(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 110046116 B
(45)授权公告日 2020.08.21

(21)申请号 201910327608.8

(22)申请日 2019.04.23

(65)同一申请的已公布的文献号
申请公布号 CN 110046116 A

(43)申请公布日 2019.07.23

(73)专利权人 上海燧原智能科技有限公司
地址 201306 上海市浦东新区南汇新城镇环湖西二路888号C楼
专利权人 上海燧原科技有限公司

(72)发明人 车驰

(74)专利代理机构 北京品源专利代理有限公司 11332
代理人 孟金喆

(51)Int.Cl.
G06F 13/28(2006.01)
G06F 16/28(2019.01)
G06N 3/04(2006.01)

(56)对比文件
CN 109190758 A,2019.01.11
CN 109255438 A,2019.01.22
CN 208766643 U,2019.04.19
CN 109324827 A,2019.02.12
CN 108875957 A,2018.11.23
US 2017213145 A1,2017.07.27

审查员 贾东曜

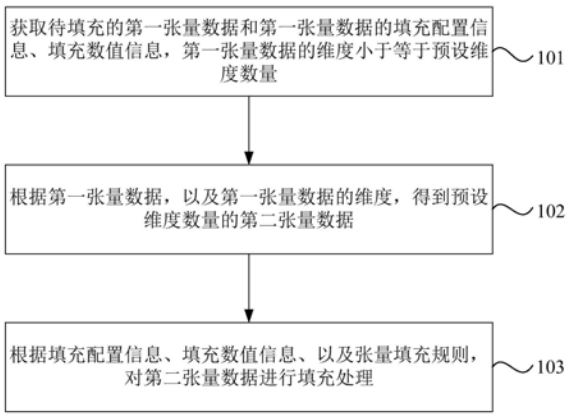
权利要求书2页 说明书12页 附图3页

(54)发明名称

一种张量填充方法、装置、设备及存储介质

(57)摘要

本发明实施例公开了一种张量填充方法、装置、设备及存储介质。其中，该方法包括：获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量；根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据；根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理。本发明实施例解决了现有技术对于维度大于2的原始张量数据，无法利用DMA操作直接完成所有维度的张量填充的问题，可以直接利用DMA操作完成预设维度数量以内任意维度的张量填充操作，提升张量填充效率，极大缩短张量填充所需的时间。



CN 110046116 B

1.一种张量填充方法,其特征在于,包括:

获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息,所述第一张量数据为输入图像数据,所述第一张量数据的维度小于等于四维;

根据所述第一张量数据,以及所述第一张量数据的维度,得到四维的第二张量数据,第二张量数据包括:第一维度的数据、第二维度的数据、第三维度的数据、以及第四维度的数据;其中,第一维度为最高维,第四维度为最低维;

根据所述第一维度的数据和所述第二维度的数据的填充配置信息,判断所述第一维度的数据和所述第二维度的数据是否需要进行填充处理;如果所述第一维度的数据和所述第二维度的数据需要进行填充处理,则通过DMA张量填充操作,根据所述填充配置信息、所述填充数值信息、以及张量填充规则,对所述第二张量数据的第三维度的数据和第四维度的数据进行填充处理;通过DMA张量转置操作,根据第一期望维度序列对填充处理后的第二张量数据进行转置处理,以使所述第二张量数据的第一维度的数据、第二维度的数据与第三维度的数据、第四维度的数据顺序互换;通过DMA张量填充操作,根据所述填充配置信息、所述填充数值信息、以及张量填充规则,对所述顺序互换后的第二张量数据的第三维度的数据和第四维度的数据进行填充处理;通过DMA张量转置操作,根据所述第一期望维度序列对填充处理后的第二张量数据进行转置处理,完成四维以内的张量填充,以使所述第二张量数据的大小符合窗口移动的要求。

2.根据权利要求1所述的方法,其特征在于,根据第一张量数据,以及所述第一张量数据的维度,得到四维的第二张量数据,包括:

当所述第一张量数据的维度小于四维时,根据所述维度和预设维度补充规则,将所述第一张量数据转换为四维的第二张量数据。

3.根据权利要求1所述的方法,其特征在于,通过DMA张量填充操作,根据所述填充配置信息、所述填充数值信息、以及张量填充规则,对所述第二张量数据的第三维度的数据和第四维度的数据进行填充处理,包括:

根据所述第三维度的数据和所述第四维度的数据的填充配置信息,判断所述第三维度的数据和所述第四维度的数据是否需要进行中部数据填充处理;

如果所述第四维度的数据不需要进行中部数据填充处理,则通过DMA张量填充操作,根据所述填充配置信息和所述填充数值信息,对所述第二张量数据进行填充处理。

4.根据权利要求3所述的方法,其特征在于,在判断所述第三维度的数据和所述第四维度的数据是否需要进行中部数据填充处理之后,还包括:

如果所述第四维度的数据需要进行中部数据填充处理,且所述第三维度的数据不需要进行中部数据填充处理,则通过DMA张量转置操作,根据第二期望维度序列对所述第二张量数据进行转置处理,以使所述第二张量数据的第三维度的数据和第四维度的数据互换顺序;

通过DMA张量填充操作,根据所述填充配置信息和所述填充数值信息,对转置处理后的第二张量数据进行填充处理;

通过DMA张量转置操作,根据所述第二期望维度序列对填充处理后的第二张量数据进行转置处理。

5.根据权利要求3所述的方法,其特征在于,在判断所述第三维度的数据和所述第四维

度的数据是否需要进行中部数据填充处理之后，还包括：

如果所述第四维度的数据需要进行中部数据填充处理，且所述第三维度的数据需要进行中部数据填充处理，则通过DMA张量填充操作，根据所述填充配置信息和所述填充数值信息，对所述第二张量数据进行第一次填充处理；

通过DMA张量转置操作，根据第二期望维度序列对第一次填充处理后的第二张量数据进行转置处理，以使所述第二张量数据的第三维度的数据和第四维度的数据互换顺序；

通过DMA张量填充操作，根据所述填充配置信息和所述填充数值信息，对转置处理后的第二张量数据进行第二次填充处理；

通过DMA张量转置操作，根据所述第二期望维度序列对第二次填充处理后的第二张量数据进行转置处理。

6.一种张量填充装置，其特征在于，包括：

数据获取模块，用于获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，所述第一张量数据为输入图像数据，所述第一张量数据的维度小于等于四维；

数据确定模块，用于根据第一张量数据，以及所述第一张量数据的维度，得到四维的第二张量数据，第二张量数据包括：第一维度的数据、第二维度的数据、第三维度的数据、以及第四维度的数据；其中，第一维度为最高维，第四维度为最低维；

数据填充模块，用于根据所述第一维度的数据和所述第二维度的数据的填充配置信息，判断所述第一维度的数据和所述第二维度的数据是否需要进行填充处理；如果所述第一维度的数据和所述第二维度的数据需要进行填充处理，则通过DMA张量填充操作，根据所述填充配置信息、所述填充数值信息、以及张量填充规则，对所述第二张量数据的第三维度的数据和第四维度的数据进行填充处理；通过DMA张量转置操作，根据第一期望维度序列对填充处理后的第二张量数据进行转置处理，以使所述第二张量数据的第一维度的数据、第二维度的数据与第三维度的数据、第四维度的数据顺序互换；通过DMA张量填充操作，根据所述填充配置信息、所述填充数值信息、以及张量填充规则，对所述顺序互换后的第二张量数据的第三维度的数据和第四维度的数据进行填充处理；通过DMA张量转置操作，根据所述第一期望维度序列对填充处理后的第二张量数据进行转置处理，完成四维以内的张量填充，以使所述第二张量数据的大小符合窗口移动的要求。

7.一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求1-5中任一所述的张量填充方法。

8.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该计算机程序被处理器执行时实现如权利要求1-5中任一所述的张量填充方法。

一种张量填充方法、装置、设备及存储介质

技术领域

[0001] 本发明实施例涉及数据处理技术，尤其涉及一种张量填充方法、装置、设备及存储介质。

背景技术

[0002] 张量填充是神经网络中，一种常用的张量操作方法。张量填充具体是指，对于张量数据，可以在指定的维度上，对该维度的数据进行前部、中间、后部的数据填充。因为给定的原始张量数据往往在大小尺寸上，不能满足神经网络中的实际需要，所以需要张量填充操作来完成对原始张量数据的拓展，获得满足神经网络的实际需求的张量数据。

[0003] 对张量的填充，会在各个维度上增加大量数据，消耗大量操作时间，因此，提升张量填充效率，对于整个神经网络的运行速度提高至关重要。目前，进行张量填充操作一般是采用直接内存访问设备(Direct MemoryAccess，DMA)实现。

[0004] 发明人在实现本发明的过程中发现，现有技术的缺陷在于，现有的DMA张量填充操作每次只能对原始张量数据的最低的两个维度的数据进行填充。对于维度大于2的原始张量数据，无法利用DMA操作直接完成所有维度的张量填充。

发明内容

[0005] 本发明实施例提供一种张量填充方法、装置、设备及存储介质，以优化现有的张量填充方法，提升张量填充效率。

[0006] 第一方面，本发明实施例提供了一种张量填充方法，包括：

[0007] 获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量；

[0008] 根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据；

[0009] 根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理。

[0010] 第二方面，本发明实施例还提供了一种张量填充装置，包括：

[0011] 数据获取模块，用于获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量；

[0012] 数据确定模块，用于根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据；

[0013] 数据填充模块，用于根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理。

[0014] 第三方面，本发明实施例还提供了一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，处理器执行计算机程序时实现如本发明实施例所述的张量填充方法。

[0015] 第四方面，本发明实施例还提供了一种计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现如本发明实施例所述的张量填充方法。

[0016] 本发明实施例的技术方案，通过获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量，并根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据，然后根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理，解决了现有技术对于维度大于2的原始张量数据，无法利用DMA操作直接完成所有维度的张量填充的问题，可以直接利用DMA操作完成预设维度数量以内任意维度的张量填充操作，提升张量填充效率，极大缩短张量填充所需的时间。

附图说明

[0017] 图1为本发明实施例一提供的一种张量填充方法的流程图；

[0018] 图2为本发明实施例二提供的一种张量填充方法的流程图；

[0019] 图3为本发明实施例三提供的一种张量填充装置的结构示意图；

[0020] 图4为本发明实施例四提供的一种计算机设备的结构示意图。

具体实施方式

[0021] 下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释本发明，而非对本发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与本发明相关的部分而非全部结构。

[0022] 另外还需要说明的是，为了便于描述，附图中仅示出了与本发明相关的部分而非全部内容。在更加详细地讨论示例性实施例之前应当提到的是，一些示例性实施例被描述成作为流程图描绘的处理或方法。虽然流程图将各项操作(或步骤)描述成顺序的处理，但是其中的许多操作可以被并行地、并发地或者同时实施。此外，各项操作的顺序可以被重新安排。当其操作完成时所述处理可以被终止，但是还可以具有未包括在附图中的附加步骤。所述处理可以对应于方法、函数、规程、子例程、子程序等等。

[0023] 为了便于理解，将本发明实施例的主要发明构思进行简述。首先，发明人针对现有技术中的主要问题：现有的DMA张量填充操作每次只能对原始张量数据的最低的两个维度的数据进行填充，对于维度大于2的原始张量数据，无法利用DMA操作直接完成所有维度的张量填充，考虑是否可以利用现有DMA张量转置操作和DMA张量填充操作，对原始张量数据进行维度顺序的调换，并对指定维度进行填充，完成所有维度的张量填充，实现直接利用DMA操作完成维度大于2的原始张量数据的所有维度的张量填充。因为是直接利用DMA操作完成填充，运行速度快，效率高，可以极大缩短张量填充所需的时间。

[0024] 基于上述思考，发明人创造性的提出，获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量，并根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据，然后根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理。由此，直接利用DMA操作完成预设维度数量以内任意维度的张量填充操作，提升张量填充效率，极大缩短张量填充所需的时间。

[0025] 实施例一

[0026] 图1为本发明实施例一提供的一种张量填充方法的流程图。本实施例可适用于对预设维度数量以内的张量数据进行填充的情况，该方法可以由本发明实施例提供的张量填充装置来执行，该装置可采用软件和/或硬件的方式实现，并一般可集成在计算机设备中。如图1所示，本实施例的方法具体包括：

[0027] 步骤101、获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量。

[0028] 其中，待填充的第一张量数据是需要进行张量填充的原始张量数据。当原始张量数据在大小尺寸上，不能满足神经网络中的实际需要时，可通过张量填充操作来完成对原始张量数据的拓展，获得满足神经网络的实际需求的张量数据。

[0029] 在一个具体实例中，计算机视觉领域中的物体识别，图像分割，人脸识别等技术可以被应用在安防，交通，人工智能等各个领域。目前主流的算法是利用深度学习的方式，训练模型来高效的实现物体识别，图像分割，人脸识别等功能。具体的，在深度学习的训练中，会利用卷积(Convolution)的方式，通过卷积核(Kernel)窗口，在输入图像中横向和纵向的移动，来提取输入图像中各个区域的信息。在窗口移动时，窗口的大小和每次窗口移动的步长，可能会和输入的图像大小不符，这个时候就需要利用张量填充，使得输入图像的大小变成符合窗口移动的要求，这样深度学习的训练才能够进行。

[0030] 输入图像数据即为一种需要进行张量填充的原始张量数据。通过张量填充操作来对输入图像数据的大小进行拓展，获得大小符合窗口移动的要求的输入图像数据，满足深度学习的训练的实际需求。

[0031] 第一张量数据的维度小于等于预设维度数量。预设维度数量可以根据需求设置。可选的，预设维度数量大于二维。例如，预设维度数量为四维。第一张量数据的维度可以为一维、二维、三维或者四维。

[0032] 预先根据业务需求，设置第一张量数据的填充配置信息、填充数值信息。填充数值信息是将要填入的数值。可以将填入的数值设置为任意数值。例如，将填入的数值设置为0。填充配置信息是第一张量数据的各个维度进行数据填充的配置信息。每一个维度有对应的配置信息。每一个配置信息都可以同时包括对应维度的前部填充信息，中部填充信息和后部填充信息。前部填充信息表示对前部填充的数值个数。中部填充信息表示对当前处理维度中每两个数字中间填充的数值个数。后部填充信息表示对后部填充的数值个数。例如，填充数值信息为0。第一张量数据的第三维度的张量数据为[5,6,7]，前部填充信息为2，中部填充信息为1，后部填充信息为3。填充后，第一张量数据的第三维度的张量数据为[0,0,5,0,6,0,7,0,0,0]。第一张量数据的第四维度的张量数据为[5,6,7]，前部填充信息为2，中部填充信息为空，后部填充信息为3。填充后，第一张量数据的第三维度的张量数据为[0,0,5,6,7,0,0,0]。

[0033] 步骤102、根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据。

[0034] 其中，在获取第一张量数据之后，确定第一张量数据的维度大小，各维度的维度大小，以及根据第一张量数据中各维度的顺序对各个维度设置标识，各维度的标识可以是序号、或者其他唯一表示一个维度数据的符号。

[0035] 示例性的，当第一张量数据为四维数据时，可以是依序设置各维度的标识(例如为序号)为0，1，2，3。即第一张量数据可表示为index(T)＝[0，1，2，3]。index(T)用于通过维度序号表示第一张量数据T。第一张量数据T的维度数量为4，分别包括标识为0，1，2，3的四个维度的数据。从左至右，0表示最高维，3表示最低维。dims(T)用于提取第一张量数据T中各维度的维度大小。例如，维度标识分别为0，1，2，3的维度大小可以是3，4，5，2，即dims(T)＝[3，4，5，2]。

[0036] 可选的，根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据，可以包括：当第一张量数据的维度等于预设维度数量时，直接将第一张量数据作为预设维度数量的第二张量数据。

[0037] 可选的，根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据，可以包括：当第一张量数据的维度小于预设维度数量时，根据维度和预设维度补充规则，将第一张量数据转换为预设维度数量的第二张量数据。

[0038] 具体的，因为是预设维度数量以内的张量填充，如果第一张量数据的维度小于预设维度数量，会将第一张量数据的高维当作1来补充，将第一张量数据转换为预设维度数量的第二张量数据。例如，预设维度数量为四维。dims(T)＝[7，8]，表示第一张量数据T只有两个维度并且大小分别为7和8。此时会将此2维当作最低维，并把剩余最高两维的大小当作1来处理，所以最终的实际结果是dims(T)＝[1，1，7，8]。

[0039] 步骤103、根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理。

[0040] 其中，现有的DMA张量填充操作每次能够同时对张量数据的最低两个维度进行填充，并且对倒数第一维只能实现前部和后部的数据填充，而对倒数第二维，可以实现前部、中部和后部的同时数据填充操作。现有的DMA张量转置操作可以实现张量数据的任意维度顺序调换。利用现有DMA张量转置操作和DMA张量填充操作，对第二张量数据进行维度顺序的调换，并根据填充配置信息、填充数值信息对第二张量数据的指定维度进行填充，完成第二张量数据的所有维度的张量填充，实现直接利用DMA操作完成预设维度数量以内的张量填充。

[0041] 在一个具体实例中，预设维度数量为四维。第二张量数据包括：第一维度的数据、第二维度的数据、第三维度的数据、以及第四维度的数据。其中，第一维度为最高维，第四维度为最低维。

[0042] 其中，获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量。根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据。根据第一维度的数据和第二维度的数据的填充配置信息，判断第一维度的数据和第二维度的数据是否需要进行填充处理。

[0043] 如果第一维度的数据和第二维度的数据需要进行填充处理，则根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据的第三维度的数据和第四维度的数据进行填充处理。根据第一期望维度序列对填充处理后的第二张量数据进行转置处理，以使第二张量数据的第一维度的数据、第二维度的数据与第三维度的数据、第四维度的数据顺序互换。根据填充配置信息、填充数值信息、以及张量填充规则，对顺序互换后的第二张量数据的第三维度的数据和第四维度的数据进行填充处理。根据第一期望维度序列对填充

处理后的第二张量数据进行转置处理。

[0044] 如果第一维度的数据和第二维度的数据不需要进行填充处理，则直接根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据的第三维度的数据和第四维度的数据进行填充处理。

[0045] 具体的，根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据的第三维度的数据和第四维度的数据进行填充处理，可以包括：根据第三维度的数据和第四维度的数据的填充配置信息，判断第三维度的数据和第四维度的数据是否需要进行中部数据填充处理。

[0046] 如果第四维度的数据不需要进行中部数据填充处理，则根据填充配置信息和填充数值信息，对第二张量数据进行填充处理。

[0047] 如果第四维度的数据需要进行中部数据填充处理，且第三维度的数据不需要进行中部数据填充处理，则根据第二期望维度序列对第二张量数据进行转置处理，以使第二张量数据的第三维度的数据和第四维度的数据互换顺序；根据填充配置信息和填充数值信息，对转置处理后的第二张量数据进行填充处理；根据第二期望维度序列对填充处理后的第二张量数据进行转置处理。

[0048] 如果第四维度的数据需要进行中部数据填充处理，且第三维度的数据需要进行中部数据填充处理，则根据填充配置信息和所述填充数值信息，对第二张量数据进行第一次填充处理；根据第二期望维度序列对第一次填充处理后的第二张量数据进行转置处理，以使第二张量数据的第三维度的数据和第四维度的数据互换顺序；根据填充配置信息和填充数值信息，对转置处理后的第二张量数据进行第二次填充处理；根据第二期望维度序列对第二次填充处理后的第二张量数据进行转置处理。

[0049] 本发明实施例提供了一种张量填充方法，通过获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量，并根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据，然后根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理，解决了现有技术对于维度大于2的原始张量数据，无法利用DMA操作直接完成所有维度的张量填充的问题，可以直接利用DMA操作完成预设维度数量以内任意维度的张量填充操作，提升张量填充效率，极大缩短张量填充所需的时间。

[0050] 实施例二

[0051] 图2为本发明实施例二提供的一种张量填充方法的流程图。本实施例可以与上述一个或者多个实施例中各个可选方案结合，在本实施例中，预设维度数量为四维；第二张量数据包括：第一维度的数据、第二维度的数据、第三维度的数据、以及第四维度的数据；其中，第一维度为最高维，第四维度为最低维。

[0052] 以及，根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理，可以包括：根据第一维度的数据和第二维度的数据的填充配置信息，判断第一维度的数据和第二维度的数据是否需要进行填充处理；如果第一维度的数据和第二维度的数据需要进行填充处理，则根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据的第三维度的数据和第四维度的数据进行填充处理；根据第一期望维度序列对填充处理后的第二张量数据进行转置处理，以使第二张量数据的第一维度的数据、第二维度

的数据与第三维度的数据、第四维度的数据顺序互换;根据填充配置信息、填充数值信息、以及张量填充规则,对顺序互换后的第二张量数据的第三维度的数据和第四维度的数据进行填充处理;根据第一期望维度序列对填充处理后的第二张量数据进行转置处理。

[0053] 如图2所示,本实施例的方法具体包括:

[0054] 步骤201、获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息,第一张量数据的维度小于等于预设维度数量。

[0055] 步骤202、根据第一张量数据,以及第一张量数据的维度,得到预设维度数量的第二张量数据。

[0056] 步骤203、根据第一维度的数据和第二维度的数据的填充配置信息,判断第一维度的数据和第二维度的数据是否需要进行填充处理:若是,则执行步骤204;若否,则执行步骤208。

[0057] 其中,如果确定第一维度的数据填充配置信息非空,或者第二维度的数据的填充配置信息非空,则判定第一维度的数据和第二维度的数据需要进行填充处理。如果确定第一维度的数据填充配置信息为空,且第二维度的数据的填充配置信息为空,则判定第一维度的数据和第二维度的数据不需要进行填充处理。

[0058] 步骤204、根据填充配置信息、填充数值信息、以及张量填充规则,对第二张量数据的第三维度的数据和第四维度的数据进行填充处理。

[0059] 其中,根据第三维度的数据和第四维度的数据的填充配置信息,判断第三维度的数据和第四维度的数据是否需要进行中部数据填充处理。具体的,根据对应的中部填充信息,判断数据是否需要进行中部数据填充处理。如果中部填充信息为0,则判定数据不需要进行中部数据填充处理;如果中部填充信息非0,则判定数据需要进行中部数据填充处理。

[0060] DMA张量填充操作每次能够同时对第二张量数据的第三维度的数据和第四维度的数据进行填充处理,并且对第四维度的数据只能实现前部和后部的数据填充,而对第三维度的数据,可以实现前部、中部和后部的同时数据填充操作。

[0061] 可选的,如果第四维度的数据不需要进行中部数据填充处理,则根据填充配置信息和填充数值信息,对第二张量数据进行填充处理。

[0062] 其中,第四维度的数据不需要进行中部数据填充处理,则直接通过DMA张量填充操作,根据填充配置信息和填充数值信息,对第二张量数据的第三维度的数据和第四维度的数据进行填充处理。

[0063] 可选的,如果第四维度的数据需要进行中部数据填充处理,且第三维度的数据不需要进行中部数据填充处理,则根据第二期望维度序列对第二张量数据进行转置处理,以使第二张量数据的第三维度的数据和第四维度的数据互换顺序;根据填充配置信息和填充数值信息,对转置处理后的第二张量数据进行填充处理;根据第二期望维度序列对填充处理后的第二张量数据进行转置处理。

[0064] 其中,通过DMA张量转置操作将第二张量数据的第三维度的数据和第四维度的数据互换顺序,将第四维度的数据调换至第三维度,第三维度的数据调换至第四维度。具体的,第二期望维度序列为[0,1,3,2]。按照第二期望维度序列表示的顺序对第二张量数据进行转置,使第二张量数据的第三维度的数据和第四维度的数据互换顺序。例如,dims(T)=[5,6,7,8]。按照第二期望维度序列[0,1,3,2]表示的顺序对第二张量数据T进行转置,得到

转置处理后的第二张量数据T’。dims(T’)＝[5,6,8,7]。

[0065] 然后可以通过DMA张量填充操作对第三维度的数据进行前部和后部的数据填充，对第四维度的数据进行前部、中部和后部的数据填充。填充完成后，根据第二期望维度序列对填充处理后的第二张量数据进行转置处理，将第二张量数据的第三维度的数据和第四维度的数据再调换回原来顺序。由此，完成对第三维度的数据和第四维度的数据的前部、中部和后部的数据填充。

[0066] 可选的，如果第四维度的数据需要进行中部数据填充处理，且第三维度的数据需要进行中部数据填充处理，则根据填充配置信息和所述填充数值信息，对第二张量数据进行第一次填充处理；根据第二期望维度序列对第一次填充处理后的第二张量数据进行转置处理，以使第二张量数据的第三维度的数据和第四维度的数据互换顺序；根据填充配置信息和填充数值信息，对转置处理后的第二张量数据进行第二次填充处理；根据第二期望维度序列对第二次填充处理后的第二张量数据进行转置处理。

[0067] 其中，在第一次填充处理中，通过DMA张量填充操作对第四维度的数据进行前部和后部的数据填充，对第三维度的数据进行前部、中部和后部的数据填充。然后将第四维度的数据调换至第三维度，第三维度的数据调换至第四维度，进行第二次填充处理，通过DMA张量填充操作对第四维度的数据进行中部的数据填充。第二次填充处理完成后，根据第二期望维度序列对填充处理后的第二张量数据进行转置处理，将第二张量数据的第三维度的数据和第四维度的数据再调换回原来顺序。由此，完成对第四维度的数据的前部、中部和后部的数据填充，以及对第三维度的数据的前部、中部和后部的数据填充。

[0068] 步骤205、根据第一期望维度序列对填充处理后的第二张量数据进行转置处理，以使第二张量数据的第一维度的数据、第二维度的数据与第三维度的数据、第四维度的数据顺序互换。

[0069] 其中，在完成对第二张量数据的第三维度的数据、第四维度的数据的填充处理后，将第二张量数据的第一维度的数据、第二维度的数据与第三维度的数据、第四维度的数据顺序互换，将第三维度的数据调换至第一维度，第四维度的数据调换至第二维度，第一维度的数据调换至第三维度，第二维度的数据调换至第四维度。第一期望维度序列为[2,3,0,1]。例如，dims(T)＝[5,6,7,8]。按照第一期望维度序列[2,3,0,1]表示的顺序对第二张量数据T进行转置，将第二张量数据T的前两维与后两维顺序调换，得到转置处理后的第二张量数据T’。dims(T’)＝[7,8,5,6]。

[0070] 步骤206、根据填充配置信息、填充数值信息、以及张量填充规则，对顺序互换后的第二张量数据的第三维度的数据和第四维度的数据进行填充处理。

[0071] 其中，将第一维度的数据调换至第三维度，第二维度的数据调换至第四维度之后，根据填充配置信息、填充数值信息、以及张量填充规则，对顺序互换后的第二张量数据的第三维度的数据和第四维度的数据进行填充处理。

[0072] 根据顺序互换后的第二张量数据的第三维度的数据和第四维度的数据的填充配置信息，判断第三维度的数据和第四维度的数据是否需要进行中部数据填充处理。

[0073] 可选的，如果第四维度的数据不需要进行中部数据填充处理，则根据填充配置信息和填充数值信息，对第二张量数据进行填充处理。

[0074] 可选的，如果第四维度的数据需要进行中部数据填充处理，且第三维度的数据不

需要进行中部数据填充处理，则根据第二期望维度序列对第二张量数据进行转置处理，以使第二张量数据的第三维度的数据和第四维度的数据互换顺序；根据填充配置信息和填充数值信息，对转置处理后的第二张量数据进行填充处理；根据第二期望维度序列对填充处理后的第二张量数据进行转置处理。

[0075] 可选的，如果第四维度的数据需要进行中部数据填充处理，且第三维度的数据需要进行中部数据填充处理，则根据填充配置信息和所述填充数值信息，对第二张量数据进行第一次填充处理；根据第二期望维度序列对第一次填充处理后的第二张量数据进行转置处理，以使第二张量数据的第三维度的数据和第四维度的数据互换顺序；根据填充配置信息和填充数值信息，对转置处理后的第二张量数据进行第二次填充处理；根据第二期望维度序列对第二次填充处理后的第二张量数据进行转置处理。

[0076] 步骤207、根据第一期望维度序列对填充处理后的第二张量数据进行转置处理。

[0077] 其中，完成对顺序互换后的第二张量数据的第三维度的数据和第四维度的数据的填充处理后，根据第一期望维度序列对填充处理后的第二张量数据进行转置处理，将第二张量数据的前两维与后两维顺序调换回原来顺序。

[0078] 由此，完成第二张量数据的四个维度的数据的填充处理，得到满足业务需求的张量数据。

[0079] 步骤208、根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据的第三维度的数据和第四维度的数据进行填充处理。

[0080] 其中，第二张量数据的前两维不需要进行填充处理，直接根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据的第三维度的数据和第四维度的数据进行填充处理。

[0081] 根据第三维度的数据和第四维度的数据的填充配置信息，判断第三维度的数据和第四维度的数据是否需要进行中部数据填充处理。

[0082] 可选的，如果第四维度的数据不需要进行中部数据填充处理，则根据填充配置信息和填充数值信息，对第二张量数据进行填充处理。

[0083] 可选的，如果第四维度的数据需要进行中部数据填充处理，且第三维度的数据不需要进行中部数据填充处理，则根据第二期望维度序列对第二张量数据进行转置处理，以使第二张量数据的第三维度的数据和第四维度的数据互换顺序；根据填充配置信息和填充数值信息，对转置处理后的第二张量数据进行填充处理；根据第二期望维度序列对填充处理后的第二张量数据进行转置处理。

[0084] 可选的，如果第四维度的数据需要进行中部数据填充处理，且第三维度的数据需要进行中部数据填充处理，则根据填充配置信息和所述填充数值信息，对第二张量数据进行第一次填充处理；根据第二期望维度序列对第一次填充处理后的第二张量数据进行转置处理，以使第二张量数据的第三维度的数据和第四维度的数据互换顺序；根据填充配置信息和填充数值信息，对转置处理后的第二张量数据进行第二次填充处理；根据第二期望维度序列对第二次填充处理后的第二张量数据进行转置处理。

[0085] 由此，完成第二张量数据的后两个维度的数据的填充处理，得到满足业务需求的张量数据。

[0086] 本发明实施例提供了一种张量填充方法，根据填充配置信息和填充数值信息，通

过DMA张量转置操作和DMA张量填充操作对张量数据进行维度顺序的调换，并对指定维度进行填充，可以直接利用DMA操作完成四维以内任意维度的张量填充操作，可以利用最多6次DMA张量转置操作加和DMA张量填充操作，或者最少1次DMA张量填充操作，完成张量数据的所有维度的张量填充，提升张量填充效率，极大缩短张量填充所需的时间。

[0087] 实施例三

[0088] 图3为本发明实施例三提供的一种张量填充装置的结构示意图。如图3所示，所述装置可以配置于计算机设备，包括：数据获取模块301、数据确定模块302以及数据填充模块303。

[0089] 其中，数据获取模块301，用于获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量；数据确定模块302，用于根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据；数据填充模块303，用于根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理。

[0090] 本发明实施例提供了一种张量填充装置，通过获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量，并根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据，然后根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理，解决了现有技术对于维度大于2的原始张量数据，无法利用DMA操作直接完成所有维度的张量填充的问题，可以直接利用DMA操作完成预设维度数量以内任意维度的张量填充操作，提升张量填充效率，极大缩短张量填充所需的时间。

[0091] 在上述各实施例的基础上，数据确定模块302可以包括：维度补充单元，用于当第一张量数据的维度小于预设维度数量时，根据维度和预设维度补充规则，将第一张量数据转换为预设维度数量的第二张量数据。

[0092] 在上述各实施例的基础上，预设维度数量可以为四维；第二张量数据可以包括：第一维度的数据、第二维度的数据、第三维度的数据、以及第四维度的数据；其中，第一维度为最高维，第四维度为最低维。

[0093] 在上述各实施例的基础上，数据填充模块303可以包括：填充判断单元，用于根据第一维度的数据和第二维度的数据的填充配置信息，判断第一维度的数据和第二维度的数据是否需要进行填充处理；第一填充单元，用于如果第一维度的数据和所述第二维度的数据需要进行填充处理，则根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据的第三维度的数据和第四维度的数据进行填充处理；第一转置单元，用于根据第一期望维度序列对填充处理后的第二张量数据进行转置处理，以使第二张量数据的第一维度的数据、第二维度的数据与第三维度的数据、第四维度的数据顺序互换；第二填充单元，用于根据填充配置信息、填充数值信息、以及张量填充规则，对顺序互换后的第二张量数据的第三维度的数据和第四维度的数据进行填充处理；第二转置单元，用于根据第一期望维度序列对填充处理后的第二张量数据进行转置处理。

[0094] 在上述各实施例的基础上，第一填充单元可以包括：填充判断子单元，用于根据第三维度的数据和第四维度的数据的填充配置信息，判断第三维度的数据和第四维度的数据是否需要进行中部数据填充处理；第一填充子单元，用于如果第四维度的数据不需要进行

中部数据填充处理，则根据填充配置信息和填充数值信息，对第二张量数据进行填充处理。

[0095] 在上述各实施例的基础上，第一填充单元可以还包括：第一转置子单元，用于如果第四维度的数据需要进行中部数据填充处理，且第三维度的数据不需要进行中部数据填充处理，则根据第二期望维度序列对第二张量数据进行转置处理，以使第二张量数据的第三维度的数据和第四维度的数据互换顺序；第二填充子单元，用于根据填充配置信息和填充数值信息，对转置处理后的第二张量数据进行填充处理；第二转置子单元，用于根据第二期望维度序列对填充处理后的第二张量数据进行转置处理。

[0096] 在上述各实施例的基础上，第一填充单元可以还包括：第三填充子单元，用于如果第四维度的数据需要进行中部数据填充处理，且第三维度的数据需要进行中部数据填充处理，则根据填充配置信息和填充数值信息，对第二张量数据进行第一次填充处理；第三转置子单元，用于根据第二期望维度序列对第一次填充处理后的第二张量数据进行转置处理，以使第二张量数据的第三维度的数据和第四维度的数据互换顺序；第四填充子单元，用于根据填充配置信息和填充数值信息，对转置处理后的第二张量数据进行第二次填充处理；第四转置子单元，用于根据第二期望维度序列对第二次填充处理后的第二张量数据进行转置处理。

[0097] 上述张量填充装置可执行本发明任意实施例所提供的张量填充方法，具备执行张量填充方法相应的功能模块和有益效果。

[0098] 实施例四

[0099] 图4为本发明实施例四提供的一种计算机设备的结构示意图。图4示出了适于用来实现本发明实施方式的示例性计算机设备412的框图。图4显示的计算机设备412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。计算机设备412可以为一种终端设备或者服务器。

[0100] 如图4所示，计算机设备412以通用计算设备的形式表现。计算机设备412的组件可以包括但不限于：一个或者多个处理器或者处理单元416，系统存储器428，连接不同系统组件(包括系统存储器428和处理单元416)的总线418。

[0101] 总线418表示几类总线结构中的一种或多种，包括存储器总线或者存储器控制器，外围总线，图形加速端口，处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说，这些体系结构包括但不限于工业标准体系结构(ISA)总线，微通道体系结构(MAC)总线，增强型ISA总线、视频电子标准协会(VESA)局域总线以及外围组件互连(PCI)总线。

[0102] 计算机设备412典型地包括多种计算机系统可读介质。这些介质可以是任何能够被计算机设备412访问的可用介质，包括易失性和非易失性介质，可移动的和不可移动的介质。

[0103] 系统存储器428可以包括易失性存储器形式的计算机系统可读介质，例如随机存取存储器(RAM)430和/或高速缓存存储器432。计算机设备412可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例，存储系统434可以用于读写不可移动的、非易失性磁介质(图4未显示，通常称为“硬盘驱动器”)。尽管图4中未示出，可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器，以及对可移动非易失性光盘(例如CD-ROM，DVD-ROM或者其它光介质)读写的光盘驱动器。在这些情况下，每个驱动器可以通过一个或者多个数据介质接口与总线418相连。系统存储器428可以包括至少

一个程序产品，该程序产品具有一组(例如至少一个)程序模块，这些程序模块被配置以执行本发明各实施例的功能。

[0104] 具有一组(至少一个)程序模块442的程序/实用工具440，可以存储在例如系统存储器428中，这样的程序模块442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。程序模块442通常执行本发明所描述的实施例中的功能和/或方法。

[0105] 计算机设备412也可以与一个或多个外部设备414(例如键盘、指向设备、显示器424等)通信，还可与一个或者多个使得用户能与该计算机设备412交互的设备通信，和/或与使得该计算机设备412能与一个或多个其它计算设备进行通信的任何设备(例如网卡，调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口422进行。并且，计算机设备412还可以通过网络适配器420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420通过总线418与计算机设备412的其它模块通信。应当明白，尽管图4中未示出，可以结合计算机设备412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。

[0106] 处理单元416通过运行存储在系统存储器428中的程序，从而执行各种功能应用以及数据处理，例如实现本发明实施例所提供的张量填充方法。也即，获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量；根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据；根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理。

[0107] 实施例五

[0108] 本发明实施例五提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现如本申请所有发明实施例提供的张量填充方法。也即，获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量；根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据；根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理。

[0109] 可以采用一个或多个计算机可读的介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机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、便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0110] 计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括——但不限于——电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、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。

[0111] 计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括——但不限于——无线、电线、光缆、RF等等，或者上述的任意合适的组合。

[0112] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到用户计算机，或者，可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0113] 注意，上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解，本发明不限于这里所述的特定实施例，对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此，虽然通过以上实施例对本发明进行了较为详细的说明，但是本发明不仅仅限于以上实施例，在不脱离本发明构思的情况下，还可以包括更多其他等效实施例，而本发明的范围由所附的权利要求范围决定。

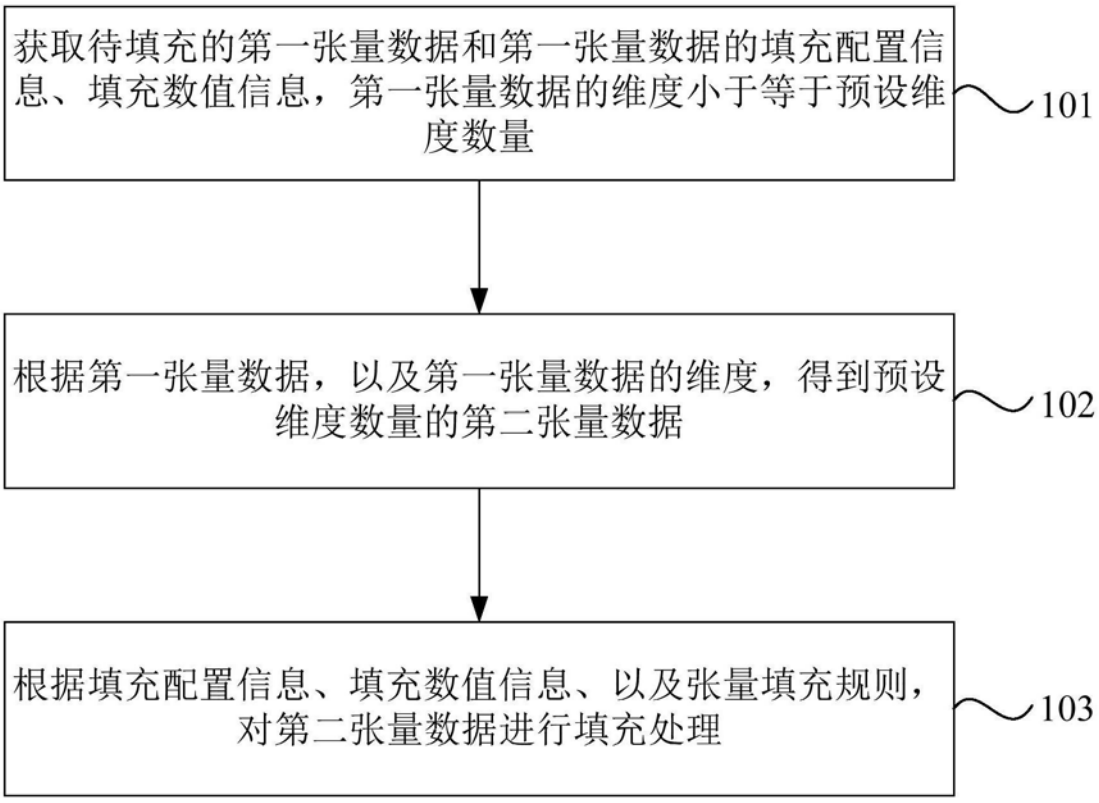

获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量
101
根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据
102
根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据进行填充处理
103

图1

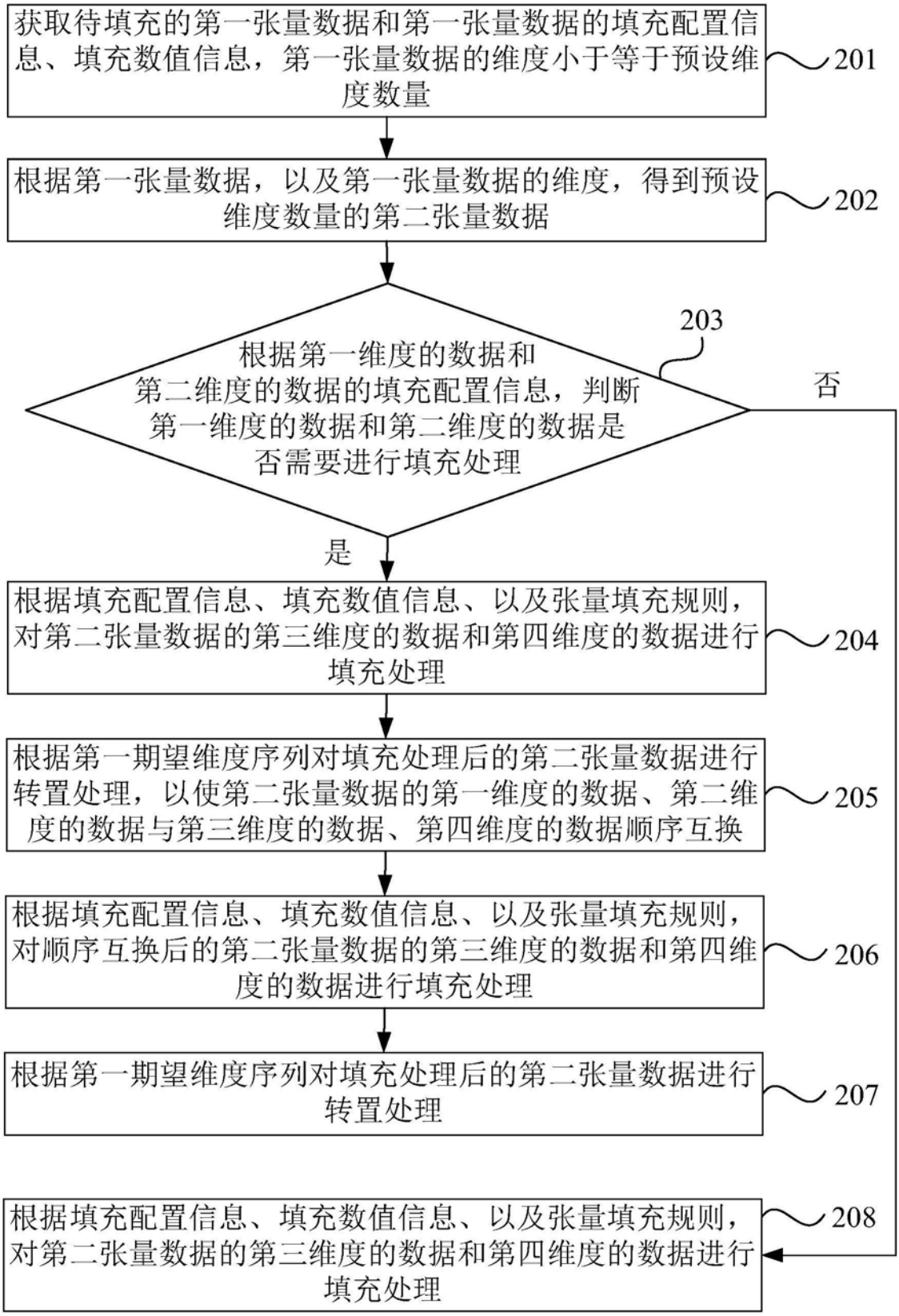

获取待填充的第一张量数据和第一张量数据的填充配置信息、填充数值信息，第一张量数据的维度小于等于预设维度数量
201
根据第一张量数据，以及第一张量数据的维度，得到预设维度数量的第二张量数据
202
根据第一维度的数据和第二维度的数据的填充配置信息，判断第一维度的数据和第二维度的数据是否需要进行填充处理
203
否
是
根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据的第三维度的数据和第四维度的数据进行填充处理
204
根据第一期望维度序列对填充处理后的第二张量数据进行转置处理，以使第二张量数据的第一维度的数据、第二维度的数据与第三维度的数据、第四维度的数据顺序互换
205
根据填充配置信息、填充数值信息、以及张量填充规则，对顺序互换后的第二张量数据的第三维度的数据和第四维度的数据进行填充处理
206
根据第一期望维度序列对填充处理后的第二张量数据进行转置处理
207
根据填充配置信息、填充数值信息、以及张量填充规则，对第二张量数据的第三维度的数据和第四维度的数据进行填充处理
208

图2

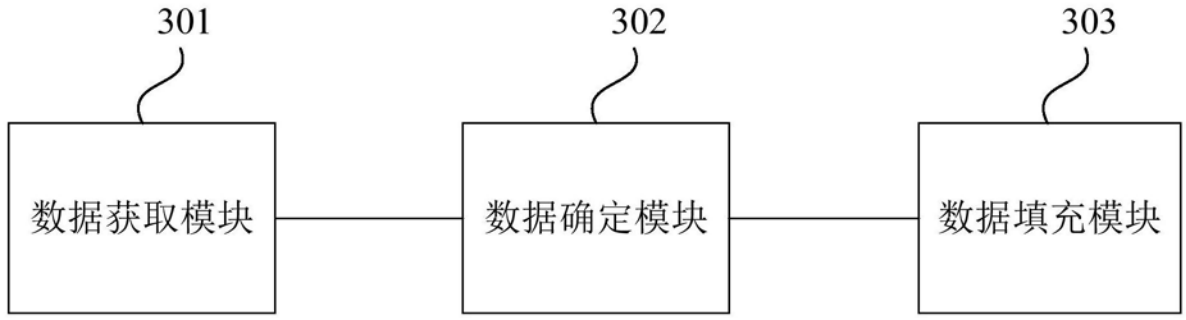

301
302
303
数据获取模块
数据确定模块
数据填充模块

图3

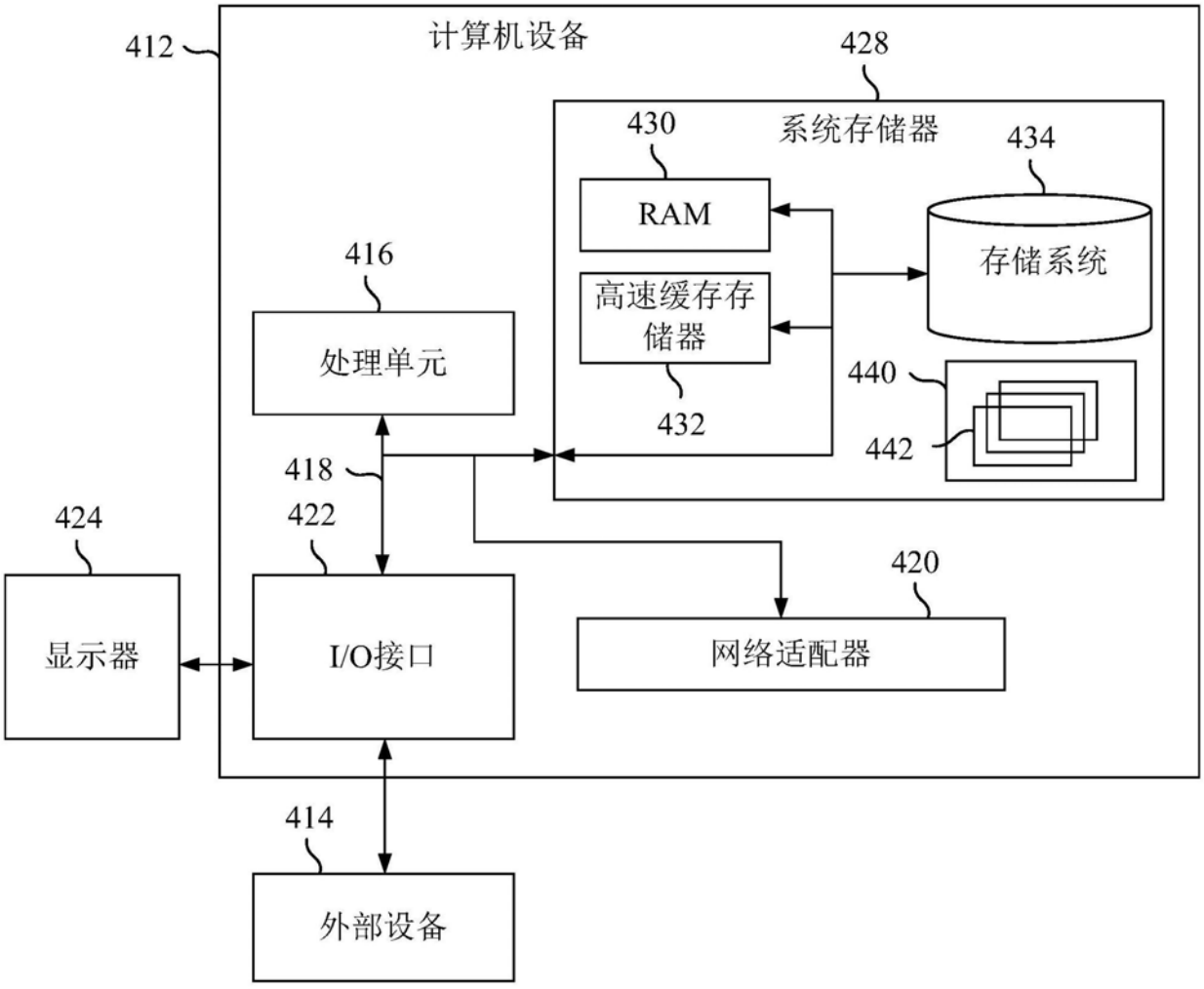

412
计算机设备
428
430
系统存储器
434
RAM
存储系统
416
高速缓存存储器
处理单元
440
442
432
418
424
422
420
显示器
I/O接口
网络适配器
414
外部设备

图4