述处理单元1210执行本说明书上述“示例性方法”部分中描述的根据本发明各种示例性实施方式的步骤。例如，所述处理单元1210可以执行如本公开实施例所示的方法。

[0173] 存储单元1220可以包括易失性存储单元形式的可读介质，例如随机存取存储单元(RAM)12201和/或高速缓存存储单元12202，还可以进一步包括只读存储单元(ROM)12203。

[0174] 存储单元1220还可以包括具有一组(至少一个)程序模块12205的程序/实用工具12204，这样的程序模块12205包括但不限于：操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。

[0175] 总线1230可以为表示几类总线结构中的一种或多种，包括存储单元总线或者存储单元控制器、外围总线、图形加速端口、处理单元或者使用多种总线结构中的任意总线结构的局域总线。

[0176] 电子设备1200也可以与一个或多个外部设备1240(例如键盘、指向设备、蓝牙设备等)通信，还可与一个或者多个使得用户能与该电子设备1200交互的设备通信，和/或与使得该电子设备1200能与一个或多个其它计算设备进行通信的任何设备(例如路由器、调制解调器等等)通信。通信可以通过输入/输出(I/O)接口1250进行。并且，电子设备1200还可以通过网络适配器1260与一个或者多个网络(例如局域网(LAN)，广域网(WAN)和/或公共网络，例如因特网)通信。如图所示，网络适配器1260通过总线1230与电子设备1200的其它模块通信。应当明白，尽管图中未示出，可以结合电子设备1200使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0177] 通过以上的实施方式的描述，本领域的技术人员易于理解，这里描述的示例实施

方式可以通过软件实现，也可以通过软件结合必要的硬件的方式来实现。因此，根据本公开实施方式的技术方案可以以软件产品的形式体现出来，该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM，U盘，移动硬盘等)中或网络上，包括若干指令以使得一台计算设备(可以是个人计算机、服务器、终端装置、或者网络设备等)执行根据本公开实施方式的方法。

[0178] 在本公开的示例性实施例中，还提供了一种计算机可读存储介质，其上存储有能够实现本说明书上述方法的程序产品。在一些可能的实施方式中，本发明的各个方面还可以实现为一种程序产品的形式，其包括程序代码，当所述程序产品在终端设备上运行时，所述程序代码用于使所述终端设备执行本说明书上述“示例性方法”部分中描述的根据本发明各种示例性实施方式的步骤。

[0179] 根据本发明的实施方式的用于实现上述方法的程序产品可以采用便携式紧凑盘只读存储器(CD-ROM)并包括程序代码，并可以在终端设备，例如个人电脑上运行。然而，本发明的程序产品不限于此，在本文件中，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。

[0180] 所述程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以为但不限于电、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。可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器(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。

[0181] 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，其中承载了可读程序代码。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。可读信号介质还可以是可读存储介质以外的任何可读介质，该可读介质可以发送、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。

[0182] 可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于无线、有线、光缆、RF等等，或者上述的任意合适的组合。

[0183] 可以以一种或多种程序设计语言的任意组合来编写用于执行本发明操作的程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、C++等，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。在涉及远程计算设备的情形中，远程计算设备可以通过任意种类的网络，包括局域网(LAN)或广域网(WAN)，连接到用户计算设备，或者，可以连接到外部计算设备(例如利用因特网服务提供商来通过因特网连接)。

[0184] 此外，上述附图仅是根据本发明示例性实施例的方法所包括的处理的示意性说明，而不是限制目的。易于理解，上述附图所示的处理并不表明或限制这些处理的时间顺序。另外，也易于理解，这些处理可以是例如在多个模块中同步或异步执行的。

[0185] 本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本公开的其

它实施方案。本申请旨在涵盖本公开的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本公开的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本公开的真正范围和构思由权利要求指出。

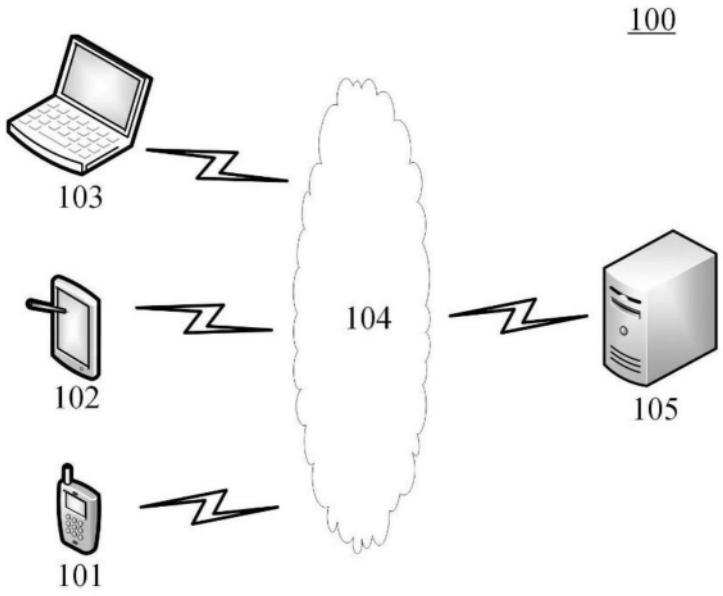

100
103
102
101
104
105

图1

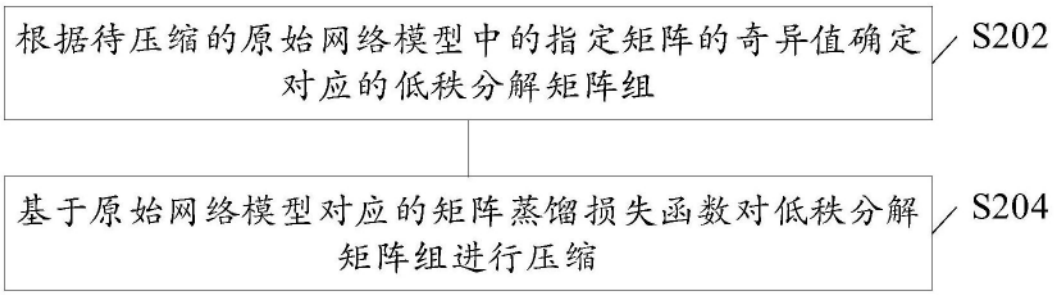

根据待压缩的原始网络模型中的指定矩阵的奇异值确定对应的低秩分解矩阵组
S202
基于原始网络模型对应的矩阵蒸馏损失函数对低秩分解矩阵组进行压缩
S204

图2

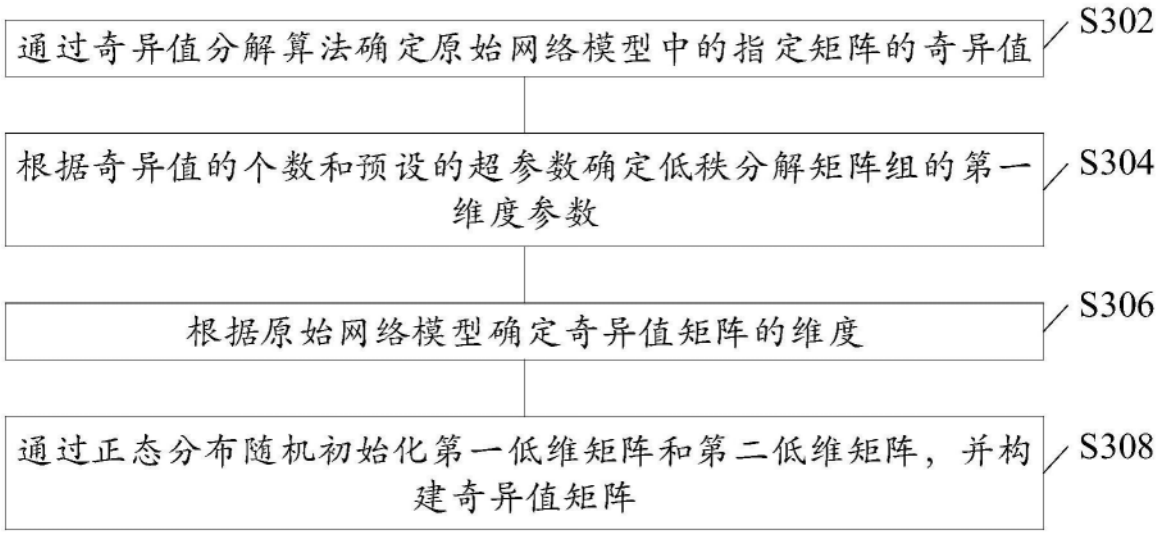

通过奇异值分解算法确定原始网络模型中的指定矩阵的奇异值
S302
根据奇异值的个数和预设的超参数确定低秩分解矩阵组的第一维度参数
S304
根据原始网络模型确定奇异值矩阵的维度
S306
通过正态分布随机初始化第一低维矩阵和第二低维矩阵，并构建奇异值矩阵
S308

图3

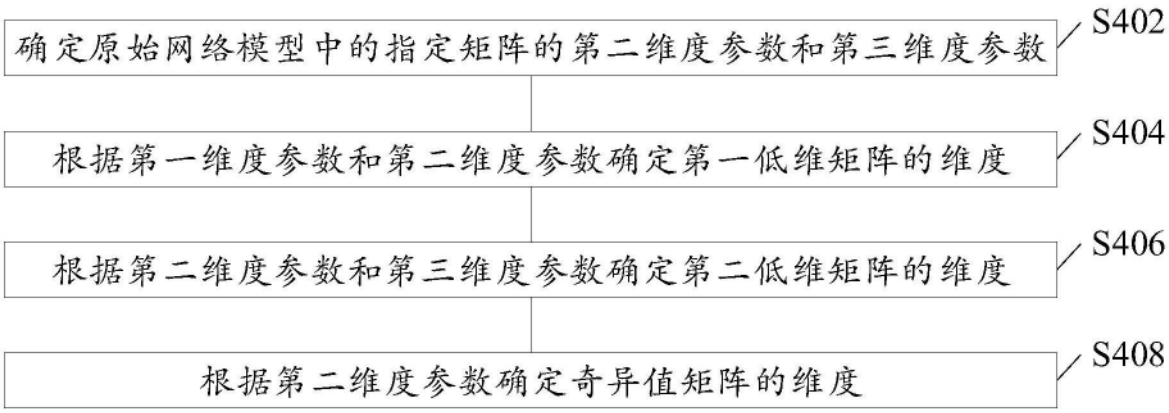

确定原始网络模型中的指定矩阵的第二维度参数和第三维度参数
S402
根据第一维度参数和第二维度参数确定第一低维矩阵的维度
S404
根据第二维度参数和第三维度参数确定第二低维矩阵的维度
S406
根据第二维度参数确定奇异值矩阵的维度
S408

图4

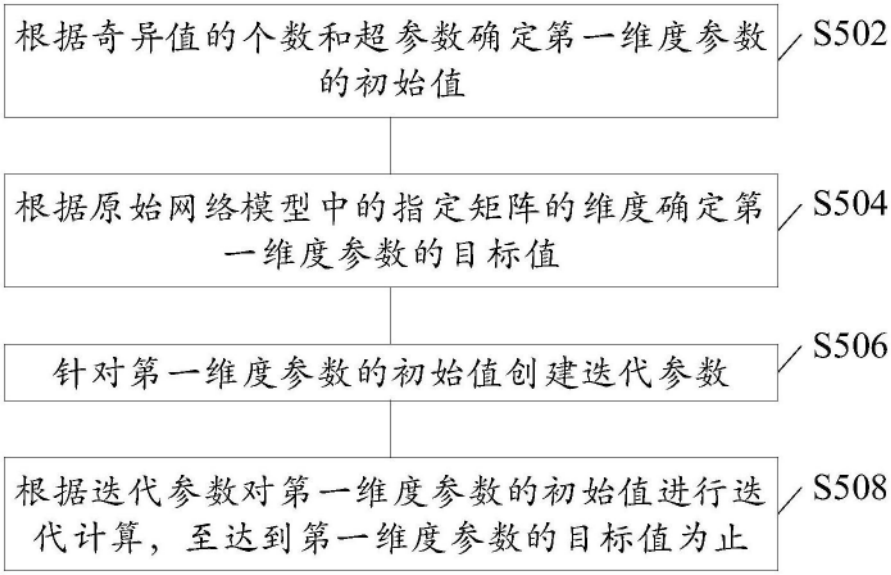

根据奇异值的个数和超参数确定第一维度参数的初始值
S502
根据原始网络模型中的指定矩阵的维度确定第一维度参数的目标值
S504
针对第一维度参数的初始值创建迭代参数
S506
根据迭代参数对第一维度参数的初始值进行迭代计算，至达到第一维度参数的目标值为止
S508

图5

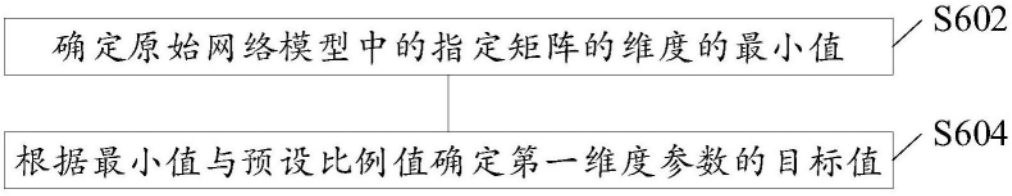

确定原始网络模型中的指定矩阵的维度的最小值
S602
根据最小值与预设比例值确定第一维度参数的目标值
S604

图6

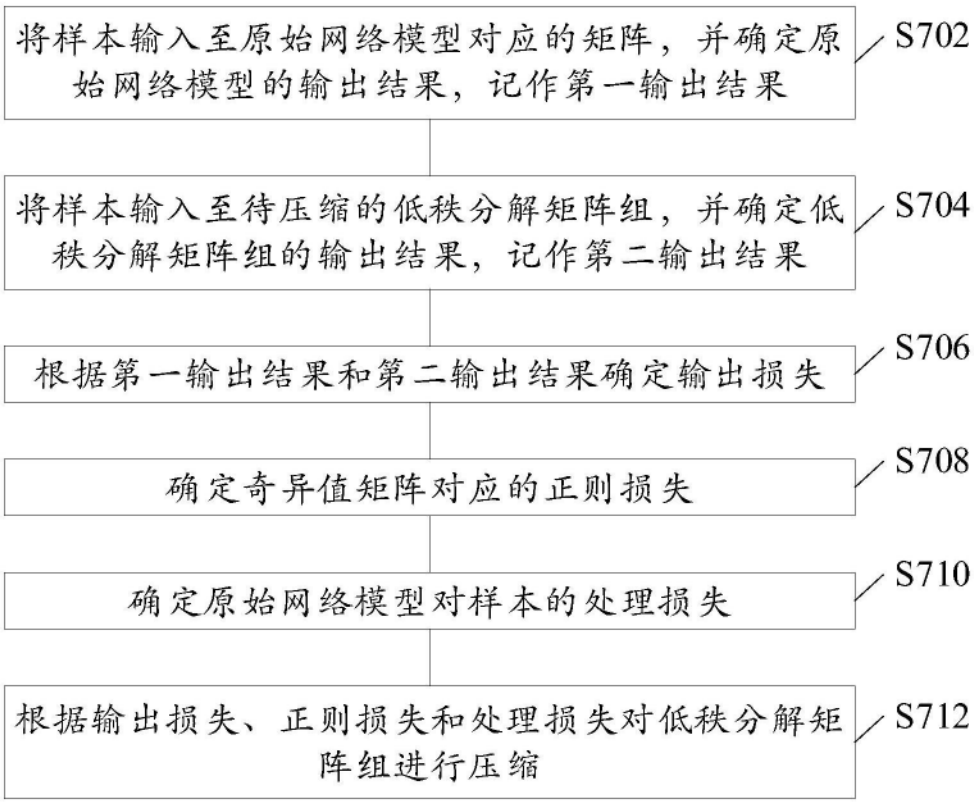

将样本输入至原始网络模型对应的矩阵，并确定原始网络模型的输出结果，记作第一输出结果
S702
将样本输入至待压缩的低秩分解矩阵组，并确定低秩分解矩阵组的输出结果，记作第二输出结果
S704
根据第一输出结果和第二输出结果确定输出损失
S706
确定奇异值矩阵对应的正则损失
S708
确定原始网络模型对样本的处理损失
S710
根据输出损失、正则损失和处理损失对低秩分解矩阵组进行压缩
S712

图7

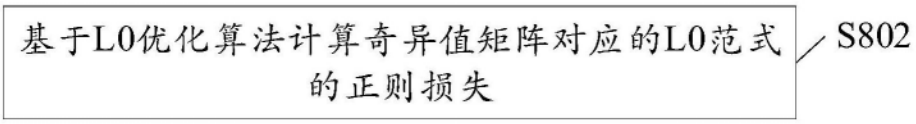

基于L0优化算法计算奇异值矩阵对应的L0范式的正则损失
S802

图8

确定样本的处理任务为分类任务
S902
确定原始网络模型对分类任务处理后的分类损失
S904

图9

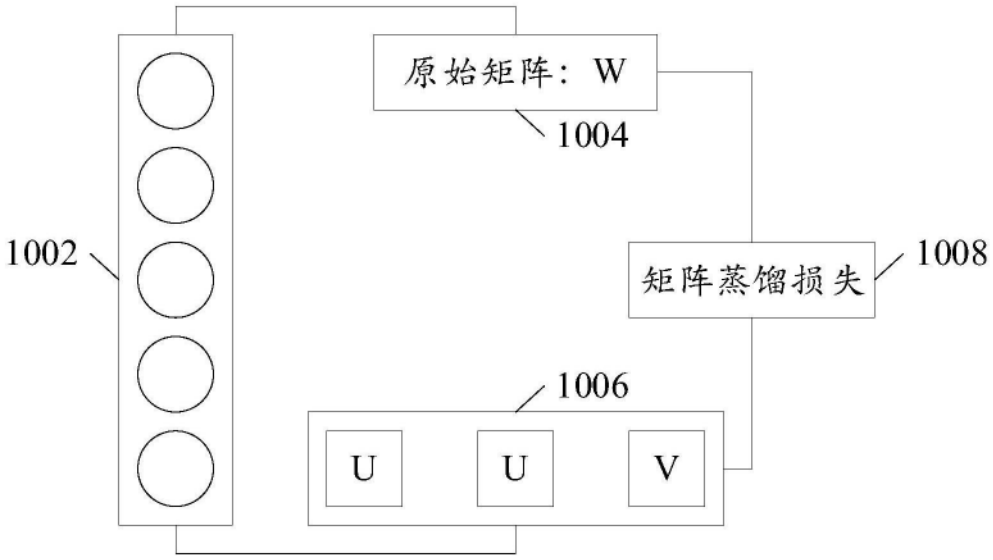

原始矩阵：W
1004
1002
矩阵蒸馏损失
1008
1006
U
U
V

图10

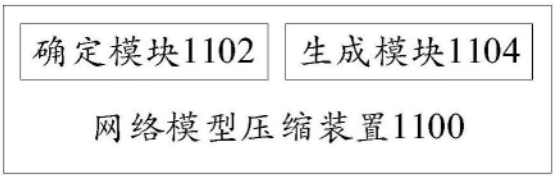

确定模块1102
生成模块1104
网络模型压缩装置1100

图11

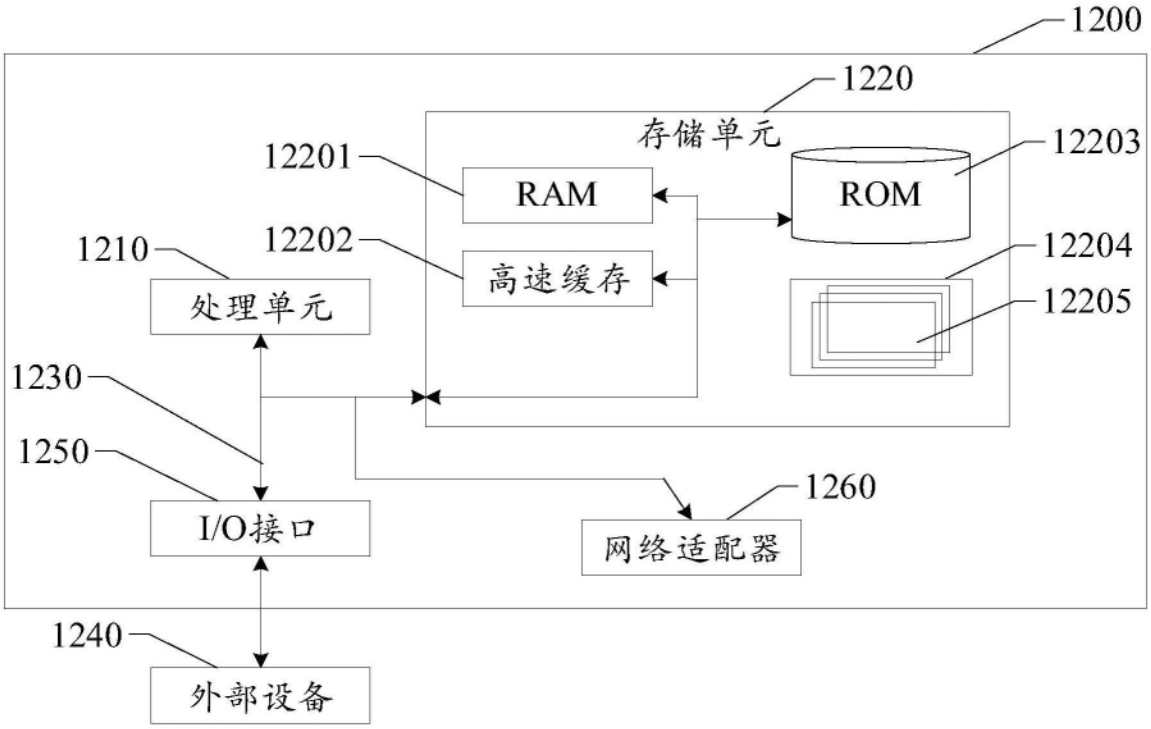

1200
1220
存储单元
12201
RAM
12203
ROM
1210
12202
高速缓存
12204
处理单元
12205
1230
1250
1260
I/O接口
网络适配器
1240
外部设备

图12