(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 108304265 B
(45)授权公告日 2022.02.01

(21)申请号 201810064209.2

(22)申请日 2018.01.23

(65)同一申请的已公布的文献号
申请公布号 CN 108304265 A

(43)申请公布日 2018.07.20

(73)专利权人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 黄凯宁 朱晓龙 梅利健 黄生辉 王一同 罗镜民

(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
代理人 刘映东

(51)Int.Cl.
G06F 9/50(2006.01)
G06N 3/04(2006.01)

(56)对比文件
CN 106447035 A,2017.02.22
CN 105809248 A,2016.07.27
JP 3816762 B2,2006.08.30
US 2017228643 A1,2017.08.10
ma1998.《卷积神经网络内存占用计算》.《CSDN博客》.2017,第1-26页.
Minsoo Rhu 等.vDNN: Virtualized Deep Neural Networks for Scalable, Memory-Efficient Neural Network Design.《IEEE Xplore》.2016,第1-13页.
Linnan Wang 等.SuperNeurons: Dynamic GPU Memory Management for Training Deep Neural Networks.《arXiv:1801.04380v1》.2018,第1-13页.

审查员 冷小素

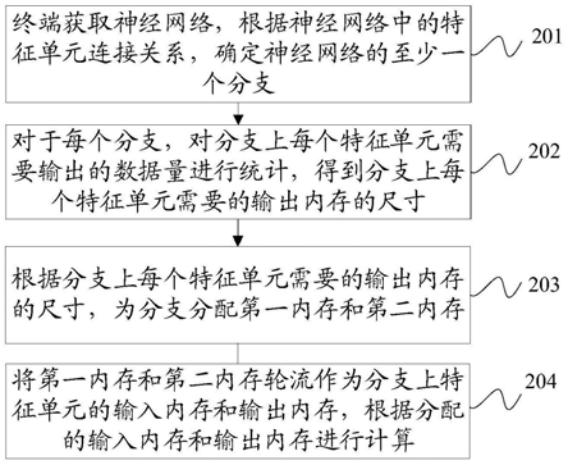
权利要求书3页 说明书10页 附图3页

(54)发明名称

内存管理方法、装置及存储介质

(57)摘要

本发明实施例公开了一种内存管理方法、装置及存储介质，属于计算机技术领域。方法包括：根据神经网络中的特征单元连接关系，确定神经网络的至少一个分支；对于每个分支，根据分支上每个特征单元需要的输出内存的尺寸，为分支分配第一内存和第二内存，第一内存尺寸不小于第二内存尺寸，第一内存尺寸和第二内存尺寸均不小于分支需要的其他内存尺寸；将第一内存和第二内存轮流作为分支上特征单元的输入内存和输出内存。本发明实施例针对神经网络中的每个分支仅分配两个内存，轮流作为特征单元的输入内存和输出内存，既可以保证计算的正常进行，而且实现了内存复用，节省了占用的内存，降低了内存需求，保证在终端上能够正常实现神经网络。



CN 108304265 B

1.一种内存管理方法,其特征在于,应用于终端中,所述方法包括:

根据神经网络中的特征单元连接关系,确定所述神经网络的至少一个分支,所述神经网络包括按照顺序排列的多个网络层,每个网络层包括至少一个特征单元,每个分支由位于不同网络层的多个特征单元连接构成;

对于每个分支,根据所述分支上每个特征单元需要的输出内存的尺寸,为所述分支分配第一内存和第二内存,第一内存尺寸不小于第二内存尺寸,所述第一内存尺寸和所述第二内存尺寸均不小于所述分支需要的其他内存尺寸;

当所述第一内存尺寸与所述第二内存尺寸相等时,将所述第一内存和所述第二内存轮流作为所述分支上特征单元的输入内存和输出内存;

当所述第一内存尺寸大于所述第二内存尺寸时,确定所述分支上需要的输出内存的尺寸为所述第一内存尺寸的指定特征单元,将所述第一内存作为所述指定特征单元的输出内存,按照所述分支上每个特征单元与所述指定特征单元之间的位置关系,将所述第一内存和所述第二内存作为所述分支上除指定特征单元以外的每个特征单元的输入内存或输出内存,以使所述第一内存和所述第二内存轮流作为所述分支上特征单元的输入内存和输出内存。

2.根据权利要求1所述的方法,其特征在于,所述根据所述分支上每个特征单元需要的输出内存的尺寸,为所述分支分配第一内存和第二内存之前,所述方法还包括:

对所述分支上每个特征单元需要输出的数据量进行统计,得到所述分支上每个特征单元需要的输出内存的尺寸。

3.根据权利要求1所述的方法,其特征在于,所述根据所述分支上每个特征单元需要的输出内存的尺寸,为所述分支分配第一内存和第二内存,包括:

按照从大到小的顺序,对所述分支上每个特征单元需要的输出内存的尺寸进行排序;

将排在第一位的尺寸作为第一内存尺寸,将排在第二位的尺寸作为第二内存尺寸;

为所述分支分配与所述第一内存尺寸匹配的第一内存和与所述第二内存尺寸匹配的第二内存。

4.根据权利要求1所述的方法,其特征在于,所述将所述第一内存和所述第二内存轮流作为所述分支上特征单元的输入内存和输出内存,包括:

将所述第一内存作为所述分支上第一特征单元的输出内存,且作为第二特征单元的输入内存,将所述第二内存作为所述第二特征单元的输出内存;

其中,所述第一特征单元为所述分支上的任一特征单元,所述第二特征单元为所述分支上所述第一特征单元的下一个特征单元。

5.根据权利要求1-4任一项所述的方法,其特征在于,所述将所述第一内存和所述第二内存轮流作为所述分支上特征单元的输入内存和输出内存之后,所述方法还包括:

在基于所述分支进行计算的过程中,将第一特征单元输出的数据存储于所述第一特征单元的输出内存中;

从所述第一特征单元的输出内存中提取所述第一特征单元输出的数据,通过第二特征单元对提取的数据进行计算,将所述第二特征单元输出的数据存储于所述第二特征单元的输出内存中;

其中,所述第一特征单元为所述分支上的任一特征单元,所述第二特征单元为所述分

支上所述第一特征单元的下一个特征单元，所述第一特征单元的输出内存与所述第二特征单元的输入内存相同。

6.一种内存管理装置，其特征在于，应用于终端中，所述装置包括：

确定模块，用于根据神经网络中的特征单元连接关系，确定所述神经网络的至少一个分支，所述神经网络包括按照顺序排列的多个网络层，每个网络层包括至少一个特征单元，每个分支由位于不同网络层的多个特征单元连接构成；

分配模块，用于对于每个分支，根据所述分支上每个特征单元需要的输出内存的尺寸，为所述分支分配第一内存和第二内存，第一内存尺寸不小于第二内存尺寸，所述第一内存尺寸和所述第二内存尺寸均不小于所述分支需要的其他内存尺寸；

所述分配模块，还用于当所述第一内存尺寸与所述第二内存尺寸相等时，将所述第一内存和所述第二内存轮流作为所述分支上特征单元的输入内存和输出内存；

所述分配模块，还用于当所述第一内存尺寸大于所述第二内存尺寸时，确定所述分支上需要的输出内存的尺寸为所述第一内存尺寸的指定特征单元，将所述第一内存作为所述指定特征单元的输出内存，按照所述分支上每个特征单元与所述指定特征单元之间的位置关系，将所述第一内存和所述第二内存作为所述分支上除指定特征单元以外的每个特征单元的输入内存或输出内存，以使所述第一内存和所述第二内存轮流作为所述分支上特征单元的输入内存和输出内存。

7.根据权利要求6所述的装置，其特征在于，所述装置还包括：

统计模块，用于对所述分支上每个特征单元需要输出的数据量进行统计，得到所述分支上每个特征单元需要的输出内存的尺寸。

8.根据权利要求6所述的装置，其特征在于，所述分配模块，包括：

排序单元，用于按照从大到小的顺序，对所述分支上每个特征单元需要的输出内存的尺寸进行排序；

第一分配单元，用于将排在第一位的尺寸作为第一内存尺寸，将排在第二位的尺寸作为第二内存尺寸，为所述分支分配与所述第一内存尺寸匹配的第一内存和与所述第二内存尺寸匹配的第二内存。

9.根据权利要求6所述的装置，其特征在于，所述分配模块，包括：

第三分配单元，用于当所述第一内存尺寸大于所述第二内存尺寸时，确定所述分支上需要的输出内存的尺寸为所述第一内存尺寸的指定特征单元；将所述第一内存作为所述指定特征单元的输出内存；按照所述分支上每个特征单元与所述指定特征单元之间的位置关系，将所述第一内存和所述第二内存作为所述分支上除所述指定特征单元以外的每个特征单元的输入内存或输出内存，以使所述第一内存和所述第二内存轮流作为所述分支上特征单元的输入内存和输出内存。

10.根据权利要求6-9任一项所述的装置，其特征在于，所述装置还包括：

第一计算模块，用于在基于所述分支进行计算的过程中，将第一特征单元输出的数据存储于所述第一特征单元的输出内存中；

第二计算模块，用于从所述第一特征单元的输出内存中提取所述第一特征单元输出的数据，通过第二特征单元对提取的数据进行计算，将所述第二特征单元输出的数据存储于所述第二特征单元的输出内存中；

其中，所述第一特征单元为所述分支上的任一特征单元，所述第二特征单元为所述分支上所述第一特征单元的下一个特征单元，所述第一特征单元的输出内存与所述第二特征单元的输入内存相同。

11.一种内存管理装置，其特征在于，所述内存管理装置包括处理器和存储器，所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集，所述指令、所述程序、所述代码集或所述指令集由所述处理器加载并执行以实现如权利要求1至5任一权利要求所述的内存管理方法中所执行的操作。

12.一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中存储有至少一条指令、至少一段程序、代码集或指令集，所述指令、所述程序、所述代码集或所述指令集由处理器加载并执行以实现如权利要求1至5任一权利要求所述的内存管理方法中所执行的操作。

内存管理方法、装置及存储介质

技术领域

[0001] 本发明实施例涉及计算机技术领域，特别涉及一种内存管理方法、装置及存储介质。

背景技术

[0002] 近年来，深度学习在语音识别和计算机视觉等领域得到了广泛的应用，并且随着深度学习的快速发展，驱动着AI(Artificial Intelligence，人工智能)算法不断涌现，并迅速地改变了技术的发展方向和人们的生活。

[0003] 深度学习离不开神经网络的实现，传统的神经网络通常在服务器端实现，然而，随着AI算法的不断发展，在服务器端实现神经网络的方式已不能满足人们日益增长的业务需求，亟需提出在终端实现神经网络的方式。

[0004] 但是，神经网络中包括多个网络层，每个网络层又包括多个特征单元，每个特征单元计算完成输出数据时均需要占用一块内存，导致实现神经网络时需要占用大量的内存。而与服务器端相比，终端的内存有限，无法满足神经网络的内存需求，导致在终端实现神经网络的方式难以实现。

发明内容

[0005] 本发明实施例提供了一种内存管理方法、装置及存储介质，可以解决相关技术的问题。所述技术方案如下：

[0006] 第一方面，提供了一种内存管理方法，应用于终端中，所述方法包括：

[0007] 根据神经网络中的特征单元连接关系，确定所述神经网络的至少一个分支，所述神经网络包括按照顺序排列的多个网络层，每个网络层包括至少一个特征单元，每个分支由位于不同网络层的多个特征单元连接构成；

[0008] 对于每个分支，根据所述分支上每个特征单元需要的输出内存的尺寸，为所述分支分配第一内存和第二内存，第一内存尺寸不小于第二内存尺寸，所述第一内存尺寸和所述第二内存尺寸均不小于所述分支需要的其他内存尺寸；

[0009] 将所述第一内存和所述第二内存轮流作为所述分支上特征单元的输入内存和输出内存。

[0010] 第二方面，提供了一种内存管理装置，应用于终端中，所述装置包括：

[0011] 确定模块，用于根据神经网络中的特征单元连接关系，确定所述神经网络的至少一个分支，所述神经网络包括按照顺序排列的多个网络层，每个网络层包括至少一个特征单元，每个分支由位于不同网络层的多个特征单元连接构成；

[0012] 分配模块，用于对于每个分支，根据所述分支上每个特征单元需要的输出内存的尺寸，为所述分支分配第一内存和第二内存，第一内存尺寸不小于第二内存尺寸，所述第一内存尺寸和所述第二内存尺寸均不小于所述分支需要的其他内存尺寸；

[0013] 所述分配模块，还用于将所述第一内存和所述第二内存轮流作为所述分支上特征

单元的输入内存和输出内存。

[0014] 第三方面，提供了一种内存管理装置，所述内存管理装置包括处理器和存储器，所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集，所述指令、所述程序、所述代码集或所述指令集由所述处理器加载并执行以实现如第一方面所述的内存管理方法中所执行的操作。

[0015] 第四方面，提供了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至少一条指令、至少一段程序、代码集或指令集，所述指令、所述程序、所述代码集或所述指令集由处理器加载并执行以实现如第一方面所述的内存管理方法中所执行的操作。

[0016] 本发明实施例提供的技术方案带来的有益效果是：

[0017] 本发明实施例提供的方法、装置及存储介质，通过根据神经网络中的特征单元连接关系，确定神经网络的至少一个分支，对于每个分支，根据分支上每个特征单元需要的输出内存的尺寸，为分支分配第一内存和第二内存，第一内存尺寸和第二内存尺寸均不小于分支需要的其他内存尺寸；将第一内存和第二内存轮流作为分支上特征单元的输入内存和输出内存。本发明实施例针对神经网络中的每个分支仅分配两个内存，轮流作为特征单元的输入内存和输出内存，既可以保证计算的正常进行，而且实现了内存复用，节省了占用的内存，降低了内存需求，保证终端的内存可以满足神经网络的内存需求，在终端上能够正常实现神经网络，从而扩展了终端的功能。

附图说明

[0018] 为了更清楚地说明本发明实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，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。

[0019] 图1是本发明实施例提供的一种神经网络的结构示意图；

[0020] 图2是本发明实施例提供的一种内存管理方法的流程图；

[0021] 图3是本发明实施例提供的一种计算流程的示意图；

[0022] 图4是本发明实施例提供的一种内存管理装置的结构示意图；

[0023] 图5是本发明实施例提供的一种终端的结构示意图。

具体实施方式

[0024] 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，所描述的实施例是本发明一部分实施例，而不是全部的实施例。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，都属于本发明保护的范围。

[0025] 在对本发明实施例进行详细说明之前，首先对本发明实施例涉及的神经网络进行如下说明：

[0026] 图1是本发明实施例提供的一种神经网络的结构示意图，参见图1，神经网络包括多个网络层，每个网络层包括至少一个特征单元，相邻的两个网络层中的特征单元可以相互连接，从而构成一个网络。实现神经网络的过程中，前一个网络层中的特征单元会进行计

算，将输出的数据输入给其连接的、位于下一个网络层中的特征单元，由下一个特征单元进行计算，以此类推，直至位于最后一个网络层的特征单元输出计算结果。

[0027] 实际应用中，该神经网络可以为CNN(Convolutional Neural Network，卷积神经网络)、RNN(Recurrent Neural Network，循环神经网络)或者DNN(DeepNeural Network，深度神经网络)等多种类型的神经网络，针对不同类型的神经网络，其所包括的网络层类型不同，网络层中的特征单元类型也不同。

[0028] 例如，CNN包括输入层、卷积层、采样层和输出层等多个网络层，每个网络层中包括特征图，特征图中包括神经元，则本发明实施例中的特征单元可以指特征图或者神经元。

[0029] 本发明实施例提供的神经网络可以应用于多种场景下，例如人脸识别、垃圾邮件筛选、信息推荐等，如在人脸识别的场景下，将采集到的图像输入到神经网络中进行计算，以判别该图像中是否包含人脸，或者在信息推荐的场景下，将多种信息输入到神经网络中进行计算，筛选出用户可能感兴趣的信息，推荐给用户。

[0030] 图2是本发明实施例提供的一种内存管理方法的流程图，该内存管理方法的执行主体为终端，参见图2，该方法包括：

[0031] 201、终端获取神经网络，根据神经网络中的特征单元连接关系，确定神经网络的至少一个分支。

[0032] 其中，神经网络包括按照顺序排列的多个网络层，每个网络层包括至少一个特征单元，相邻的两个网络层中的特征单元可以相互连接，从而构成一个网络。对于位于某一网络层的某一特征单元来说，该特征单元可以与所在网络层的上一个网络层中的一个或多个特征单元连接，也可以与所在网络层的下一个网络层中的一个或多个特征单元连接。

[0033] 因此，在神经网络中存在着很多分支，可以根据多个特征单元的连接关系将神经网络划分为至少一个分支，每个分支由位于不同网络层的特征单元连接构成。

[0034] 例如，参见图1，可以将神经网络划分为“11-21-31-41”、“12-22-32-41”，“13-23-32-41”等分支。

[0035] 202、对于每个分支，对分支上每个特征单元需要输出的数据量进行统计，得到分支上每个特征单元需要的输出内存的尺寸。

[0036] 对于每个分支来说，在基于该分支进行计算的过程中需要占用内存。针对该分支中相邻的两个特征单元来说，前一个特征单元会将计算以后输出的数据存储于内存中，下一个特征单元从该内存中提取前一个特征单元输出的数据，并根据提取的数据进行计算，该内存即为前一个特征单元的输出内存，下一个特征单元的输入内存。

[0037] 相关技术中会在服务器端实现神经网络，计算过程中，会为每个特征单元分配输入内存和输出内存，导致占用了大量的内存。由于服务器端的内存较大，神经网络的计算过程不会受到内存大小的限制，不会影响服务器端的正常运行。

[0038] 而本发明实施例中，考虑到终端的内存有限，若要在终端上实现神经网络则容易占用过多的内存，而且占用内存过多时操作系统可能会回收已分配的内存，这会影响到终端的正常运行。因此，为了节省内存，在实际进行计算之前，将神经网络划分为至少一个分支，对于神经网络的每个分支仅分配两个内存，以分配的两个内存轮流作为输出内存和输入内存，而不再为每个特征单元均分配输入内存和输出内存。

[0039] 为此，在分配内存时，先确定分支上每个特征单元需要输出的数据量，该数据量即

为对应的特征单元需要的输出内存的尺寸，以此获取到分支上每个特征单元需要的输出内存的尺寸。

[0040] 203、根据分支上每个特征单元需要的输出内存的尺寸，为分支分配第一内存和第二内存。

[0041] 其中，第一内存尺寸不小于第二内存尺寸，第一内存尺寸和第二内存尺寸均不小于分支需要的其他内存尺寸。

[0042] 为了保证分配的内存足够存储分支上每个特征单元输出的数据，可以在获取到每个特征单元需要的输出内存的尺寸后，按照从大到小的顺序，对每个特征单元需要的输出内存的尺寸进行排序，将排在第一位的尺寸作为第一内存尺寸，将排在第二位的尺寸作为第二内存尺寸，从而为该分支分配与第一内存尺寸匹配的第一内存和与第二内存尺寸匹配的第二内存。

[0043] 例如，参见图1，分支“11-21-31-41”上4个特征单元需要输出的数据量依次为20M、30M、20M、50M，说明特征单元11需要的输出内存的尺寸为20M，特征单元21需要的输出内存的尺寸为30M，特征单元31需要的输出内存的尺寸为20M，特征单元41需要的输出内存的尺寸为50M，则经过排序后可以确定第一内存尺寸为50M，第二内存尺寸为30M，因此为该分支分配一块50M的内存和一块30M的内存。

[0044] 204、将第一内存和第二内存轮流作为分支上特征单元的输入内存和输出内存，根据分配的输入内存和输出内存进行计算。

[0045] 在第一种可能实现方式中，当第一内存尺寸与第二内存尺寸相等时，表示该分支上至少有两个特征单元需要的输出内存的尺寸最大，该最大值等于第一内存尺寸，而其他特征单元需要的输出内存的尺寸均不大于第一内存尺寸。则分配内存的规则为：每个特征单元的输入内存和输出内存不同，前一个特征单元的输出内存为下一个特征单元的输入内存。

[0046] 按照该规则进行分配时，只需为某一个特征单元选取第一内存作为输出内存，或者选取第二内存作为输出内存，即可根据这些特征单元之间的位置关系确定其他特征单元的输入内存和输出内存是第一内存还是第二内存。

[0047] 以第一特征单元为例，将第一内存作为第一特征单元的输出内存，且作为第二特征单元的输入内存，将第二内存作为第二特征单元的输出内存；其中，第一特征单元为分支上的任一特征单元，第二特征单元为分支上第一特征单元的下一个特征单元。

[0048] 在第二种可能方式中，当第一内存尺寸大于第二内存尺寸时，表示该分支上需要的最大输出内存的尺寸与次大的输出内存尺寸不同，则该分支上只有一个特征单元需要的输出内存的尺寸最大，且最大值为第一内存尺寸并大于第二内存尺寸，因此如果将第二内存作为这个特征单元的输出内存会导致该特征单元输出的数据量大于第二内存尺寸，进而导致数据无法正常存储。

[0049] 为了避免上述情况，分配内存的规则为：每个特征单元的输入内存和输出内存不同，前一个特征单元的输出内存为下一个特征单元的输入内存，且需要的输出内存的尺寸为第一内存尺寸的特征单元的输出内存应当为第一内存。

[0050] 因此，先确定分支上需要的输出内存的尺寸为第一内存尺寸的指定特征单元，将第一内存作为指定特征单元的输出内存，再按照分支上每个特征单元与指定特征单元之间

的位置关系,将第一内存和第二内存作为分支上除指定特征单元以外的每个特征单元的输入内存或输出内存,以使第一内存和第二内存轮流作为分支上特征单元的输入内存和输出内存。

[0051] 例如,如果指定特征单元在分支上的排列顺序为偶数,则将第一内存作为分支上偶数位的特征单元的输出内存,将第二内存作为分支上奇数位的特征单元的输出内存。如果指定特征单元在分支上的排列顺序为奇数,则将第一内存作为分支上奇数位的特征单元的输出内存,将第二内存作为分支上偶数位的特征单元的输出内存。

[0052] 在一种可能实现方式中,在上述分配内存过程中,终端可以调用OpenCL(Open Computing Language,开放运算语言)提供的clCreateBuffer函数来分配内存。

[0053] 确定该神经网络中每个分支上的每个特征单元的输入内存和输出内存之后,即可根据分配好的输入内存和输出内存,基于该神经网络进行计算。在计算过程中,以第一特征单元为例,将第一特征单元输出的数据存储于第一特征单元的输出内存中,再从第一特征单元的输出内存中提取第一特征单元输出的数据,通过第二特征单元对提取的数据进行计算,将第二特征单元输出的数据存储于第二特征单元的输出内存中。其中,第一特征单元为任一分支上的任一特征单元,第二特征单元为在该分支上第一特征单元的下一个特征单元,第一特征单元的输出内存与第二特征单元的输入内存相同。

[0054] 基于上述图1所示的神经网络,以分支“11-21-31-41”为例,将第一内存作为特征单元21和特征单元41的输入内存,作为特征单元11和特征单元31的输出内存,将第二内存作为特征单元11和特征单元31的输入内存,作为特征单元21和特征单元41的输出内存。则参见图3,计算过程包括以下步骤:

[0055] 1、在第二内存中输入数据a,则特征单元11从第二内存中提取数据a,进行计算后得到数据b,将数据b存储于第一内存中。

[0056] 2、特征单元21从第一内存中提取数据b,进行计算后得到数据c,将数据c存储于第二内存中,此时第二内存中的数据由a变成c。

[0057] 3、特征单元31从第二内存中提取数据c,进行计算后得到数据d,将数据d存储于第一内存中,此时第一内存中的数据由b变成d。

[0058] 4、特征单元41从第一内存中提取数据d,进行计算后得到数据e,存储于第二内存中,此时第二内存中的数据由d变成e。

[0059] 需要说明的是,神经网络中的任两个分支可能会完全分离,也可能会相互交叉,具有共同的特征单元。那么,对于神经网络中的某一个特征单元来说,该特征单元可能会位于多个分支上,针对不同的分支,该特征单元会设置不同的输入内存和输出内存。

[0060] 也即是,该特征单元的输入端可能会连接位于所在网络层的前一个网络层的多个上层特征单元,从而构成多个分支,则针对不同的分支为该特征单元分配不同的输入内存。计算时该特征单元会从多个输入内存中分别提取数据,根据提取的多个数据进行计算。

[0061] 该特征单元的输出端也可能会连接位于所在网络层的下一个网络层的多个下层特征单元,从而构成多个分支,针对不同的分支为该特征单元分配不同的输出内存。计算时该特征单元将计算出的数据分别存储于分配的多个输出内存中,供多个下层特征单元提取。

[0062] 需要说明的是,实际应用中可以在终端上针对神经网络部署运算框架,该运算框

架针对终端平台开发,对编译环境和硬件环境没有过多要求,编译较为容易,调用该运算框架时,可以实现本发明实施例提供的分配内存和计算的步骤,从而在终端上实现神经网络的计算。考虑到神经网络中会包括大量的并行运算,而CPU(Central Processing Unit,中央处理器)的性能难以满足运算要求,GPU(Graphic Processing Unit,图形处理器)由于其结构特性,更适合进行并行运算,因此,在终端上实现神经网络时可以使用CPU实现对内存的优化,并使用GPU进行核心运算,从而能够提高计算效率,更为符合神经网络的技术发展趋势。

[0063] 相应地,针对该神经网络,终端可以通过CPU执行上述分配内存的步骤,确定每个特征单元的输入内存和输出内存,并下发计算任务,该计算任务中可以包括每个特征单元的输入内存和输出内存,则终端通过GPU领取CPU下发的任务,按照该任务中分配好的内存进行计算。

[0064] 在一种可能实现方式中,CPU为GPU下发计算任务时,可以调用OpenCL提供的clCreateCommandQueue函数创建任务队列,调用OpenCL提供的addToCommandQueue函数在任务队列中添加任务,由GPU从创建的任务队列中领取任务并进行计算。

[0065] 本发明实施例提供的方法,通过根据神经网络中的特征单元连接关系,确定神经网络的至少一个分支,对于每个分支,根据分支上每个特征单元需要的输出内存的尺寸,为分支分配第一内存和第二内存,第一内存尺寸和第二内存尺寸均不小于分支需要的其他内存尺寸;将第一内存和第二内存轮流作为分支上特征单元的输入内存和输出内存。本发明实施例针对神经网络中的每个分支仅分配两个内存,轮流作为特征单元的输入内存和输出内存,既可以保证计算的正常进行,而且实现了内存复用,节省了占用的内存,降低了内存需求,保证终端的内存可以满足神经网络的内存需求,在终端上能够正常实现神经网络,从而扩展了终端的功能。

[0066] 图4是本发明实施例提供的一种内存管理装置的结构示意图。参见图4,该装置应用于终端中,该装置包括:

[0067] 确定模块401,用于执行上述实施例中确定神经网络的至少一个分支的步骤。

[0068] 分配模块402,用于执行上述实施例中为分支分配第一内存和第二内存的步骤。

[0069] 分配模块402,还用于执行上述实施例中将第一内存和第二内存轮流作为分支上特征单元的输入内存和输出内存的步骤。

[0070] 可选地,该装置还包括:

[0071] 统计模块,用于执行上述实施例中对分支上每个特征单元需要输出的数据量进行统计的步骤。

[0072] 可选地,分配模块402,包括:

[0073] 排序单元,用于执行上述实施例中对分支上每个特征单元需要的输出内存的尺寸进行排序的步骤;

[0074] 第一分配单元,用于执行上述实施例中确定第一内存尺寸和第二内存尺寸,并分配第一内存和第二内存的步骤。

[0075] 可选地,分配模块402,包括:

[0076] 第二分配单元,用于执行上述实施例中当第一内存尺寸与第二内存尺寸相等时确定每个特征单元的输入内存和输出内存的步骤。

[0077] 可选地，分配模块402，包括：

[0078] 第三分配单元，用于执行上述实施例中当第一内存尺寸大于第二内存尺寸时确定每个特征单元的输入内存和输出内存的步骤。

[0079] 可选地，该装置还包括：

[0080] 第一计算模块，用于执行上述实施例中将第一特征单元输出的数据存储于第一特征单元的输出内存中的步骤；

[0081] 第二计算模块，用于执行上述实施例中从第一特征单元的输出内存中提取数据后计算并存储于第二特征单元的输出内存中的步骤。

[0082] 需要说明的是：上述实施例提供的内存管理装置在管理内存时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将终端的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的内存管理装置与内存管理方法实施例属于同一构思，其具体实现过程详见方法实施例，这里不再赘述。

[0083] 图5示出了本发明一个示例性实施例提供的终端500的结构框图。该终端500可以是便携式移动终端，比如：智能手机、平板电脑、MP3播放器(Moving Picture Experts Group Audio Layer III，动态影像专家压缩标准音频层面3)、MP4(Moving Picture Experts Group Audio Layer IV，动态影像专家压缩标准音频层面4)播放器、笔记本电脑或台式电脑。终端500还可能被称为用户设备、便携式终端、膝上型终端、台式终端等其他名称。

[0084] 通常，终端500包括有：处理器501和存储器502。

[0085] 处理器501可以包括一个或多个处理核心，比如4核心处理器、5核心处理器等。处理器501可以采用DSP(Digital Signal Processing，数字信号处理)、FPGA(Field－Programmable Gate Array，现场可编程门阵列)、PLA(Programmable Logic Array，可编程逻辑阵列)中的至少一种硬件形式来实现。处理器501也可以包括主处理器和协处理器，主处理器是用于对在唤醒状态下的数据进行处理的处理器，也称CPU(Central Processing Unit，中央处理器)；协处理器是用于对在待机状态下的数据进行处理的低功耗处理器。在一些实施例中，处理器501可以在集成有GPU(Graphics Processing Unit，图像处理器)，GPU用于负责显示屏所需要显示的内容的渲染和绘制。一些实施例中，处理器501还可以包括AI(Artificial Intelligence，人工智能)处理器，该AI处理器用于处理有关机器学习的计算操作。

[0086] 存储器502可以包括一个或多个计算机可读存储介质，该计算机可读存储介质可以是非暂态的。存储器502还可包括高速随机存取存储器，以及非易失性存储器，比如一个或多个磁盘存储设备、闪存存储设备。在一些实施例中，存储器502中的非暂态的计算机可读存储介质用于存储至少一个指令，该至少一个指令用于被处理器501所执行以实现本申请中方法实施例提供的内存管理方法。

[0087] 在一些实施例中，终端500还可选包括有：外围设备接口503和至少一个外围设备。处理器501、存储器502和外围设备接口503之间可以通过总线或信号线相连。各个外围设备可以通过总线、信号线或电路板与外围设备接口503相连。具体地，外围设备包括：射频电路504、触摸显示屏505、摄像头506、音频电路507、定位组件508和电源509中的至少一种。

[0088] 外围设备接口503可被用于将I/O(Input/Output,输入/输出)相关的至少一个外围设备连接到处理器501和存储器502。在一些实施例中,处理器501、存储器502和外围设备接口503被集成在同一芯片或电路板上;在一些其他实施例中,处理器501、存储器502和外围设备接口503中的任意一个或两个可以在单独的芯片或电路板上实现,本实施例对此不加以限定。

[0089] 射频电路504用于接收和发射RF(Radio Frequency,射频)信号,也称电磁信号。射频电路504通过电磁信号与通信网络以及其他通信设备进行通信。射频电路504将电信号转换为电磁信号进行发送,或者,将接收到的电磁信号转换为电信号。可选地,射频电路504包括:天线系统、RF收发器、一个或多个放大器、调谐器、振荡器、数字信号处理器、编解码芯片组、用户身份模块卡等等。射频电路504可以通过至少一种无线通信协议来与其它终端进行通信。该无线通信协议包括但不限于:城域网、各代移动通信网络(2G、3G、4G及5G)、无线局域网和/或WiFi(Wireless Fidelity,无线保真)网络。在一些实施例中,射频电路504还可以包括NFC(Near Field Communication,近距离无线通信)有关的电路,本申请对此不加以限定。

[0090] 显示屏505用于显示UI(User Interface,用户界面)。该UI可以包括图形、文本、图标、视频及其它们的任意组合。当显示屏505是触摸显示屏时,显示屏505还具有采集在显示屏505的表面或表面上方的触摸信号的能力。该触摸信号可以作为控制信号输入至处理器501进行处理。此时,显示屏505还可以用于提供虚拟按钮和/或虚拟键盘,也称软按钮和/或软键盘。在一些实施例中,显示屏505可以为一个,设置终端500的前面板;在另一些实施例中,显示屏505可以为至少两个,分别设置在终端500的不同表面或呈折叠设计;在再一些实施例中,显示屏505可以是柔性显示屏,设置在终端500的弯曲表面上或折叠面上。甚至,显示屏505还可以设置成非矩形的不规则图形,也即异形屏。显示屏505可以采用LCD(Liquid Crystal Display,液晶显示屏)、OLED(OrganicLight-Emitting Diode,有机发光二极管)等材质制备。

[0091] 摄像头组件506用于采集图像或视频。可选地,摄像头组件506包括前置摄像头和后置摄像头。通常,前置摄像头设置在终端的前面板,后置摄像头设置在终端的背面。在一些实施例中,后置摄像头为至少两个,分别为主摄像头、景深摄像头、广角摄像头、长焦摄像头中的任意一种,以实现主摄像头和景深摄像头融合实现背景虚化功能、主摄像头和广角摄像头融合实现全景拍摄以及VR(Virtual Reality,虚拟现实)拍摄功能或者其它融合拍摄功能。在一些实施例中,摄像头组件506还可以包括闪光灯。闪光灯可以是单色温闪光灯,也可以是双色温闪光灯。双色温闪光灯是指暖光闪光灯和冷光闪光灯的组合,可以用于不同色温下的光线补偿。

[0092] 音频电路507可以包括麦克风和扬声器。麦克风用于采集用户及环境的声波,并将声波转换为电信号输入至处理器501进行处理,或者输入至射频电路504以实现语音通信。出于立体声采集或降噪的目的,麦克风可以为多个,分别设置在终端500的不同部位。麦克风还可以是阵列麦克风或全向采集型麦克风。扬声器则用于将来自处理器501或射频电路504的电信号转换为声波。扬声器可以是传统的薄膜扬声器,也可以是压电陶瓷扬声器。当扬声器是压电陶瓷扬声器时,不仅可以将电信号转换为人类可听见的声波,也可以将电信号转换为人类听不见的声波以进行测距等用途。在一些实施例中,音频电路507还可以包括

耳机插孔。

[0093] 定位组件508用于定位终端500的当前地理位置，以实现导航或LBS(Location Based Service，基于位置的服务)。定位组件508可以是基于美国的GPS(Global Positioning System，全球定位系统)、中国的北斗系统、俄罗斯的格雷纳斯系统或欧盟的伽利略系统的定位组件。

[0094] 电源509用于为终端500中的各个组件进行供电。电源509可以是交流电、直流电、一次性电池或可充电电池。当电源509包括可充电电池时，该可充电电池可以支持有线充电或无线充电。该可充电电池还可以用于支持快充技术。

[0095] 在一些实施例中，终端500还包括有一个或多个传感器510。该一个或多个传感器510包括但不限于：加速度传感器511、陀螺仪传感器512、压力传感器513、指纹传感器514、光学传感器515以及接近传感器516。

[0096] 加速度传感器511可以检测以终端500建立的坐标系的三个坐标轴上的加速度大小。比如，加速度传感器511可以用于检测重力加速度在三个坐标轴上的分量。处理器501可以根据加速度传感器511采集的重力加速度信号，控制触摸显示屏505以横向视图或纵向视图进行用户界面的显示。加速度传感器511还可以用于游戏或者用户的运动数据的采集。

[0097] 陀螺仪传感器512可以检测终端500的机体方向及转动角度，陀螺仪传感器512可以与加速度传感器511协同采集用户对终端500的3D动作。处理器501根据陀螺仪传感器512采集的数据，可以实现如下功能：动作感应(比如根据用户的倾斜操作来改变UI)、拍摄时的图像稳定、游戏控制以及惯性导航。

[0098] 压力传感器513可以设置在终端500的侧边框和/或触摸显示屏505的下层。当压力传感器513设置在终端500的侧边框时，可以检测用户对终端500的握持信号，由处理器501根据压力传感器513采集的握持信号进行左右手识别或快捷操作。当压力传感器513设置在触摸显示屏505的下层时，由处理器501根据用户对触摸显示屏505的压力操作，实现对UI界面上的可操作性控件进行控制。可操作性控件包括按钮控件、滚动条控件、图标控件、菜单控件中的至少一种。

[0099] 指纹传感器514用于采集用户的指纹，由处理器501根据指纹传感器514采集到的指纹识别用户的身份，或者，由指纹传感器514根据采集到的指纹识别用户的身份。在识别出用户的身份为可信身份时，由处理器501授权该用户执行相关的敏感操作，该敏感操作包括解锁屏幕、查看加密信息、下载软件、支付及更改设置等。指纹传感器514可以被设置终端500的正面、背面或侧面。当终端500上设置有物理按键或厂商Logo时，指纹传感器514可以与物理按键或厂商标志集成在一起。

[0100] 光学传感器515用于采集环境光强度。在一个实施例中，处理器501可以根据光学传感器515采集的环境光强度，控制触摸显示屏505的显示亮度。具体地，当环境光强度较高时，调高触摸显示屏505的显示亮度；当环境光强度较低时，调低触摸显示屏505的显示亮度。在另一个实施例中，处理器501还可以根据光学传感器515采集的环境光强度，动态调整摄像头组件506的拍摄参数。

[0101] 接近传感器516，也称距离传感器，通常设置在终端500的前面板。接近传感器516用于采集用户与终端500的正面之间的距离。在一个实施例中，当接近传感器516检测到用户与终端500的正面之间的距离逐渐变小时，由处理器501控制触摸显示屏505从亮屏状态

切换为息屏状态；当接近传感器516检测到用户与终端500的正面之间的距离逐渐变大时，由处理器501控制触摸显示屏505从息屏状态切换为亮屏状态。

[0102] 本领域技术人员可以理解，图5中示出的结构并不构成对终端500的限定，可以包括比图示更多或更少的组件，或者组合某些组件，或者采用不同的组件布置。

[0103] 本发明实施例还提供了一种内存管理装置，该内存管理装置包括处理器和存储器，存储器中存储有至少一条指令、至少一段程序、代码集或指令集，指令、程序、代码集或指令集由处理器加载并执行以实现上述实施例的内存管理方法中所执行的操作。

[0104] 本发明实施例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中存储有至少一条指令、至少一段程序、代码集或指令集，该指令、该程序、该代码集或该指令集由处理器加载并执行以实现上述实施例的内存管理方法中所执行的操作。

[0105] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

[0106] 以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

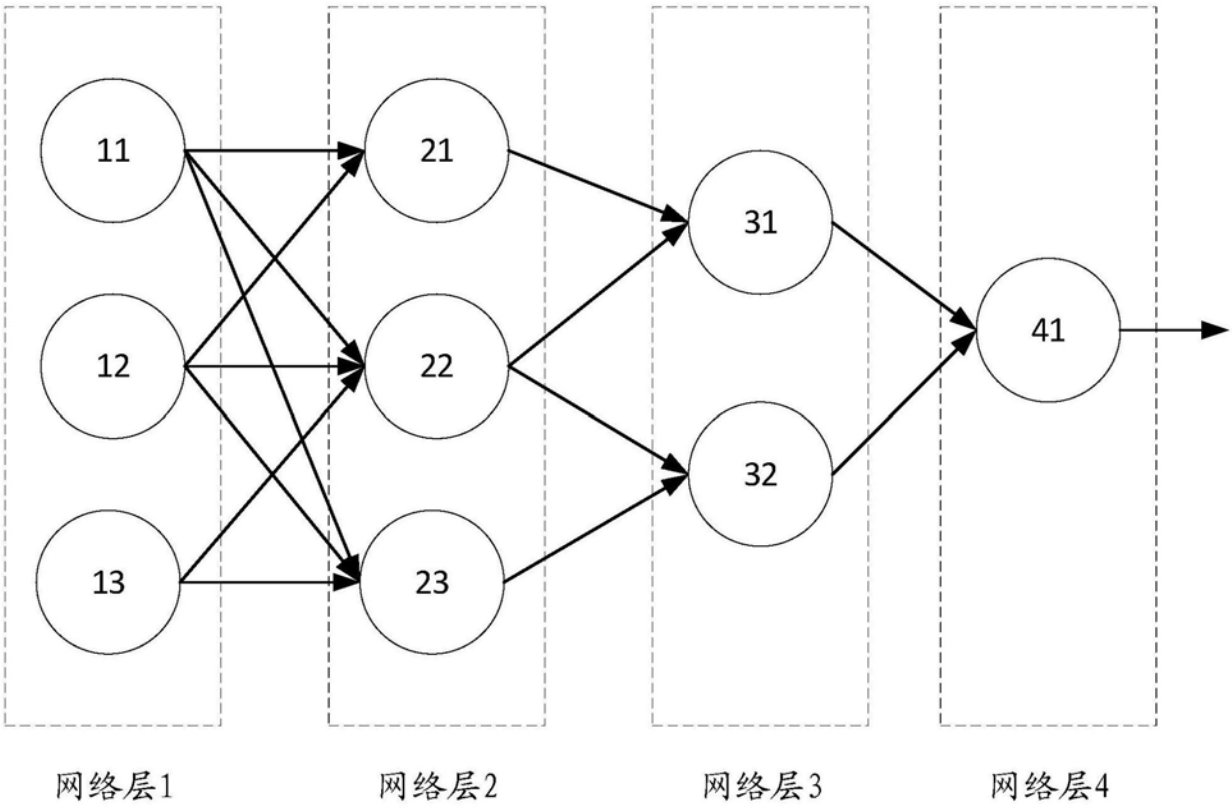

11
12
13
21
22
23
31
32
41
网络层1
网络层2
网络层3
网络层4

图1

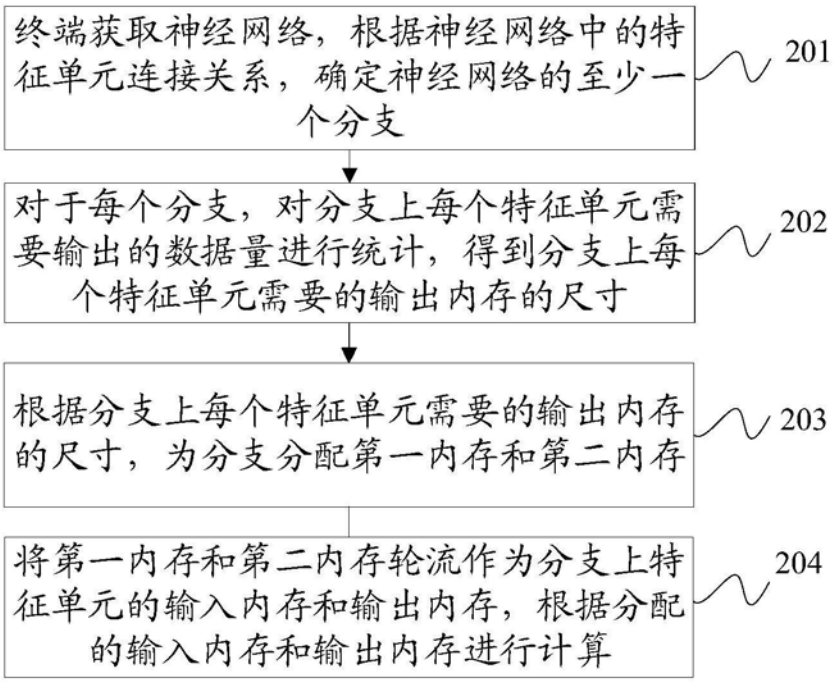

终端获取神经网络，根据神经网络中的特征单元连接关系，确定神经网络的至少一个分支
201
对于每个分支，对分支上每个特征单元需要输出的数据量进行统计，得到分支上每个特征单元需要的输出内存的尺寸
202
根据分支上每个特征单元需要的输出内存的尺寸，为分支分配第一内存和第二内存
203
将第一内存和第二内存轮流作为分支上特征单元的输入内存和输出内存，根据分配的输入内存和输出内存进行计算
204

图2

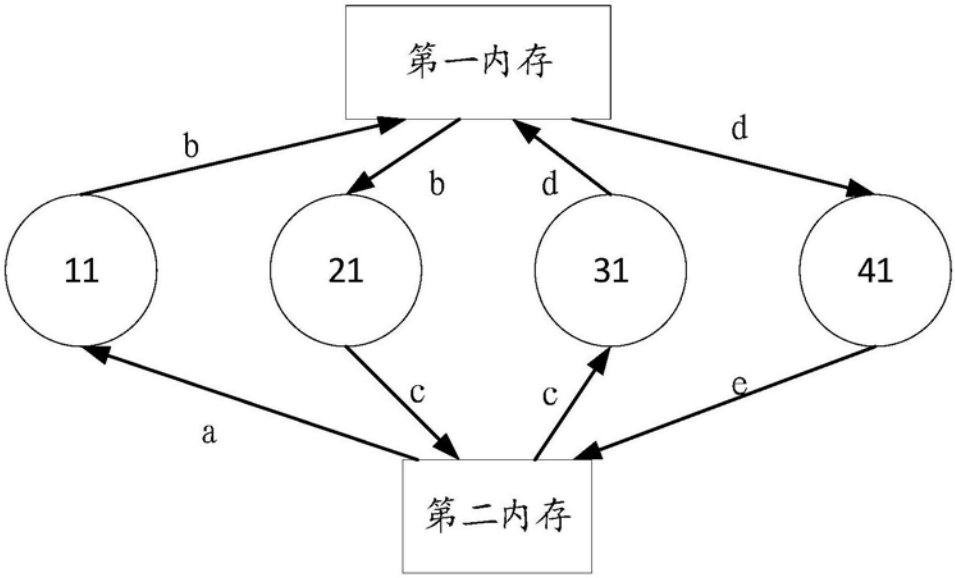


图3

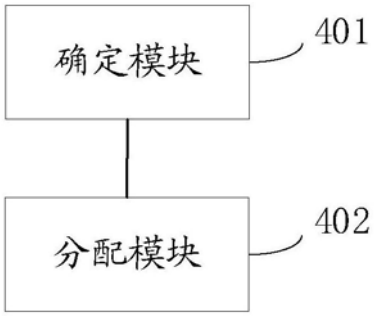


图4

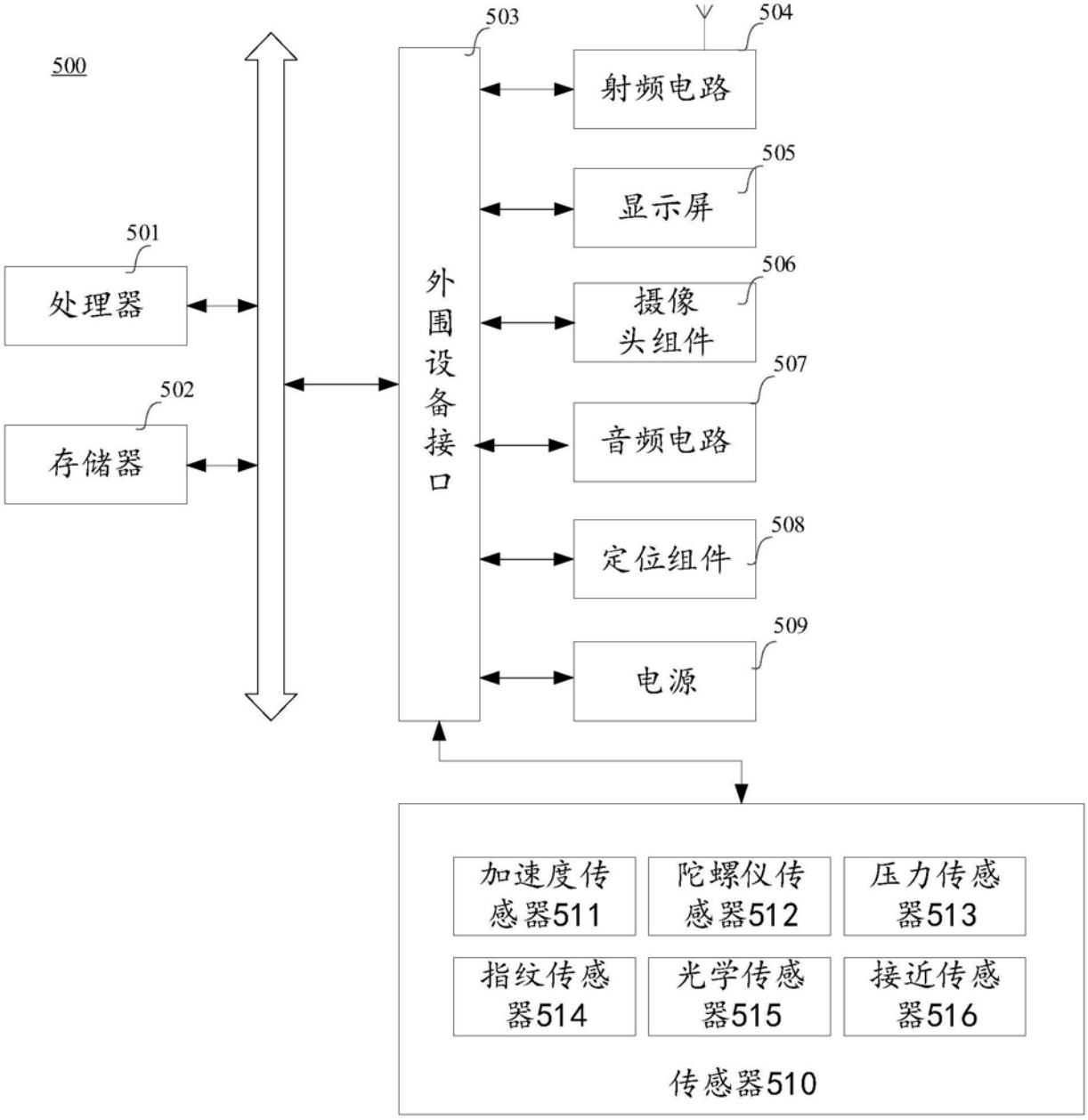

500
501
处理器
502
存储器
503
外围设备接口
504
射频电路
505
显示屏
506
摄像头组件
507
音频电路
508
定位组件
509
电源
加速度传感器511
陀螺仪传感器512
压力传感器513
指纹传感器514
光学传感器515
接近传感器516
传感器510

图5