(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111506522 A
(43)申请公布日 2020.08.07

(21)申请号 201910099854.2

(22)申请日 2019.01.31

(71)申请人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱

(72)发明人 张阳明 袁信 杨春 吴月敏

(74)专利代理机构 北京太合九思知识产权代理有限公司 11610
代理人 张爱

(51)Int.Cl.
G06F 12/0811(2016.01)
G06N 3/04(2006.01)

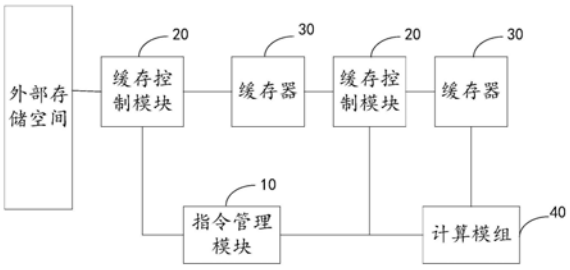
权利要求书4页 说明书11页 附图6页

(54)发明名称

数据处理设备及方法

(57)摘要

本申请实施例提供一种数据处理设备及方法。其中,数据处理设备包括多级缓存控制模块、多级缓存器以及计算模组,计算模组与多级缓存器中的末级缓存器连接。基于这种设备结构,多级缓存控制模块可控制数据在多级缓存器之间移动,计算模组可在末级缓存器提供的数据的支持下,进行数据处理操作。进而,实现了数据移动过程和数据处理过程并行的技术效果,有利于提升数据处理设备的数据处理效率。



CN 111506522 A

1.一种数据处理设备,其特征在于,包括:

指令管理模块、计算模组、多级缓存控制模块以及所述多级缓存控制模块对应的多级缓存器;

其中,所述指令管理模块,用于向所述多级缓存控制模块下发数据更新指令,以及向所述计算模组下发计算指令;

所述多级缓存控制模块,用于根据所述数据更新指令,获取激活数据和模型数据,并控制所述激活数据和所述模型数据在所述多级缓存器之间移动;

所述计算模组,与所述多级缓存器中的末级缓存器连接,用于根据所述计算指令,从所述末级缓存器中获取目标模型数据和目标激活数据,以进行数据处理操作。

2.根据权利要求1所述的设备,其特征在于,所述多级缓存控制模块,包括:多级模型数据缓存控制模块以及多级激活数据缓存控制模块;

所述多级缓存器包括:与所述多级模型数据缓存控制模块对应的多级模型数据缓存器,以及与所述多级激活数据缓存控制模块对应的多级激活数据缓存器。

3.根据权利要求2所述的设备,其特征在于,所述多级模型数据缓存控制模块,包括:一级模型数据缓存控制模块和二级模型数据缓存控制模块;

所述多级模型数据缓存器包括:一级模型数据缓存器和二级模型数据缓存器;

其中,所述一级模型数据缓存控制模块,用于:根据所述指令管理模块下发的第一数据移动指令,在所述一级模型数据缓存器处于可写状态时,从外部存储空间获取第一模型数据,并写入所述一级模型数据缓存器中;

所述二级模型数据缓存控制模块,用于:根据所述指令管理模块下发的第二数据移动指令,在所述一级模型数据缓存器处于可读状态,且所述二级模型数据缓存器处于可写状态时,从所述一级模型数据缓存器中获取第二模型数据,并写入所述二级模型数据缓存器中。

4.根据权利要求3所述的设备,其特征在于,所述多级激活数据缓存控制模块,包括:一级激活数据缓存控制模块和二级激活数据缓存控制模块;

所述多级激活数据缓存器包括:一级激活数据缓存器和二级激活数据缓存器;

其中,所述一级激活数据缓存控制模块,用于:根据所述指令管理模块下发的第三数据移动指令,在所述一级激活数据缓存器处于可写状态时,从外部存储空间获取第一激活数据,并写入所述一级激活数据缓存器中;

所述二级激活数据缓存控制模块,用于:根据所述指令管理模块下发的第四数据移动指令,在所述一级激活数据缓存器处于可读状态,且所述二级激活数据缓存器处于可写状态时,从所述一级激活数据缓存器中获取第二激活数据,并写入所述二级激活数据缓存器中。

5.根据权利要求4所述的设备,其特征在于,所述计算模组,包括:

向量矩阵计算模块、第一结果缓存器、多功能计算模块、第二结果缓存器、结果回写模块;

所述向量矩阵计算模块,用于根据所述指令管理模块下发的向量矩阵计算指令,从所述二级模型数据缓存器中获取所述目标模型数据,从所述二级激活数据缓存器中获取所述目标激活数据,以进行神经网络模型中的向量矩阵计算,并将所述向量矩阵计算的结果写

入所述第一结果缓存器；

所述多功能计算模块，用于根据所述指令管理模块下发的多功能计算指令，从所述第一结果缓存器中读取所述向量矩阵计算的结果中的第一结果，以根据所述第一结果进行所述神经网络模型中的多功能计算，并将所述多功能计算的结果写入所述第二结果缓存器；

所述结果回写模块，用于根据所述指令管理模块下发的结果回写指令，从所述第二结果缓存器读取所述多功能计算的结果中的第二结果，并将所述第二结果回写至所述多级激活数据缓存器，以更新所述多级激活数据缓存器中的激活数据，或者将所述第二结果写入所述外部存储空间。

6.根据权利要求5所述的设备，其特征在于，所述设备还包括：一端连接于所述一级模型数据缓存控制模块，另一端连接于所述多功能计算模块的偏置参数缓存器；

所述一级模型数据缓存控制模块，还用于：在所述偏置参数缓存器处于可写状态时，从所述外部存储空间获取偏置参数，并写入所述偏置参数缓存器中；

所述多功能计算模块，还用于：从所述偏置参数缓存器读取目标偏置参数，并根据所述目标偏置参数和所述第一结果进行所述神经网络模型中的多功能计算。

7.根据权利要求6所述的设备，其特征在于，所述计算模组还包括：与所述多功能计算模块连接的临时缓存模块；

所述临时缓存模块，用于存放所述多功能计算模块进行多功能计算时的临时计算结果；

所述多功能计算模块，还用于：从所述临时缓存模块中读取所述多功能计算的临时计算结果中的第三结果，并根据所述目标偏置参数、所述第一结果以及所述第三结果，进行所述神经网络模型中的多功能计算。

8.根据权利要求1-7任一项所述的设备，其特征在于，所述模型数据，包括：

所述神经网络模型各计算层所需的权重参数和/或偏置参数；

所述激活数据，包括：所述神经网络模型各计算层所需的待处理数据。

9.根据权利要求1-7任一项所述的设备，其特征在于，所述多级缓存器中，前一级缓存器的存储容量大于相邻的后一级缓存器，后一级缓存器的访问带宽大于相邻的前一级缓存器。

10.一种数据处理方法，其特征在于，包括：

指令管理模块响应于计算任务，向多级缓存控制模块下发数据更新指令以及向计算模组下发计算指令；

所述多级缓存控制模块根据所述数据更新指令，获取激活数据和模型数据，并控制所述激活数据和所述模型数据在与所述多级缓存控制模块对应的多级缓存器之间移动；

所述计算模组根据所述计算指令，从所述多级缓存器中的末级缓存器中获取目标模型数据和目标激活数据，以进行数据处理操作。

11.根据权利要求10所述的方法，其特征在于，所述多级缓存控制模块根据所述数据更新指令，获取激活数据和模型数据，并控制所述激活数据和所述模型数据在与所述多级缓存控制模块对应的多级缓存器之间移动，包括：

所述多级缓存控制模块中的多级模型数据缓存控制模块，获取模型数据，并控制所述模型数据在与所述多级模型数据缓存控制模块对应的多级模型数据缓存器之间移动；以

及，

所述多级缓存控制模块中的多级激活数据缓存控制模块，获取所述激活数据，并控制所述激活数据，在与所述多级激活数据缓存控制模块对应的多级激活数据缓存器之间移动。

12.根据权利要求11所述的方法，其特征在于，所述多级缓存控制模块中的多级模型数据缓存控制模块，获取模型数据，并控制所述模型数据在与所述多级模型数据缓存控制模块对应的多级模型数据缓存器之间移动，包括：

所述多级模型数据缓存控制模块中的一级模型数据缓存控制模块，根据所述指令管理模块下发的第一数据移动指令，在与所述一级模型数据缓存控制模块对应的一级模型数据缓存器处于可写状态时，从外部存储空间获取第一模型数据，并写入所述一级模型数据缓存器中；

所述多级模型数据缓存控制模块中的二级模型数据缓存控制模块，根据所述指令管理模块下发的第二数据移动指令，在所述一级模型数据缓存器处于可读状态，且与所述二级模型数据缓存控制模块对应的二级模型数据缓存器处于可写状态时，从所述一级模型数据缓存器中获取第二模型数据，并写入所述二级模型数据缓存器中。

13.根据权利要求11或12所述的方法，其特征在于，所述多级缓存控制模块中的多级激活数据缓存控制模块，获取所述激活数据，并控制所述激活数据，在与所述多级激活数据缓存控制模块对应的多级激活数据缓存器之间移动，包括：

所述多级激活数据缓存控制模块中的一级激活数据缓存控制模块，根据所述指令管理模块下发的第三数据移动指令，在与所述一级激活数据缓存控制模块对应的一级激活数据缓存器处于可写状态时，从外部存储空间获取第一激活数据，并写入所述一级激活数据缓存器中；

所述多级激活数据缓存控制模块中的二级激活数据缓存控制模块，根据所述指令管理模块下发的第四数据移动指令，在所述一级激活数据缓存器处于可读状态，且与所述二级激活数据缓存控制模块对应的二级激活数据缓存器处于可写状态时，从所述一级激活数据缓存器中获取第二激活数据，并写入所述二级激活数据缓存器中。

14.根据权利要求13所述的方法，其特征在于，计算模组根据所述计算指令，从所述多级缓存器中的末级缓存器中获取目标模型数据和目标激活数据，以进行数据处理操作，包括：

所述计算模组中的向量矩阵计算模块根据所述指令管理模块下发的向量矩阵计算指令，从所述二级模型数据缓存器中获取所述目标模型数据，从所述二级激活数据缓存器中获取所述目标激活数据，以进行神经网络模型中的向量矩阵计算，并将所述向量矩阵计算的结果写入所述计算模组中的第一结果缓存器；

所述计算模组中的多功能计算模块根据所述指令管理模块下发的多功能计算指令，从所述第一结果缓存器中读取所述向量矩阵计算的结果中的第一结果，以根据所述第一结果进行所述神经网络模型中的多功能计算，并将所述多功能计算的结果写入所述计算模组中的第二结果缓存器；

所述计算模组中的结果回写模块根据所述指令管理模块下发的结果回写指令，从所述第二结果缓存器读取所述多功能计算的结果中的第二结果，并将所述第二结果回写至所述

多级激活数据缓存器，以更新所述多级激活数据缓存器中的激活数据，或者将所述第二结果写入所述外部存储空间。

15.根据权利要求14所述的方法，其特征在于，还包括：

所述一级模型数据缓存控制模块在所述多级缓存器中的偏置参数缓存器处于可写状态时，从所述外部存储空间获取偏置参数，并写入所述偏置参数缓存器中；

所述多功能计算模块从所述第一结果缓存器中读取所述向量矩阵计算的结果中的第一结果，以根据所述第一结果进行所述神经网络模型中的多功能计算，包括：

所述多功能计算模块从所述偏置参数缓存器读取目标偏置参数，并根据所述目标偏置参数和所述第一结果，进行所述神经网络模型中的多功能计算。

16.根据权利要求15所述的方法，其特征在于，所述多功能计算模块，进行所述神经网络模型中的多功能计算，还包括：

所述多功能计算模块，将所述多功能计算的临时计算结果写入所述计算模组中的临时缓存模块中；

所述多功能计算模块从所述偏置参数缓存器读取目标偏置参数，并根据所述目标偏置参数和所述第一结果，进行所述神经网络模型中的多功能计算，包括：

所述多功能计算模块从所述临时缓存模块中，读取所述多功能计算的临时计算结果中第三结果；

所述多功能计算模块根据所述目标偏置参数、所述第一结果以所述第三结果，进行所述神经网络模型中的多功能计算。

数据处理设备及方法

技术领域

[0001] 本申请涉及机器学习领域,尤其涉及一种数据处理设备及方法。

背景技术

[0002] 随着人工智能技术的发展,机器学习算法的复杂度日益提升。传统的计算芯片,例如CPU(Central Processing Unit,中央处理模块)、GPU(Graphics Processing Unit,图形处理器)等,已经无法满足迅速增长的计算力需求。因此,有待提出一种新的解决方案。

发明内容

[0003] 本申请的多个方面提供一种数据处理设备及方法,用以提升机器学习算法的计算效率。

[0004] 本申请实施例提供一种数据处理设备,包括:指令管理模块、计算模组、多级缓存控制模块以及所述多级缓存控制模块对应的多级缓存器;其中,所述指令管理模块,用于向所述多级缓存控制模块下发数据更新指令,以及向所述计算模组下发计算指令;所述多级缓存控制模块,用于根据所述数据更新指令,获取激活数据和模型数据,并控制所述激活数据和所述模型数据在所述多级缓存器之间移动;所述计算模组,与所述多级缓存器中的末级缓存器连接,用于根据所述计算指令,从所述末级缓存器中获取目标模型数据和目标激活数据,以进行数据处理操作。

[0005] 本申请实施例还提供一种数据处理方法,包括:指令管理模块响应于计算任务,向多级缓存控制模块下发数据更新指令以及向计算模组下发计算指令;所述多级缓存控制模块根据所述数据更新指令,获取激活数据和模型数据,并控制所述激活数据和所述模型数据在与所述多级缓存控制模块对应的多级缓存器之间移动;所述计算模组根据所述计算指令,从所述多级缓存器中的末级缓存器中获取目标模型数据和目标激活数据,以进行数据处理操作。

[0006] 本申请实施例提供的数据处理设备及方法中,数据处理设备包括多级缓存控制模块、多级缓存器以及计算模组,计算模组与多级缓存器中的末级缓存器连接。基于这种设备结构,多级缓存控制模块可控制数据在多级缓存器之间移动,计算模组可在末级缓存器提供的数据的支持下,进行数据处理操作。进而,实现了数据移动过程和数据处理过程并行的技术效果,有利于提升数据处理设备的数据处理效率。

附图说明

[0007] 此处所说明的附图用来提供对本申请的进一步理解,构成本申请的一部分,本申请的示意性实施例及其说明用于解释本申请,并不构成对本申请的不当限定。在附图中:

[0008] 图1为本申请一示例性实施例提供的数据处理设备的结构示意图;

[0009] 图2a为本申请另一示例性实施例提供的数据处理设备的结构示意图;

[0010] 图2b为本申请又一示例性实施例提供的数据处理设备的结构示意图;

[0011] 图2c为本申请一示例性实施例提供的神经网络模型的计算过程抽象示意图；

[0012] 图2d为本申请另一示例性实施例提供的神经网络模型的计算过程抽象示意图；

[0013] 图2e为本申请又一示例性实施例提供的数据处理设备的结构示意图；

[0014] 图2f为本申请又一示例性实施例提供的数据处理设备的结构示意图；

[0015] 图2g为本申请又一示例性实施例提供的数据处理设备的结构示意图；

[0016] 图3为本申请一示例性实施例提供的数据处理方法的流程示意图。

具体实施方式

[0017] 为使本申请的目的、技术方案和优点更加清楚，下面将结合本申请具体实施例及相应的附图对本申请技术方案进行清楚、完整地描述。显然，所描述的实施例仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，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，都属于本申请保护的范围。

[0018] 针对现有技术中，传统的计算芯片无法满足迅速增长的计算力需求的技术问题，在本申请一些实施例中，提供了一种解决方案，以下结合附图，详细说明本申请各实施例提供的技术方案。

[0019] 图1为本申请一示例性实施例提供的数据处理设备的结构示意图，如图1所示，该数据处理设备100包括：指令管理模块10、多级缓存控制模块20、多级缓存器30以及计算模组40。

[0020] 其中，指令管理模块10，与多级缓存控制模块20以及计算模组40连接，用于向多级缓存控制模块20下发数据更新指令，以及向计算模组40下发计算指令。在一些实施例中，指令管理模块10可在接收到计算任务时，下发上述指令；在另一些实施例中，指令管理模块10可在设定的定时事件触发下，下发上述指令。

[0021] 在数据处理设备100中，多级缓存控制模块20，与指令管理模块10以及多级缓存器30连接，用于根据指令管理模块10下发的数据更新指令，获取激活数据和模型数据，并控制激活数据和模型数据在多级缓存器30之间移动。

[0022] 其中，模型数据，指的是数据处理设备100进行数据处理时所需的任何计算参数、配置参数和/或指令。例如，数据处理设备100进行加权计算时，模型数据可包括加权计算所需的权重。当然，数据处理设备100被应用于其他数据处理场景时，模型数据还可实现为其他数据处理过程所需的各类参数或者指令，本实施例不做限制。

[0023] 在数据处理设备100中，多级缓存控制模块20与多级缓存器30可一一对应，每级缓存控制模块可控制与其对应的缓存器中缓存的数据是否进行更新。例如，首级(一级)缓存控制模块与首级(一级)缓存器对应，且可控制首级(一级)缓存器中缓存的数据是否进行更新。二级缓存控制模块与二级缓存器对应，且可控制二级缓存器中缓存的数据是否进行更新。

[0024] 在本实例中，多级缓存器30可包含两级缓存器、三级缓存器甚至更多级缓存器。缓存器30的级数和缓存控制模块20的级数相同，多级缓存器30之间，可通过多级缓存控制模块20串行连接。通常，首级(一级)缓存器通过首级(一级)缓存控制模块连接于外部存储空间，二级缓存器通过二级缓存控制模块连接于一级缓存器，以此类推。其中，末级缓存器一方面通过末级缓存控制模块连接于相邻的前一级缓存器，另一方面连接于计算模组40，以

支持计算模组40的计算过程。

[0025] 在数据处理设备100中，计算模组40与多级缓存器30中的末级缓存器连接，用于根据指令管理模块10下发的计算指令，从该末级缓存器中获取目标模型数据和目标激活数据，以进行数据处理操作。

[0026] 其中，计算模组40可包括一个或多个计算模块，每个计算模块可对应一个缓存器，以缓存该计算模块的计算结果。

[0027] 其中，目标模型数据，指的是模型数据中，实时参与计算模组40的数据处理过程的部分数据；目标激活数据，指的激活数据中，实时参与计算模组40的数据处理过程的部分数据。

[0028] 例如，计算模组40执行神经网络模型中的第一计算层对应的数据处理操作时，目标激活数据可以为输入第一计算层的样本数据，目标模型数据可以为神经网络模型第一计算层的各种模型参数。再例如，计算模组40执行神经网络模型中的第二计算层对应的数据处理操作时，目标激活数据可以为第一计算层的计算结果，目标模型数据可以为神经网络模型第二计算层的各种模型参数。

[0029] 本实施例中，数据处理设备包括多级缓存控制模块、多级缓存器以及计算模组，计算模组与多级缓存器中的末级缓存器连接。基于这种设备结构，多级缓存控制模块可控制数据在多级缓存器之间移动，计算模组可在末级缓存器提供的数据的支持下，进行数据处理操作。进而，实现了数据移动过程和数据处理过程并行的技术效果，有利于提升数据处理设备的数据处理效率。

[0030] 需要说明的是，在本申请的上述或者下述实施例中，数据处理设备100可采用专门应用的集成电路(Application Specific Integrated Circuit,ASIC)芯片实现；或者采用eASIC芯片实现；或者采用可编程逻辑器件实现，例如现场可编程门阵列(Field-Programmable Gate Array,FPGA)是在可编程阵列逻辑器件(Programmable Array Logic,PAL)、通用阵列逻辑器件(General Array Logic,GAL)、复杂可编程逻辑器件(Complex Programmable Logic Device,CPLD)等，本申请实施例包含但不限于此。

[0031] 在一些示例性的实施例中，多级缓存器30中，前一级缓存器的存储容量大于相邻的后一级缓存器，后一级缓存器的访问带宽大于相邻的前一级缓存器。也就是说，在多级缓存器30中，首级(一级)缓存器通常具有最大的存储容量，末级缓存器通常具有最大的访问带宽。基于这种实施方式，首级(一级)缓存控制模块可从外部存储空间中获取足够多的激活数据以及模型数据存放在首级缓存器中，以提高外部存储空间的访问带宽的使用效率。末级缓存器具有最大的访问带宽，可满足计算模组40的数据输入带宽需求。

[0032] 在一些示例性的实施方式中，如图2a所示，多级缓存控制模块20可包括：多级模型数据缓存控制模块20A以及多级激活数据缓存控制模块20B；多级缓存器30包括：与多级模型数据缓存控制模块20A对应的多级模型数据缓存器30A，以及与多级激活数据缓存控制模块20B对应的多级激活数据缓存器30B。

[0033] 基于此，多级模型数据缓存控制模块20A以及多级模型数据缓存器30A可组成一路数据缓存通道，用于获取并移动模型数据；多级激活数据缓存控制模块20B和多级激活数据缓存器30B可组成另一路数据缓存通道，用于获取并移动激活数据。两条数据缓存通道并行，一方面有利于对参加数据处理的数据进行区分，另一方面，有利于提升数据传输的速

率。

[0034] 可选地，在一些实施例中，数据处理设备100可应用于实现神经网络模型的相关计算。在这种实施方式中，模型数据包括但不限于：神经网络模型各计算层所需的权重参数和/或偏置参数。除了神经网络模型各计算层所需的权重参数和/或偏置参数之外，模型数据还可包括参与神经网络模型计算的其他各类数据，此处不赘述。激活数据，包括但不限于：神经网络模型各计算层所需的待处理数据，例如输入样本、前一层的计算结果等。

[0035] 可选地，在本实施例中，模型数据缓存控制模块20A和模型数据缓存器30A的级数，可以为两级、三级或者更多级，本实施例不做限制。考虑到降低成本的需求，在一些实施例中，可在数据处理设备100上设置两级模型数据缓存控制模块20A和两级模型数据缓存器30A。

[0036] 如图2b所示，在这种实施方式中，多级模型数据缓存控制模块20A包括：一级模型数据缓存控制模块201A和二级模型数据缓存控制模块202A。多级模型数据缓存器30A包括：一级模型数据缓存器301A和二级模型数据缓存器302A。

[0037] 参考图2b所示，一级模型数据缓存控制模块201A的一端可与外部存储空间(或者外部处理器)连接，另一端与一级模型数据缓存器301A连接。

[0038] 在一些实施例中，指令管理模块10可向一级模型数据缓存控制模块201A下发第一数据移动指令。一级模型数据缓存控制模块201A，根据该第一数据移动指令，可在一级模型数据缓存器301A处于可写状态时，从外部存储空间获取第一模型数据，并写入一级模型数据缓存器301A中。其中，第一模型数据，指的是模型数据中，由外部存储空间(或者外部处理器)中移动至一级模型数据缓存器301A中的部分。

[0039] 参考图2b所示，二级模型数据缓存控制模块202A的一端与一级模型数据缓存器301A连接，另一端与二级模型数据缓存器302A连接。在一些实施例中，指令管理模块10可向二级模型数据缓存控制模块202A下发第二数据移动指令。二级模型数据缓存控制模块202A，根据该第二数据移动指令，可在一级模型数据缓存器301A处于可读状态，且二级模型数据缓存器302A处于可写状态时，从一级模型数据缓存器301A中获取第二模型数据，并写入二级模型数据缓存器302A中。其中，第二模型数据，指的是模型数据中，由一级模型数据缓存器301A移动至二级模型数据缓存器302A中的部分。

[0040] 在图2a以及2b中，为清晰示意其他模块之间的连接关系，涉及到指令管理模块10的具体连接方式并未在图2a以及2b中进行示意，但并不意味着指令管理模块10与其他模块之间无连接。可选地，指令管理模块10与可接收其指令的各模块存在连接关系，此处不再赘述。

[0041] 需要说明的是，在上述实施例中，数据在一级模型数据缓存器301A和二级模型数据缓存器302A之间移动时，可采用pingpong缓存方式实现。在这种基于pingpong缓存的数据移动方式中，数据在一级模型数据缓存器301A和二级模型数据缓存器302A中，可分两路(ping路和pong路)进行读写操作，以下将具体说明。

[0042] 假设，一级模型数据缓存器301A中的ping路缓存空间处于可读状态，pong路缓存空间处于可写状态；二级模型数据缓存器302A中的ping路缓存空间处于可写状态，pong路缓存空间处于可读状态。

[0043] 此时，一级模型数据缓存控制模块201A可从外部存储空间获取第一模型数据，并

写入一级模型数据缓存器301A中的pong路缓存空间中。与此同时，二级模型数据缓存控制模块202A可从一级模型数据缓存器301A中的ping路缓存空间中获取第二模型数据，并写入二级模型数据缓存器302A的ping路缓存空间中。

[0044] 当一级模型数据缓存器301A中的pong路缓存空间中缓存足够的数据，且其状态为可读状态时，一级模型数据缓存器301A中的ping路缓存空间的数据已被读取完毕，且其状态为可写状态，此时，一级模型数据缓存控制模块201A可从外部存储空间获取第一模型数据，并写入一级模型数据缓存器301A中的ping路缓存空间中。二级模型数据缓存控制模块202A可从一级模型数据缓存器301A中的pong路缓存空间中获取第二模型数据，并写入二级模型数据缓存器302A的pong路缓存空间中。

[0045] 基于上述实施方式，一级模型数据缓存控制模块201A无需等待一级模型数据缓存器301A中所有的数据均被读取完毕再从外部存储空间中读取第一数据，二级模型数据缓存控制模块202A也无需等待二级模型数据缓存器302A中所有的数据均被读取完毕再从一级模型数据缓存器301A中读取第二数据。这种减少数据移动过程中花费的等待时间的方式，进一步提升了对后续的数据处理效率。

[0046] 如图2b所示，在一些可选的实施方式中，多级激活数据缓存控制模块20B包括：一级激活数据缓存控制模块201B和二级激活数据缓存控制模块202B。多级激活数据缓存器30B包括：一级激活数据缓存器301B和二级激活数据缓存器302B。

[0047] 参考图2b所示，一级激活数据缓存控制模块201B的一端与外部存储空间连接，另一端与一级激活数据缓存器301B连接。在一些实施例中，指令管理模块10可向一级激活数据缓存控制模块201B下发第三数据移动指令。一级激活数据缓存控制模块201B，根据该第三数据移动指令，可在一级激活数据缓存器301B处于可写状态时，从外部存储空间获取第一激活数据，并写入一级激活数据缓存器301B中。其中，第一激活数据，指的是激活数据中，由外部存储空间中移动至一级激活数据缓存器301B中的部分。

[0048] 参考图2b所示，二级激活数据缓存控制模块202B的一端与一级激活数据缓存器301B连接，另一端与二级激活数据缓存器302B连接。在一些实施例中，指令管理模块10可向二级激活数据缓存控制模块202B下发第四数据移动指令。二级激活数据缓存控制模块202B，根据该第四数据移动指令，可在一级激活数据缓存器301B处于可读状态，且二级激活数据缓存器302B处于可写状态时，从一级激活数据缓存器301B中获取第二激活数据，并写入二级激活数据缓存器302B中。其中，第二激活数据，指的是激活数据中，由一级激活数据缓存器301B移动至二级激活数据缓存器302B中的部分。

[0049] 在上述实施例中，数据在一级激活数据缓存器301B和二级激活数据缓存器302B之间移动时，一级激活数据缓存器301B和二级激活数据缓存器302B也可采用上述实施例记载的pingpong缓存的方式实现两路缓存通道并行，此处不再赘述。

[0050] 在本实施例中，计算模组40主要用于执行神经网络模型的相关计算操作。其中，神经网络模型可包括：MLP(Multi-Layer Perceptron，多层感知机)、CNN(Convolutional Neural Networks，卷积神经网络)、RNN(Recurrent Neural Network，循环神经网络)、LSTM(Long Short-Term Memory，长短期记忆网络)等，本申请实施例包含但不限于此。

[0051] 需要说明的是，当本申请实施例提供的数据处理设备100用于执行CNN模型相关的计算操作时，模型数据还包括CNN模型参数中的特征图(feaure map)。在这种应用场景下，

一级模型数据缓存控制模块201A在将外部存储空间中的特征图移动到一级模型数据缓存器301A中时，可执行按照卷积核将特征图展开成矩阵的运算(image2col)，以便于后续的矩阵向量计算；或者，二级模型数据缓存控制模块202A，在将一级模型数据缓存器301A中的特征图移动到二级模型数据缓存器302A中时，可执行按照卷积核将特征图展开成矩阵的运算(image2col)，以便于后续的矩阵向量计算，不再赘述。

[0052] 在对神经网络模型进行研究的过程中发现，神经网络模型的计算过程，可被抽象为两部分计算：矩阵向量计算和后处理多样运算。

[0053] 以MLP模型为例，如图2c所示，MLP模型的计算过程包括矩阵的乘运算(MatMul)以及由添加偏置参数运算(BiasAdd)和激活运算(Activation)组成的后处理多样运算。其中，激活运算包括但不限于：与ReLU(Rectified Linear Unit，线性整流函数)、Sigmoid、Tanh等激活函数相关的运算。

[0054] 以LSTM模型为例，如图2d所示，LSTM模型的计算过程包括矩阵的乘运算(MatMul)以及由偏置参数运算(BiasAdd)和激活运算(Activation)、添加元素运算(ElementAdd)和元素相乘运算(ElementMult)组成的后处理多样运算。其中，激活运算包括但不限于：sigmoid、tanh等激活函数相关的运算。

[0055] 以CNN模型为例，CNN模型中的后处理多样运算包括但不限于：与激活函数相关的运算、池化运算(Pooling)和规范化(Norm)运算。

[0056] 基于上述，在一些示例性的实施方式中，如图2b所示，计算模组40，可包括：向量矩阵计算模块401、第一结果缓存器402、多功能计算模块403、第二结果缓存器404、结果回写模块405。

[0057] 在一些实施例中，指令管理模块10可向向量矩阵计算模块401下发向量矩阵计算指令。向量矩阵计算模块401，用于根据该向量矩阵计算指令，从二级模型数据缓存器302A中获取目标模型数据，从二级激活数据缓存器302B中获取目标激活数据，以进行神经网络模型中的向量矩阵计算，并将向量矩阵计算的结果写入第一结果缓存器402。

[0058] 在一些实施例中，指令管理模块10可向多功能计算模块403下发多功能计算指令。多功能计算模块403，用于从第一结果缓存器402中读取该向量矩阵计算的结果中的第一结果，以根据该第一结果进行神经网络模型中的多功能计算，并将多功能计算的结果写入第二结果缓存器403。其中，第一结果，指的是第一结果缓存器402中保存的向量矩阵计算的结果中，用于参加多功能计算的部分或全部向量矩阵计算的结果。

[0059] 在一些实施例中，指令管理模块10可向结果回写模块405下发结果回写指令。

[0060] 在一典型的情况下中，写入第二结果缓存器403中的多功能计算的结果，为神经网络模型中前一计算层的输出结果，该输出结果可作为相邻的后一计算层的激活数据。

[0061] 在这种情况下，结果回写模块405，可根据该结果回写指令，从第二结果缓存器读取多功能计算的结果中的第二结果，并将该第二结果回写至多级激活数据缓存器30B，以更新多级激活数据缓存器30B中的激活数据。或者，将该第二结果回写至外部存储空间，以更新外部存储空间中保存的激活数据。基于此，前一计算层的输出结果，可通过多级激活数据缓存器30B作为相邻的后一计算层的激活数据。

[0062] 可选地，如图2b所示，本实施例中，结果回写模块405与一级激活数据缓存器301B、二级激活数据缓存器302B以及外部存储空间连接，可将第二结果回写至一级激活数据缓存

器301B、二级激活数据缓存器302B或者外部存储空间中，本实施例不做限制。

[0063] 需要说明的是，在上述各实施例中，模型数据包含的神经网络模型各计算层所需的权重参数以及偏置参数，权重参数主要用于支持向量矩阵计算模块401执行的向量矩阵计算，偏置参数主要用于支持多功能计算模块403执行的BiasAdd运算。

[0064] 在一些实施例中，可通过多级模型数据缓存控制模块20A及其对应的多级模型数据缓存器30A对偏置参数进行搬移，再经由向量矩阵计算模块401到达第一结果缓存器402，进而多功能计算模块403可从第一结果缓存器402中获取到计算所需的偏置参数。

[0065] 在另一些实施例中，可在多级缓存器30中设置偏置参数缓存器303。如图2e所示。偏置参数缓存器303的一端连接于一级模型数据缓存控制模块201A，另一端连接于多功能计算模块403。

[0066] 在这种实施方式中，一级模型数据缓存控制模块201A，可在偏置参数缓存器303处于可写状态时，从外部存储空间获取偏置参数，并写入偏置参数缓存器303中。

[0067] 基于此，多功能计算模块403，可从偏置参数缓存器303读取目标偏置参数，并根据目标偏置参数和第一结果，进行神经网络模型中的多功能计算。

[0068] 在这种实施方式中，用于支持多功能计算模块403执行的BiasAdd运算的偏置参数，可在一级模型数据缓存控制模块201A的控制下，由外部存储空间移动至偏置参数缓存器303，不再经过上述多级模型数据缓存器30A、向量矩阵计算模块401以及第一结果缓存器402，进而，极大节省了上述各缓存器的访问带宽以及存储容量。

[0069] 需要说明的是，在一些其他的实施例中，偏置参数可经过多级激活数据缓存控制模块20B和对应的多级激活数据缓存器30B进行搬移。如图2f所示，二级激活数据缓存器302B可与多功能计算模块403连接。在这种实施方式中，可通过一级激活数据缓存控制模块201B将偏置参数从外部存储器搬移到一级激活数据缓存器301B，再通过二级激活数据缓存控制模块202B将偏置参数从一级激活数据缓存器301B搬移到二级激活数据缓存器302B。进而，多功能计算模块403可直接从二级激活数据缓存器302B中获取到计算所需的偏置参数，不再经过上述向量矩阵计算模块401以及第一结果缓存器402，节省了访问带宽以及存储空间。

[0070] 需要说明的是，在一些其他的实施例中，如图2g所示，二级激活数据缓存控制模块202B可直接与第一结果缓存器402连接。进而，二级激活数据缓存控制模块202B可将偏置参数从一级激活数据缓存器301B搬移到第一结果缓存器402。多功能计算模块403可直接从第一结果缓存器402中获取到计算所需的偏置参数，不再经过二级激活数据缓存器302B以及向量矩阵计算模块401。

[0071] 需要说明的是，在一些实施例中，如图2e、2f、2g所示，计算模组40还包括：与多功能计算模块403连接的临时缓存模块406。临时缓存模块406，用于存放多功能计算模块403进行多功能计算时的临时计算结果。

[0072] 基于此，在一些实施例中，多功能计算模块403在进行神经网络模型中的多功能计算时，还可从临时缓存模块406中读取该多功能计算的临时计算结果中的第三结果，并根据获取到的目标偏置参数、该第一结果以及该第三结果，进行神经网络模型中的多功能计算。其中，第三结果指的是，临时缓存模块406中保存的多功能计算的临时计算结果中，用于继续参加多功能计算的部分或全部临时计算结果。

[0073] 例如，以LSTM模型为例，多功能计算模块403可将BiasAdd运算的结果写入临时缓存模块406。接着，多功能计算模块403可在进行ElementAdd运算之后，从临时缓存模块406中读取BiasAdd运算的结果，再基于ElementAdd运算和BiasAdd运算的结果进行激活函数运算。

[0074] 在上述各实施例中，由指令管理模块10管理以及下发的指令包括：第一数据移动指令、第二数据移动指令、第三数据移动指令、第四数据移动指令、向量矩阵计算指令、多功能计算指令、结果回写指令。

[0075] 在一些实施例中，第一数据移动指令，可实现为：LoadL2WeightBuffer(LL2WB)指令，该指令可由一级模型数据缓存控制模块201A执行，可实现将权重参数从外部存储空间移动到一级模型数据缓存器301A。

[0076] 第二数据移动指令，可实现为：LoadL1WeightBuffer(LL1WB)指令，该指令可由二级模型数据缓存控制模块202A执行，可实现将权重参数从一级模型数据缓存器301A移动到二级模型数据缓存器302A中。

[0077] 第三数据移动指令，可实现为：LoadL2ActivationBuffer(LL2AB)指令，该指令可由一级激活数据缓存控制模块201B执行，可实现将激活函数从外部存储空间移动到一级激活数据缓存器301B。

[0078] 第四数据移动指令，可实现为：LoadL1ActivationBuffer(LL1AB)指令，该指令可由二级激活数据缓存控制模块202B执行，可实现将激活函数从一级模型数据缓存器301B移动到二级模型数据缓存器302B中。

[0079] 向量矩阵计算指令可实现为：MartixVectorComputing(MVP)指令，该指令由向量矩阵计算模块401执行，可实现向量矩阵的乘法和加法运算，以累加矩阵相乘的结果。

[0080] 多功能计算指令可实现为MiscellaneousFunctionsComputing(MFC)，该指令由向多功能计算模块403执行，可完成各种激活函数运算、BiasAdd运算、element-add和element-mult运算等。

[0081] 结果回写指令可实现为UploadResult(UR)指令，该指令可由结果回写模块405执行，以将多功能计算的结果回写至多级激活数据缓存器30B或者外部存储空间。

[0082] 上述指令形成的指令集，可保存在外部存储空间、外部处理器或者指令管理模块10处，以供调用，不再赘述。

[0083] 本申请实施例除了提供上述数据处理设备之外，还提供一种数据处理方法，以下将结合附图进行说明。

[0084] 图3是本发明一示例性实施例提供的数据处理方法的流程示意图，如图3所示，该方法包括：

[0085] 步骤301、指令管理模块响应于计算任务，向多级缓存控制模块下发数据更新指令以及向计算模组下发计算指令。

[0086] 步骤302、多级缓存控制模块根据数据更新指令，获取激活数据和模型数据，并控制激活数据和模型数据在与多级缓存控制模块对应的多级缓存器之间移动。

[0087] 步骤303、计算模组根据计算指令，从多级缓存器中的末级缓存器中获取目标模型数据和目标激活数据，以进行数据处理操作。

[0088] 在一些示例性的实施方式中，多级缓存控制模块根据数据更新指令，获取激活数

据和模型数据，并控制激活数据和模型数据在与多级缓存控制模块对应的多级缓存器之间移动的一种方式，包括：多级缓存控制模块中的多级模型数据缓存控制模块，获取模型数据，并控制模型数据在与多级模型数据缓存控制模块对应的多级模型数据缓存器之间移动；以及，多级缓存控制模块中的多级激活数据缓存控制模块，获取激活数据，并控制激活数据，在与多级激活数据缓存控制模块对应的多级激活数据缓存器之间移动。

[0089] 在一些示例性的实施方式中，多级缓存控制模块中的多级模型数据缓存控制模块，获取模型数据，并控制模型数据在与多级模型数据缓存控制模块对应的多级模型数据缓存器之间移动的一种方式，包括：多级模型数据缓存控制模块中的一级模型数据缓存控制模块，根据指令管理模块下发的第一数据移动指令，在与一级模型数据缓存控制模块对应的一级模型数据缓存器处于可写状态时，从外部存储空间获取第一模型数据，并写入一级模型数据缓存器中；多级模型数据缓存控制模块中的二级模型数据缓存控制模块，根据指令管理模块下发的第二数据移动指令，在一级模型数据缓存器处于可读状态，且与二级模型数据缓存控制模块对应的二级模型数据缓存器处于可写状态时，从一级模型数据缓存器中获取第二模型数据，并写入二级模型数据缓存器中。

[0090] 在一些示例性的实施方式中，多级缓存控制模块中的多级激活数据缓存控制模块，获取激活数据，并控制激活数据，在与多级激活数据缓存控制模块对应的多级激活数据缓存器之间移动的一种方式，包括：多级激活数据缓存控制模块中的一级激活数据缓存控制模块，根据指令管理模块下发的第三数据移动指令，在与一级激活数据缓存控制模块对应的一级激活数据缓存器处于可写状态时，从外部存储空间获取第一激活数据，并写入一级激活数据缓存器中；多级激活数据缓存控制模块中的二级激活数据缓存控制模块，根据指令管理模块下发的第四数据移动指令，在一级激活数据缓存器处于可读状态，且与二级激活数据缓存控制模块对应的二级激活数据缓存器处于可写状态时，从一级激活数据缓存器中获取第二激活数据，并写入二级激活数据缓存器中。

[0091] 在一些示例性的实施方式中，计算模组根据计算指令，从多级缓存器中的末级缓存器中获取目标模型数据和目标激活数据，以进行数据处理操作的一种方式，包括：计算模组中的向量矩阵计算模块根据指令管理模块下发的向量矩阵计算指令，从二级模型数据缓存器中获取目标模型数据，从二级激活数据缓存器中获取目标激活数据，以进行神经网络模型中的向量矩阵计算，并将向量矩阵计算的结果写入计算模组中的第一结果缓存器；计算模组中的多功能计算模块根据指令管理模块下发的多功能计算指令，从第一结果缓存器中读取向量矩阵计算的结果中的第一结果，以根据第一结果进行神经网络模型中的多功能计算，并将多功能计算的结果写入计算模组中的第二结果缓存器；计算模组中的结果回写模块根据指令管理模块下发的结果回写指令，从第二结果缓存器读取多功能计算的结果中的第二结果，并将第二结果回写至多级激活数据缓存器，以更新多级激活数据缓存器中的激活数据，或者将第二结果写入外部存储空间。

[0092] 在一些示例性的实施方式中，还包括：一级模型数据缓存控制模块在多级缓存器中的偏置参数缓存器处于可写状态时，从外部存储空间获取偏置参数，并写入偏置参数缓存器中；多功能计算模块从第一结果缓存器中读取向量矩阵计算的结果中的第一结果，以根据第一结果进行神经网络模型中的多功能计算的一种方式，包括：多功能计算模块从偏置参数缓存器读取目标偏置参数，并根据目标偏置参数和第一结果进行神经网络模型中的

多功能计算。

[0093] 在一些示例性的实施方式中，多功能计算模块在进行神经网络模型中的多功能计算时，还包括：多功能计算模块将多功能计算的临时计算结果写入计算模组中的临时缓存模块中；多功能计算模块从偏置参数缓存器读取目标偏置参数，并根据目标偏置参数和第一结果，进行神经网络模型中的多功能计算的一种方式，包括：多功能计算模块从临时缓存模块中，读取多功能计算的临时计算结果中第三结果；多功能计算模块根据目标偏置参数、第一结果以第三结果，进行神经网络模型中的多功能计算。

[0094] 本实施例中，数据处理设备包括多级缓存控制模块、多级缓存器以及计算模组，计算模组与多级缓存器中的末级缓存器连接。基于这种设备结构，多级缓存控制模块可控制数据在多级缓存器之间移动，计算模组可在末级缓存器提供的数据的支持下，进行数据处理操作。进而，实现了数据移动过程和数据处理过程并行的技术效果，有利于提升数据处理设备的数据处理效率。

[0095] 需要说明的是，上述实施例所提供方法的各步骤的执行主体均可以是同一设备，或者，该方法也由不同设备作为执行主体。比如，步骤301至步骤304的执行主体可以为设备A；又比如，步骤301和302的执行主体可以为设备A，步骤303的执行主体可以为设备B；等等。

[0096] 另外，在上述实施例及附图中的描述的一些流程中，包含了按照特定顺序出现的多个操作，但是应该清楚了解，这些操作可以不按照其在本文中出现的顺序来执行或并行执行，操作的序号如301、302等，仅仅是用于区分开各个不同的操作，序号本身不代表任何的执行顺序。另外，这些流程可以包括更多或更少的操作，并且这些操作可以按顺序执行或并行执行。需要说明的是，本文中的“第一”、“第二”等描述，是用于区分不同的消息、设备、模块等，不代表先后顺序，也不限定“第一”和“第二”是不同的类型。

[0097] 本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明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，本发明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。

[0098] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0099] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0100] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一

个方框或多个方框中指定的功能的步骤。

[0101] 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。

[0102] 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。

[0103] 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、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暂存电脑可读媒体(transitory media)，如调制的数据信号和载波。

[0104] 还需要说明的是，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、商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、商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、商品或者设备中还存在另外的相同要素。

[0105] 以上所述仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。

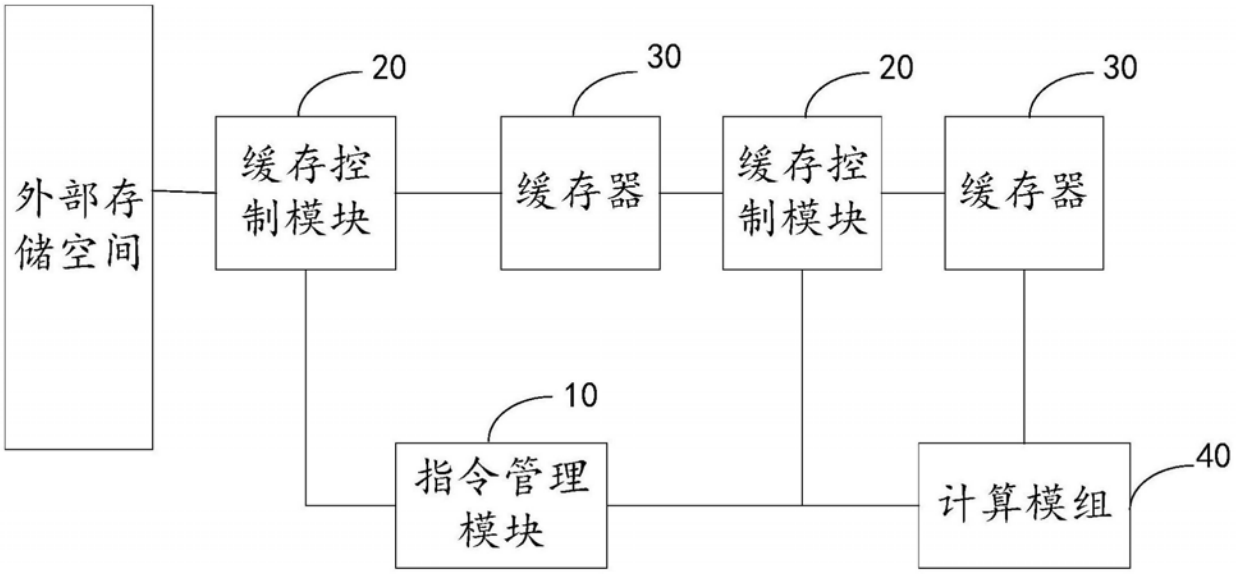

外部存储空间
20
缓存控制模块
30
缓存器
20
缓存控制模块
30
缓存器
10
指令管理模块
计算模组
40

图1

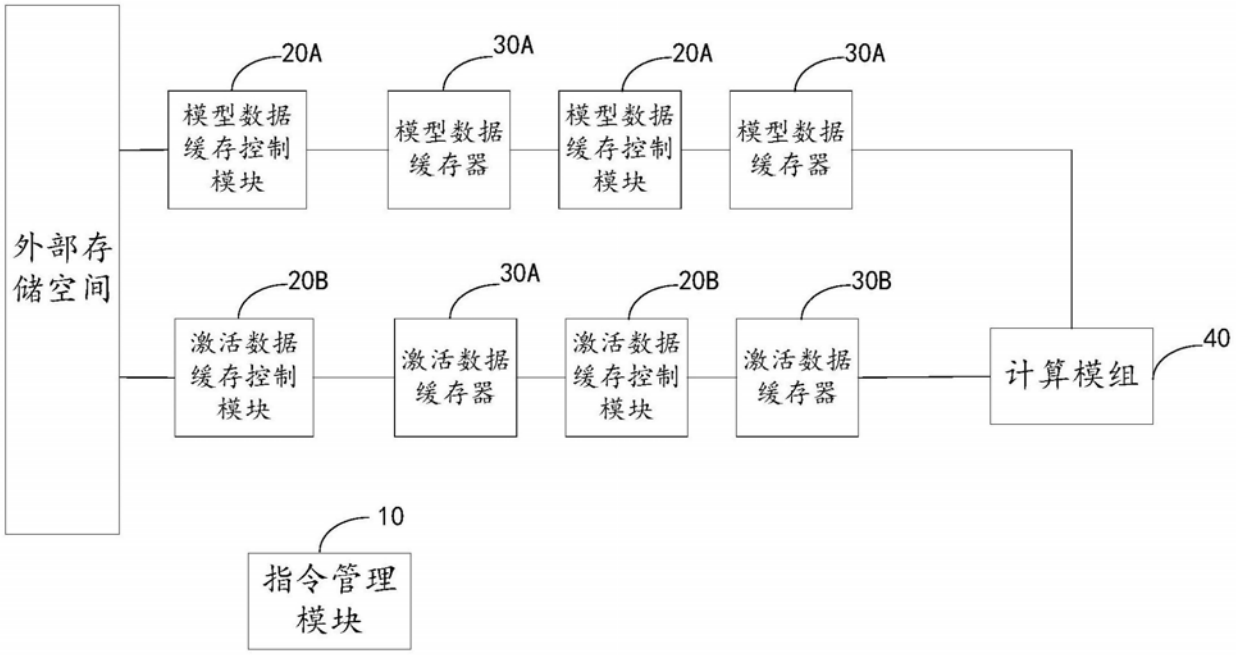

外部存储空间
20A
模型数据缓存控制模块
30A
模型数据缓存器
20A
模型数据缓存控制模块
30A
模型数据缓存器
20B
激活数据缓存控制模块
30A
激活数据缓存器
20B
激活数据缓存控制模块
30B
激活数据缓存器
计算模组
40
10
指令管理模块

图2a

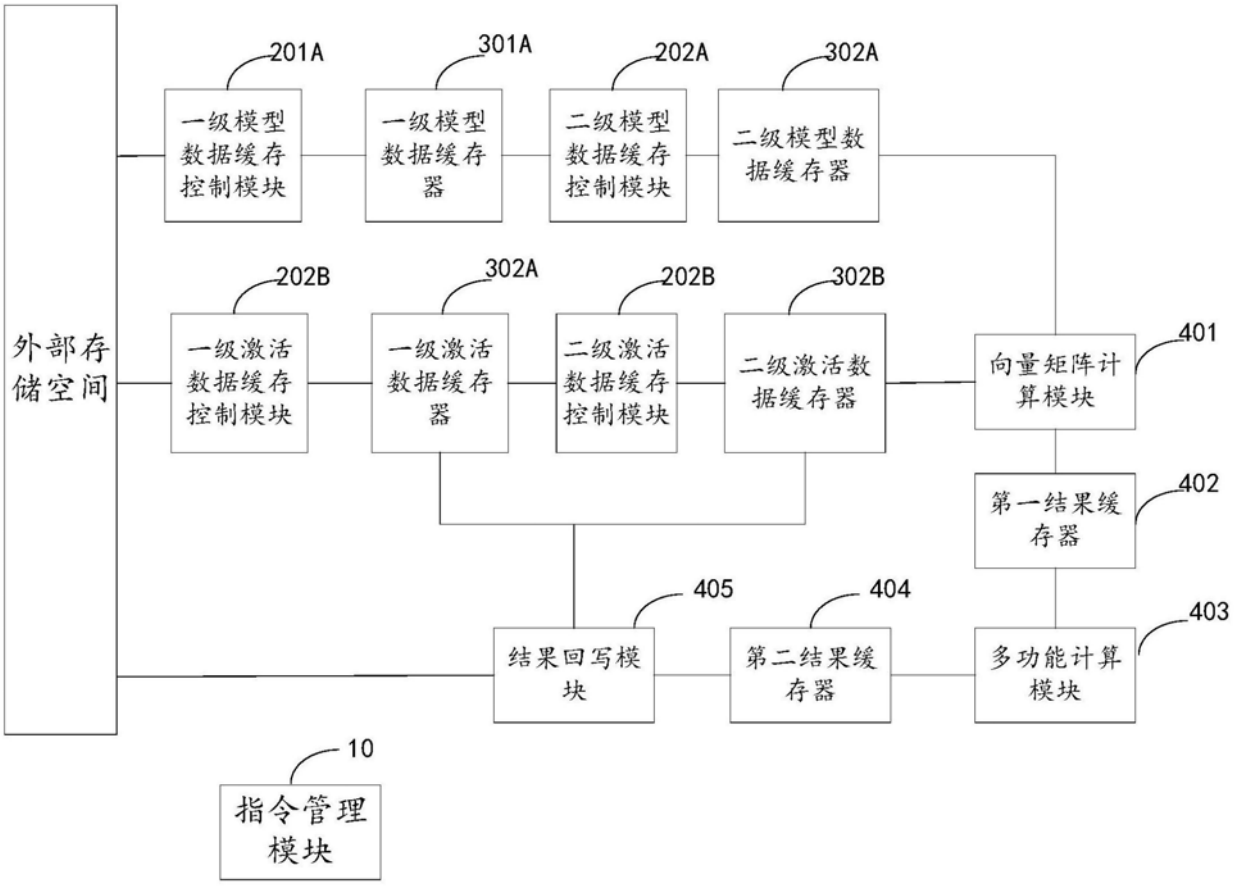

外部存储空间
201A
一级模型数据缓存控制模块
301A
一级模型数据缓存器
202A
二级模型数据缓存控制模块
302A
二级模型数据缓存器
202B
一级激活数据缓存控制模块
302A
一级激活数据缓存器
202B
二级激活数据缓存控制模块
302B
二级激活数据缓存器
401
向量矩阵计算模块
402
第一结果缓存器
403
多功能计算模块
404
第二结果缓存器
405
结果回写模块
10
指令管理模块

图2b

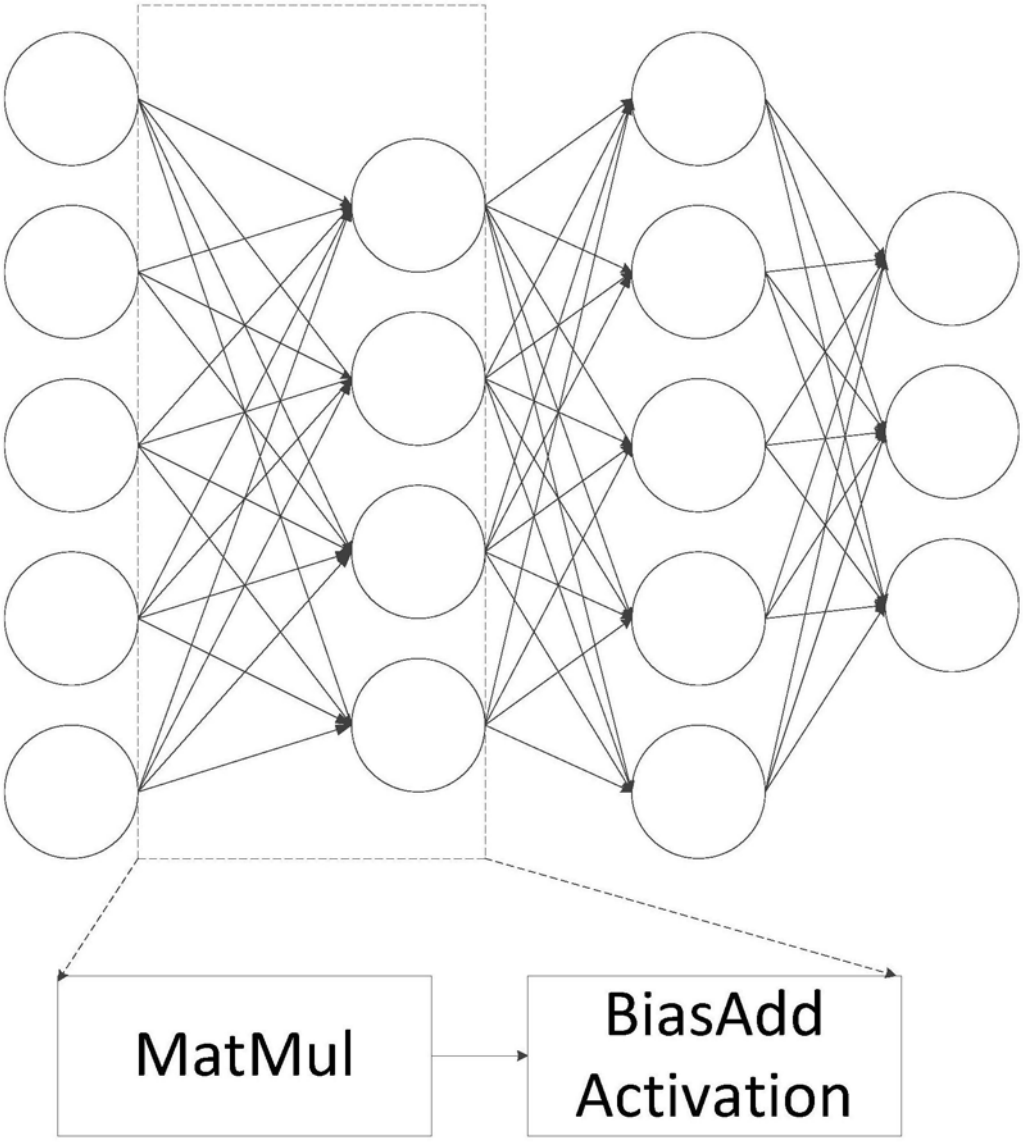


图2c

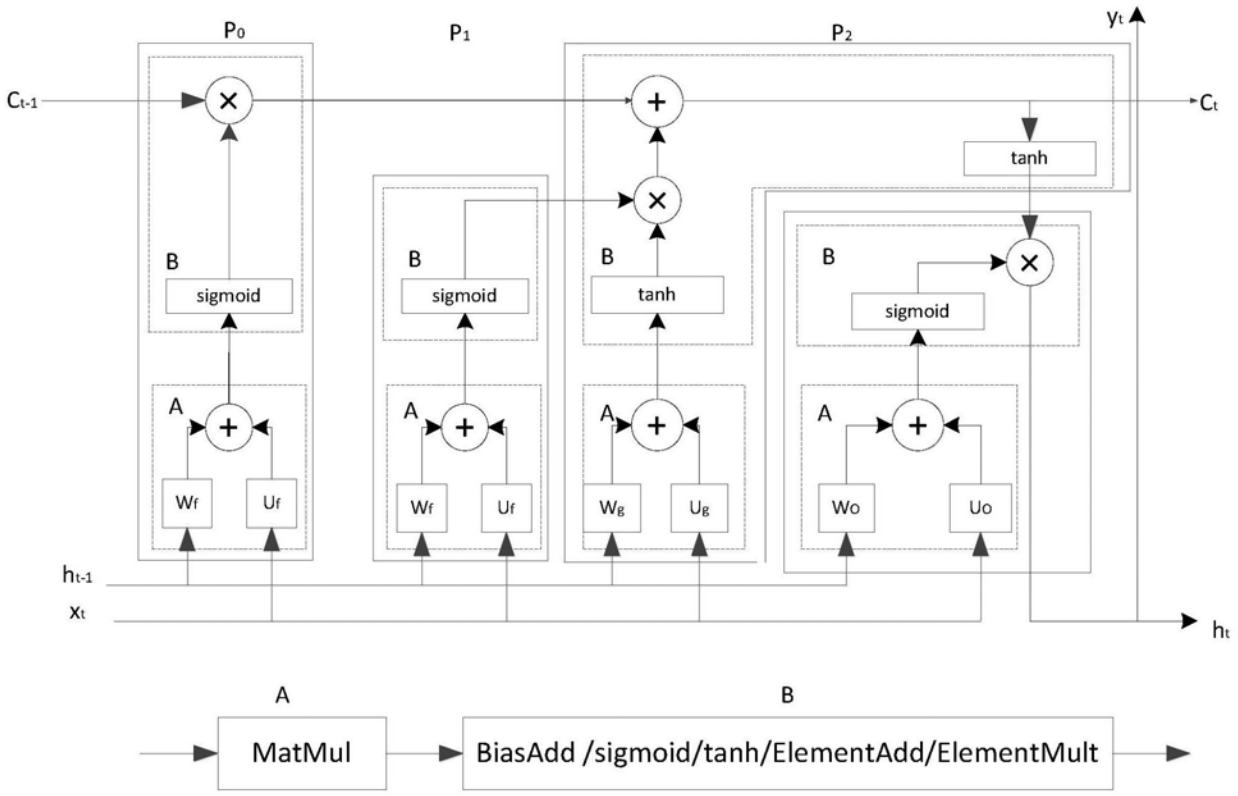

P0
P1
P2
yt
Ct-1
Ct
tanh
B
sigmoid
A
Wf
Uf
tanh
Wg
Ug
sigmoid
Wo
Uo
ht-1
xt
ht
A
B
MatMul
BiasAdd /sigmoid/tanh/ElementAdd/ElementMult

图2d

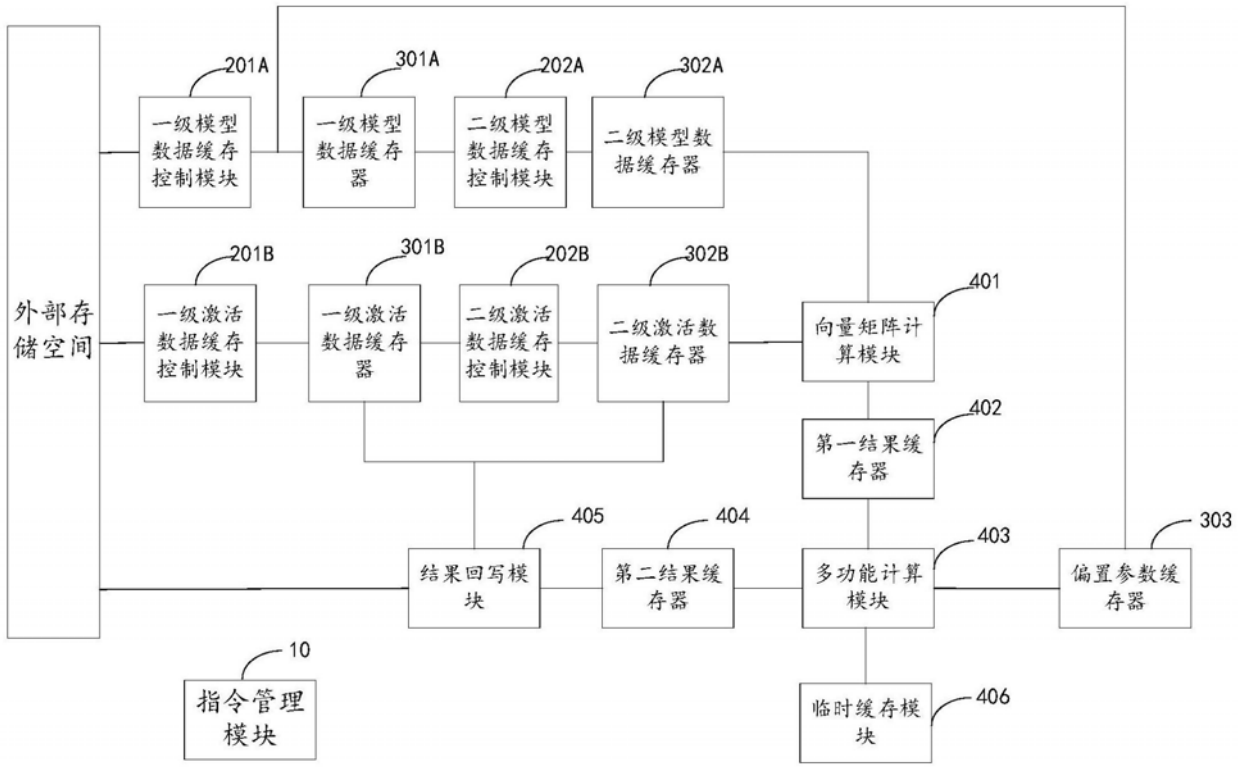

外部存储空间
201A
一级模型数据缓存控制模块
301A
一级模型数据缓存器
202A
二级模型数据缓存控制模块
302A
二级模型数据缓存器
201B
一级激活数据缓存控制模块
301B
一级激活数据缓存器
202B
二级激活数据缓存控制模块
302B
二级激活数据缓存器
401
向量矩阵计算模块
402
第一结果缓存器
405
结果回写模块
404
第二结果缓存器
403
多功能计算模块
303
偏置参数缓存器
10
指令管理模块
406
临时缓存模块

图2e

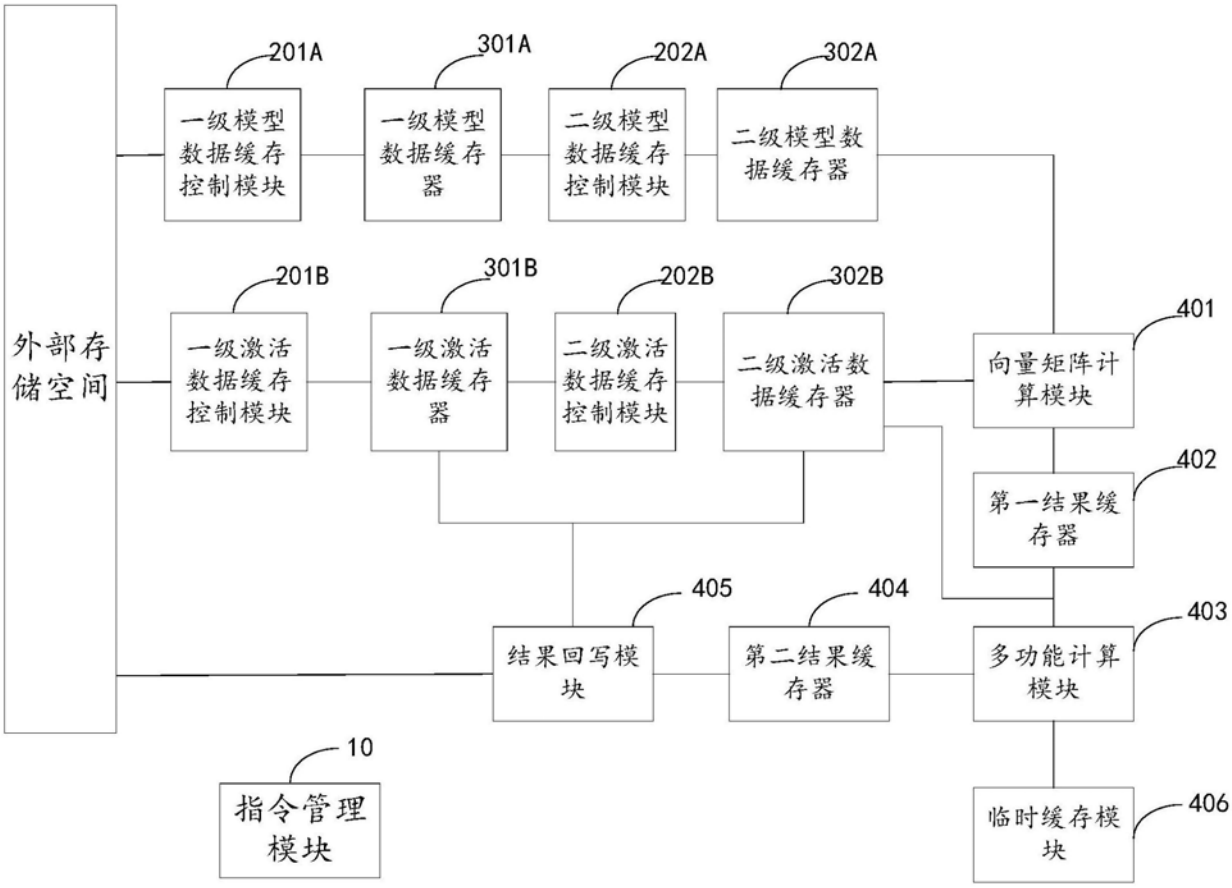

外部存储空间
201A
一级模型数据缓存控制模块
301A
一级模型数据缓存器
202A
二级模型数据缓存控制模块
302A
二级模型数据缓存器
201B
一级激活数据缓存控制模块
301B
一级激活数据缓存器
202B
二级激活数据缓存控制模块
302B
二级激活数据缓存器
401
向量矩阵计算模块
402
第一结果缓存器
403
多功能计算模块
404
第二结果缓存器
405
结果回写模块
406
临时缓存模块
10
指令管理模块

图2f

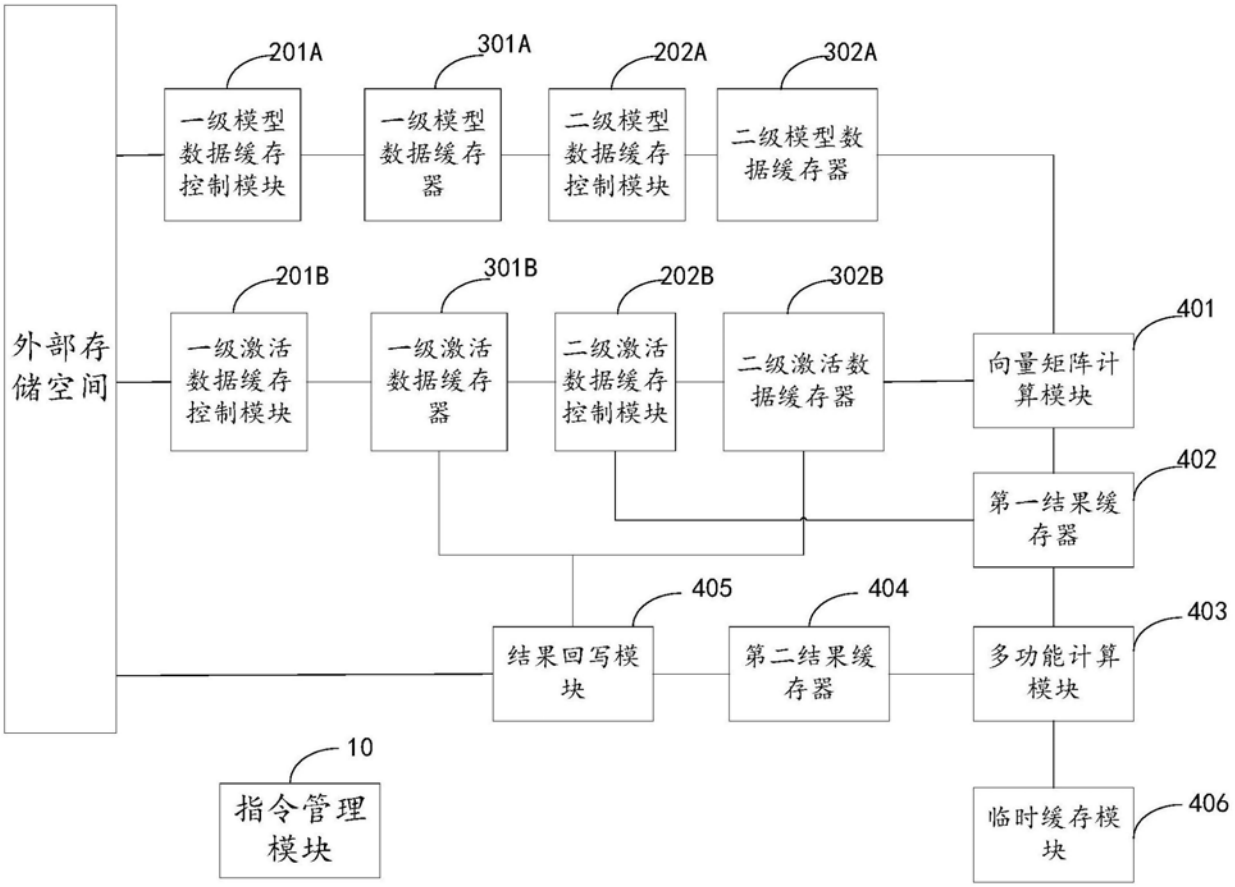

外部存储空间
201A
一级模型数据缓存控制模块
301A
一级模型数据缓存器
202A
二级模型数据缓存控制模块
302A
二级模型数据缓存器
201B
一级激活数据缓存控制模块
301B
一级激活数据缓存器
202B
二级激活数据缓存控制模块
302B
二级激活数据缓存器
401
向量矩阵计算模块
402
第一结果缓存器
405
结果回写模块
404
第二结果缓存器
403
多功能计算模块
10
指令管理模块
406
临时缓存模块

图2g

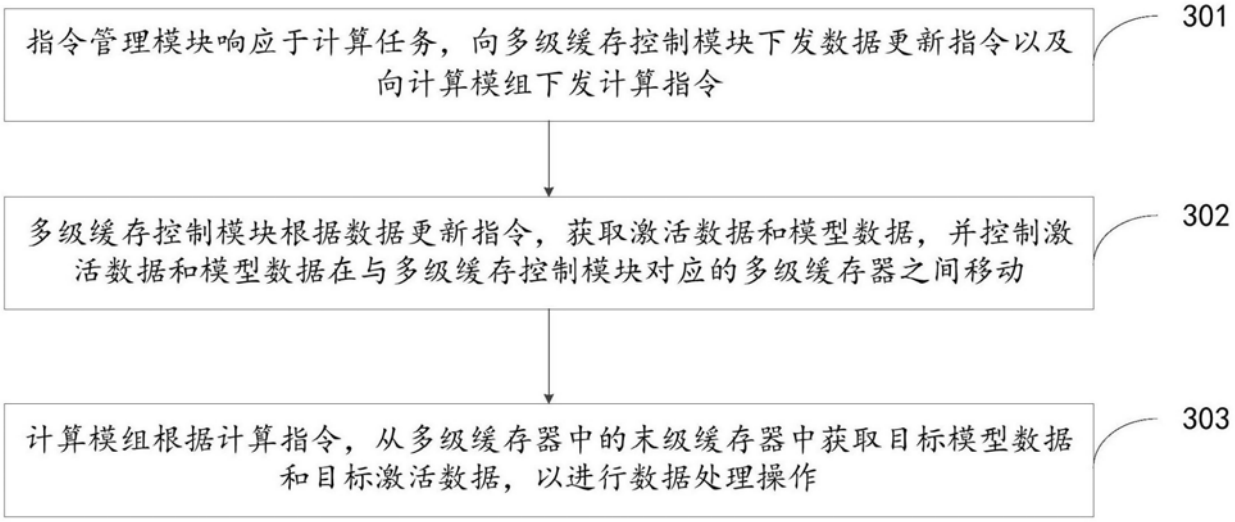

指令管理模块响应于计算任务，向多级缓存控制模块下发数据更新指令以及向计算模组下发计算指令
301
多级缓存控制模块根据数据更新指令，获取激活数据和模型数据，并控制激活数据和模型数据在与多级缓存控制模块对应的多级缓存器之间移动
302
计算模组根据计算指令，从多级缓存器中的末级缓存器中获取目标模型数据和目标激活数据，以进行数据处理操作
303

图3