(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116909984 A
(43)申请公布日 2023.10.20

(21)申请号 202310865940.6

(22)申请日 2018.04.06

(30)优先权数据
62/486,432 2017.04.17 US
15/694,663 2017.09.01 US

(62)分案原申请数据
201880025504.5 2018.04.06

(71)申请人 微软技术许可有限责任公司
地址 美国华盛顿州

(72)发明人 G·彼得 C·B·麦克布赖德
A·A·安巴德卡 K·D·塞多拉
B·博布罗夫 L·M·瓦尔

(74)专利代理机构 北京市金杜律师事务所 11256
专利代理师 赵林琳 张鑫

(51)Int.Cl.
G06F 15/80(2006.01)
G06F 9/50(2006.01)
G06F 13/16(2006.01)
G06F 12/08(2016.01)
G06F 9/38(2018.01)
G06N 3/0464(2023.01)
G06N 3/048(2023.01)
G06N 3/08(2023.01)

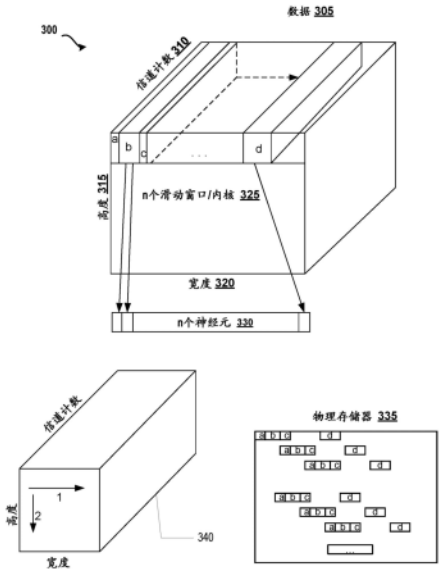
权利要求书3页 说明书13页 附图9页

(54)发明名称

使用虚拟化数据迭代器对神经网络进行数据处理性能增强

(57)摘要

本申请涉及使用虚拟化数据迭代器对神经网络进行数据处理性能增强的系统、方法、计算机可读存储介质。使用虚拟化硬件迭代器，由NN/DNN处理的数据可以被遍历并且被配置为优化操作的数目以及存储器利用率，以增强NN/DNN的整体性能。可操作地，迭代器控制器可以生成用于由NN/DNN执行的指令，指令表示一个或多个期望的迭代器操作类型并执行一个或多个迭代器操作。可以根据所选择的迭代器操作来将数据迭代，并将数据通信到NN/DD的一个或多个神经元处理器来进行处理并输出到目标存储器。迭代器操作可以应用于并行的各种数据量(例如，二进制大对象)或相同容量的多个切片。



CN 116909984 A

1.一种系统,包括:

至少一个处理器;

至少一个硬件迭代器,被配置为执行用于一个或多个数据迭代功能的并行处理的多个实例;以及

与所述至少一个处理器通信的至少一个存储器,所述至少一个存储器具有存储在其上的计算机可读指令,当所述计算机可读指令由所述至少一个处理器执行时,使得所述系统:

接收一个或多个初始化参数,所述初始化参数包括表示将由神经网络环境处理的数据的维度的数据;

从所述神经网络环境的存储器组件加载所述数据;

接收一个或多个指令,以根据一个或多个迭代器操作类型和所述一个或多个初始化参数并且使用一个或多个所选择的遍历模式来遍历加载的所述数据;

对所述一个或多个指令进行处理,以根据一个或多个迭代器操作类型和所述一个或多个初始化参数选择加载的所述数据的一个或多个部分;以及

将加载的所述数据的所述一个或多个部分通信到所述神经网络环境的一个或多个处理组件,

其中所述硬件迭代器使用向所述存储器组件中的所述数据的逻辑映射被虚拟化,所述逻辑映射使用指示所述数据的物理存储器寻址、维度和容量的参数被确定。

2.根据权利要求1所述的系统,其中加载的所述数据的所述一个或多个部分是相等部分。

3.根据权利要求1所述的系统,其中所述计算机可读指令还使得所述系统根据由接收的所述一个或多个初始化参数限定的选择数据量来遍历所述数据,所述参数包括包含以下的数据:数据高度、数据宽度、信道数目、内核数目和一个或多个其他数据描述符。

4.根据权利要求3所述的系统,其中所述计算机可读指令还使得所述系统利用一个或多个滑动窗口遍历所述数据,所述滑动窗口可操作来选择所述数据量的一个或多个数据元素作为通信到所述一个或多个处理组件的所述一个或多个部分。

5.根据权利要求4所述的系统,其中所述计算机可读指令还使得所述系统使用一个或多个滑动窗口来遍历加载的所述数据,所述一个或多个滑动窗口跨越加载的所述数据的数据维度边界。

6.根据权利要求1所述的系统,其中所述计算机可读指令还使得所述系统将一个或多个数据填充插入到加载的所述数据。

7.根据权利要求1所述的系统,其中所述计算机可读指令还使得所述系统生成表示由所述神经网络环境的一个或多个神经处理器处理的数据的数据的输出量,所述输出量根据由所述神经网络环境的控制器组件接收的一个或多个指令被生成。

8.一种用于在计算设备中使用迭代器的增强数据处理的计算机实现的方法,所述迭代器被配置为执行用于一个或多个数据迭代功能的并行处理的多个实例,所述方法包括:

从所述计算设备的存储器组件加载数据;

接收一个或多个初始化参数,所述初始化参数指示所述数据的维度作为数据量,所述数据量包括高度、宽度、信道数目、内核数目、或者一个或多个其他数据描述符中的任一个;

接收一个或多个指令,以根据一个或多个迭代器操作类型和所述一个或多个初始化参

数并且使用一个或多个所选择的遍历模式来遍历加载的所述数据；

对所述一个或多个指令进行处理，以根据一个或多个迭代器操作类型和所述一个或多个初始化参数选择加载的所述数据的一个或多个部分；以及

将加载的所述数据的所述一个或多个部分通信到所述计算设备的一个或多个处理组件，

其中所述迭代器使用向所述存储器组件中的所述数据的逻辑映射被虚拟化，所述逻辑映射使用指示所述数据的物理存储器寻址、维度和容量的参数被确定。

9.根据权利要求8所述的计算机实现的方法，其中加载的所述数据的所述一个或多个部分是相等部分。

10.根据权利要求8所述的计算机实现的方法，其中加载的所述数据的所述遍历利用可操作以跨越所述数据的数据维度边界的一个或多个滑动窗口。

11.根据权利要求8所述的计算机实现的方法，还包括：

将填充子量插入加载的所述数据中。

12.根据权利要求11所述的计算机实现的方法，其中所述填充子量由接收的所述一个或多个指令以及由接收的所述一个或多个初始化参数来限定。

13.根据权利要求8所述的计算机实现的方法，还包括：

针对由所述一个或多个处理组件处理的数据，生成输出数据量，所述输出数据量由接收的所述一个或多个指令限定。

14.根据权利要求13所述的计算机实现的方法，还包括：

丢弃所生成的所述输出数据量的一个或多个部分。

15.根据权利要求8所述的计算机实现的方法，还包括：将加载的所述数据的遍历位置作为内部状态数据存储在所述计算设备的存储器组件中。

16.一种计算机可读存储介质，具有存储在其上的计算机可执行指令，当所述计算机可执行指令由计算设备的一个或多个处理器执行时，使得所述计算设备：

实例化迭代器，所述迭代器被配置为执行用于一个或多个数据迭代功能的并行处理的多个实例；

接收一个或多个初始化参数，所述初始化参数包括表示待处理的数据的维度的数据，所述数据的所述维度包括表示所述数据的容量的数据，所述一个或多个初始化参数包括待处理的所述数据的物理存储器寻址；

利用所述一个或多个初始化参数加载所述数据；

接收一个或多个指令，以根据一个或多个迭代器操作类型和所述一个或多个初始化参数并且使用一个或多个所选择的遍历模式来遍历加载的所述数据；

对所述一个或多个指令进行处理，以根据所述一个或多个迭代操作类型和所述一个或多个初始化参数选择加载的所述数据的一个或多个部分；以及

将加载的所述数据的所述一个或多个部分通信到一个或多个处理组件；

其中所述迭代器使用向所述数据的逻辑映射被虚拟化，所述逻辑映射使用指示所述数据的物理存储器寻址、维度和容量的参数被确定。

17.根据权利要求16所述的计算机可读存储介质，其中所述一个或多个指令包括用于存储表示加载的所述数据的所述遍历的内部状态的数据的指令，其中所述计算机可执行指

令还使得所述计算设备：

将所述内部状态数据存储在所述计算设备的存储器组件中。

18.根据权利要求17所述的计算机可读存储介质，其中所述计算机可执行指令还使得所述计算设备：

将附加数据量插入到加载的所述数据。

19.根据权利要求16所述的计算机可读存储介质，其中所述计算机可执行指令还使得所述计算设备：

生成表示由所述一个或多个处理组件处理的数据的输出数据量。

20.根据权利要求16所述的计算机可读存储介质，其中所述计算机可执行指令还使得所述计算设备：

利用加载的所述数据的逻辑数据映射来遍历加载的所述数据，加载的所述数据的所述遍历包括将一个或多个滑动窗口应用于所述逻辑数据映射，以将加载的所述数据的一部分与一个或多个物理存储器地址相关联。

使用虚拟化数据迭代器对神经网络进行数据处理性能增强

[0001] 本申请为发明名称为“使用虚拟化数据迭代器对神经网络进行数据处理性能增强”的原中国发明专利申请的分案申请。原申请的申请号为201880025504.5、PCT申请号为PCT/US2018/026353，原申请的申请日为2018年4月6日、PCT国际申请进入国家阶段日为2019年10月16日。

背景技术

[0002] 在人工神经网络(NN)中，神经元是用于对大脑中的生物神经元进行建模的基本单元。人工神经元的模型包括输入向量和权重向量的内积，权重向量被添加到应用非线性的偏置。对于深度神经网络(DNN)(例如，如示例性DNN模块所表示的)，神经元可以被紧密地映射到人工神经元。

[0003] 在跨NN或DNN处理数据时，执行示例性处理操作的控制器需要对大量数据进行迭代，以应用可以影响整体NN或DNN性能的特定操作，从而导致关键时延而损害期望的所述处理目标(例如，标识示例性输入数据中的对象和/或对象特征，图像、声音、地理坐标等)。通常地，现有NN和DNN在执行各种操作时花费可避免的处理时间(例如，每秒浮点/定点操作(GFlops/s))和存储器空间(例如，每秒传递的字节数(GBytes/s))。具体地，当前实践不标识输入/数据的关键特征和/或向NN或DNN的协作组件提供关于如何最好地处理这样的输入数据来避免这样的性能问题的指令。与NN或DNN中的低效数据处理相关联的性能影响包括NN或DNN的本地和外部存储器组件之间的低效管理。这样的低效数据管理需要附加的、通常可以避免的计算/神经处理器操作，从而对整体NN/DNN性能进一步施压。

[0004] 更有利的NN/DNN将部署一个或多个虚拟化硬件迭代器，一个或多个虚拟化硬件迭代器可操作地允许NN/DNN组件、能够指定数据的维度来进行迭代并向NN/DNN组件提供一个或多个指令来将数据迭代。为了支持附加的扩展，可以将硬件迭代器部署为交替/同时运行的单个/多个实例，和/或可以表征数据来允许并行处理操作。

[0005] 关于这些考虑和其他考虑，呈现了本文所公开的内容。

发明内容

[0006] 本文所描述的技术提供了在示例性神经网络(NN)和/或深度神经网络(DNN)环境中利用的一个或多个硬件迭代器的虚拟化，其中迭代器(例如，表示为示例性NN和/或DNN环境的迭代器控制器组件)可操作地允许改进整体性能并优化存储器管理的数据处理。在一个说明性实现中，示例性DNN环境可以包括一个或多个处理块(例如，计算机处理单元-CPU)、存储器控制器、高带宽结构(例如，在示例性DNN模块和DNN环境的协作组件之间传递数据和/或数据元素的数据总线)、迭代器控制器、操作控制器，以及DNN模块。在说明性实现中，示例性DNN模块可以包括示例性DNN状态控制器、描述符列表控制器(DLC)、dMA(DDMA)、DMA流激活(DSA)、操作控制器、加载控制器和存储控制器。

[0007] 在一个说明性操作中，NN/DNN环境的操作控制器可操作地将大量数据迭代，以应用一个或多个期望的数据处理操作(例如，卷积、最大池化、标量乘/加、求和、完全连接等)。

在说明性操作中，参与用户可以指定正在迭代的数据的维度以及关于如何迭代数据以供NN/DNN计算环境使用的配置。说明性地，本文描述的迭代器可以在交替运行的多个实例中可操作地执行，以适应一个或多个输入数据和/或单个输入数据的一个或多个部分的并行处理。说明性地，可以将迭代器的内部状态保存到专用协作存储器中，以供示例性迭代器控制器使用，以在将数据迭代以由示例性NN/DNN环境的神经元处理器进行处理时生成指令。

[0008] 在一个说明性实现中，将由NN/DNN环境处理的数据可以表示为二进制大对象(blob)。通常，二进制大对象表示存储器中需要进行迭代的数据。每个二进制大对象可以维持由各种维度(例如，宽度、高度、信道数目、内核数目和其他可用维度单位)限定的逻辑映射形状。在一个说明性操作中，迭代器可遍历多维二进制大对象(例如，如逻辑数据映射限定)或这样的二进制大对象的较小N维度切片，其中N是维度数(例如，对于表示具有宽度、高度和信道数目的图像的3D二进制大对象，N＝3)。在遍历二进制大对象时，迭代器控制器可以生成一个或多个指令，一个或多个指令包括但不限于：用于将数据从源存储器加载到(多个)处理单元(例如，神经元处理器)的加载指令、或用于将由(多个)处理单元产生的数据存储到目标存储器(例如，NN/DNN环境的协作存储器组件)的存储指令。在说明性操作中，迭代器控制器能够同时产生读取/写入多个数据的指令。

[0009] 说明性地，迭代器控制器可以暴露若干操作(操作类型)以供NN/DNN环境执行，操作包括但不限于：初始化控制，允许跨越特定二进制大对象实例的迭代；填充控制，允许从当前位置处的示例性二进制大对象的数据提取；跳过控制，可以在不提取任何数据的情况下，允许内部数据指针的推进；二进制大对象结束标志，允许向参与用户通信完成二进制大对象遍历；重置控制，可以允许在第一数据位置处处理二进制大对象；保存控制，可以允许将迭代器控制器的内部状态存储到外部存储器位置以供将来重用；以及恢复控制，可以允许存储在协作外部存储器上的内部状态被恢复来供NN/DNN环境使用。根据处理数据的需要，可以由NN/DNN环境来实现和执行附加控制和/或标志。

[0010] 在一个说明性操作中，迭代器控制器可以使用一个或多个所选择的遍历模式向NN/DNN环境提供指令来遍历二进制大对象存储器，遍历模式可以特定于每个迭代器控制器操作并且可以由迭代器控制器进行进一步配置。在说明性操作中，迭代器控制器可以允许在由所存储的内部状态描述的位置处读取/写入数据以及随后的数据处理。可以从存储器读取数据/向存储器写入数据的顺序可以异步，并且可以取决于迭代器控制器操作类型和针对特定操作类型的配置数据。此外，实现每个层类型的操作控制器利用一个或多个迭代器类型，以从存储器提取数据来进行处理，并且然后在处理完成之后将数据保存到存储器。

[0011] 此外，迭代器控制器可以允许相同迭代器操作类型的多个实例被执行。在一个说明性操作中，多个迭代器操作类型可以被并行地执行，可操作地通过在存储器中存储/检索迭代器操作类型的状态来实现。作为对NN/DNN环境性能的进一步增强，当使用相同迭代器类型的单个实例时，迭代器可以被配置为跳过保存/恢复操作并返回到初始状态。

[0012] 在一个说明性操作中，示例性迭代器控制器可以提供执行二进制大对象迭代器操作类型的指令(即，二进制大对象迭代器)。在说明性操作中，二进制大对象迭代器可以具有一个或多个指令来遍历部分连续数据的区块中的二进制大对象，和/或具有一个或多个指令来遍历用于与一个或多个神经元处理器通信的命名容量。容量可以表示需要由n个神经元处理二进制大对象来产生n个输出值的数据。该数据可以是连续的并且可以顺序处理，或

也可以是不连续的并且可以多次处理。

[0013] 由二进制大对象迭代器读取的数据可以被通信到一个或多个神经元处理器来进行处理。可操作地，神经元处理器可以维持先进先出(FIFO)总线，在FIFO总线中，神经元处理器可以可操作地接收它将处理的数据。二进制大对象迭代器可以通过在每个神经元处理器FIFO中放置选定量(例如，等量)的值来通信所遍历的二进制大对象数据，以允许将二进制大对象数据分配到n个神经元处理器FIFO中。

[0014] 特定于迭代器控制器操作类型的填充命令，二进制大对象迭代器可以从协作存储器读取多个二进制大对象容量宽度的数据行，并且对于每个容量行，它可以将从一个或多个神经元处理器映射的值分布到相应神经元处理器FIFO。可以读取相同容量行上的数据，然后将其分布到多个神经元处理器FIFO。

[0015] 特定于具有宽度不是容量宽度的倍数的二进制大对象，可操作地，迭代器控制器可以提供指令来处理可以跨越容量宽度行中的多个行的容量。在一个说明性操作中，由迭代器控制器提供的指令允许二进制大对象迭代器使用多个容量宽度行跨越操作来遍历二进制大对象。

[0016] 在一个说明性操作中，二进制大对象迭代器可以允许在运行时向二进制大对象动态地添加填充容量。操作控制器可以指定在二进制大对象的每个维度上使用多少填充，并且类似于二进制大对象迭代器存在于协作物理存储器中，二进制大对象迭代器可以跟踪该填充容量。然后填充可以用于下一操作。

[0017] 在一个说明性操作中，示例性迭代器控制器可以提供执行输出迭代器操作类型的指令(例如，输出迭代器)。在一个说明性操作中，输出迭代器可以将由一个或多个神经元处理器产生的数据存储到输出二进制大对象中。输出迭代器可以被配置有输出二进制大对象的维度和产生数据的n个神经元处理器，以生成用于将经处理的数据存储到目标存储器中的一个或多个指令。说明性地，输出迭代器可以写入完整二进制大对象或仅写入二进制大对象的某些部分。

[0018] 应当理解，尽管关于系统进行了描述，但是上述主题还可以实现为计算机控制的装置、计算机进程、计算系统、或者诸如计算机可读介质和/或专用芯片组的制品。通过阅读以下详细描述和对相关附图的回顾，这些特征和各种其他特征将显而易见。提供本发明内容是为了以简化的形式介绍一些概念，这些概念将在下面的具体实施方式中进一步描述。

[0019] 本发明内容不旨在标识所要求保护的主题的关键特征或必要特征，也不旨在将本发明内容用于限制所要求保护的主题的范围。此外，所要求保护的主题不限于解决在本公开的任何部分中提到的任何缺点或所有缺点的实现。

附图说明

[0020] 参考附图描述了具体实现。在附图中，附图标记的最左边(多个)数字标识首次出现附图标记的图。不同图中的相同附图标记表示相似或相同的项。对多个项中的各个项的参考可以使用具有字母序列的字母的附图标记来指代每个单独的项。对项的通用参考可以使用不具有字母序列的特定附图标记。

[0021] 图1图示了根据本文所描述的系统和方法的示例性神经网络计算环境的框图。

[0022] 图2图示了利用一个或多个虚拟化硬件迭代器的示例性神经网络环境的框图。

[0023] 图3图示了根据本文所描述的系统和方法的说明性逻辑数据映射中表示的示例性输入数据的框图。

[0024] 图4图示了在说明性逻辑数据映射中表示的示例性输入数据的框图,示出了使用说明性n个滑动窗口来操作,以跨越说明性逻辑数据映射的一个或多个行。

[0025] 图5图示了根据本文所描述的系统和方法的、在说明性逻辑数据映射中表示的示例性输入数据的框图,示出了使用说明性n个滑动窗口来操作,以跨越说明性逻辑数据映射的一个或多个行,从而允许数据填充作为处理增强。

[0026] 图6是根据本文所描述的系统和方法的、用于在说明性神经网络计算环境中进行处理的数据迭代的说明性过程的流程图。

[0027] 图7是根据本文所描述的系统和方法的、用于在说明性神经网络计算环境中利用所选择的逻辑数据映射协议进行处理的数据迭代的说明性过程的流程图,逻辑数据映射协议可操作来提供数据填充和经跨越的滑动窗口。

[0028] 图8示出了能够执行本文所描述的方法的计算机的说明性计算机架构的附加细节。

[0029] 图9示出了根据本文所描述的系统和方法协作的说明性计算设备的附加细节。

具体实施方式

[0030] 以下详细描述描述了提供在示例性神经网络(NN)和/或深度神经网络(DNN)环境中使用的一个或多个硬件迭代器的虚拟化的技术。通常,迭代器(例如,被表示为示例性NN和/或DNN环境的迭代器控制器组件)可操作地允许改进整体性能并优化存储器管理的数据处理。在一个说明性实现中,示例性DNN环境可以包括一个或多个处理块(例如,计算机处理单元CPU)、存储器控制器、高带宽结构(例如,在示例性DNN模块和DNN环境的协作组件之间传递数据和/或数据元素的数据总线)、迭代器控制器、操作控制器和DNN模块。在说明性实现中,示例性DNN模块可以包括示例性DNN状态控制器、描述符列表控制器(DLC)、dMA(DDMA)、DMA流激活(DSA)、操作控制器、加载控制器和存储控制器。

[0031] 在说明性操作中,NN/DNN环境的操作控制器可操作地将大量数据迭代,以应用一个或多个期望的数据处理操作(例如,卷积、最大池化、标量乘加、求和、完全连接等)。在说明性操作中,参与用户可以指定正在迭代的数据维度以及用于迭代数据以供NN/DNN计算环境使用的配置。说明性地,本文所描述的迭代器可以在独立地运行的多个实例中可操作地执行,以适应一个或多个输入数据和/或单个输入数据的一个或多个部分的并行处理。说明性地,迭代器的内部状态可以被保存到专用协作存储器中,以供示例性迭代器控制器使用,以在将数据迭代来由示例性NN/DNN环境的神经元处理器进行处理时生成指令。

[0032] 在一个说明性实现中,将由NN/DNN环境处理的数据可以表示为二进制大对象。通常地,二进制大对象表示存储器中需要进行迭代的数据。每个二进制大对象可以维持由各种维度(例如,宽度、高度、信道数目、内核数目和其他可用维度单位)限定的逻辑映射形状。在一个说明性操作中,迭代器可遍历多维二进制大对象(例如,如由逻辑数据映射限定)或这样的二进制大对象的较小N维度切片,其中N是维度数(例如,对于表示具有宽度、高度和信道数的图像的3D二进制大对象,N=3)。在遍历二进制大对象时,迭代器控制器可以生成一个或多个指令,一个或多个指令包括但不限于:用于将数据从源存储器加载到(多个)处

理单元(例如,神经元处理器)的加载指令、或用于将由(多个)处理单元产生的数据存储到目标存储器(例如,NN/DNN环境的协作存储器组件)的存储指令。在说明性操作中,迭代器控制器能够同时产生读取/写入多个数据的指令。

[0033] 说明性地,迭代器控制器可以暴露若干操作(操作类型)以供NN/DNN环境执行,操作包括但不限于:初始化控制,允许跨越特定二进制大对象实例的迭代;填充控制,允许从当前位置处的示例性二进制大对象的数据提取;跳过控制,可以在不提取任何数据的情况下,允许内部数据指针的推进;二进制大对象结束标志,允许向参与用户通信完成二进制大对象遍历;重置控制,可以允许在第一数据位置处处理二进制大对象;保存控制,可以允许将迭代器控制器的内部状态存储到外部存储器位置以供将来重用;以及恢复控制,可以允许存储在协作外部存储器上的内部状态被恢复来供NN/DNN环境使用。根据处理数据的需要,可以由NN/DNN环境来实现和执行附加控制和/或标志。

[0034] 在一个说明性操作中,迭代器控制器可以使用一个或多个所选择的遍历模式向NN/DNN环境提供指令来遍历二进制大对象存储器,遍历模式可以特定于每个迭代器控制器操作并且可以由迭代器控制器进行进一步配置。在说明性操作中,迭代器控制器可以允许在由所存储的内部状态描述的位置处读取/写入数据以及随后的数据处理。可以从存储器读取数据/向存储器写入数据的顺序可以异步,并且可以取决于迭代器控制器操作类型和针对所提供的操作类型的配置数据。此外,实现每个层类型的操作控制器利用一个或多个迭代器类型,以从存储器提取数据来进行处理,并且然后在处理完成之后将数据保存到存储器。

[0035] 此外,迭代器控制器可以允许相同迭代器操作类型的多个实例被执行。在一个说明性操作中,多个迭代器操作类型可以被并行地执行,可操作地通过在存储器中存储/检索迭代器操作类型的状态来实现。作为对NN/DNN环境性能的进一步增强,当使用相同迭代器类型的单个实例时,迭代器可以被配置为跳过保存/恢复操作并返回到初始状态。

[0036] 在一个说明性操作中,示例性迭代器控制器可以提供执行二进制大对象迭代器操作类型的指令(即,二进制大对象迭代器)。在说明性操作中,二进制大对象迭代器可以具有一个或多个指令来遍历部分连续数据的区块中的二进制大对象,和/或具有一个或多个指令来遍历用于与一个或多个神经元处理器通信的命名容量。容量可以表示需要由n个神经元处理二进制大对象来产生n个输出值的数据。该数据可以是连续的并且可以顺序处理,或也可以是不连续的并且可以多次处理。

[0037] 由二进制大对象迭代器读取的数据可以被通信到一个或多个神经元处理器来进行处理。可操作地,神经元处理器可以维持先进先出(FIFO)总线,在FIFO总线中,神经元处理器可以可操作地接收它将处理的数据。二进制大对象迭代器可以通过在每个神经元处理器FIFO中放置选定量(例如,等量)的值来通信所遍历的二进制大对象数据(还被称为“所遍历的二进制大对象数据”或“所选择的数据”),以允许将二进制大对象数据分配到n个神经元处理器FIFO中。在一些配置中,本文所公开的技术可以包括处理一个或多个指令以根据一个或多个迭代器操作类型和一个或多个初始化参数来选择所加载的数据的一个或多个部分,然后将所加载的数据的一个或多个部分通信到神经网络环境的一个或多个处理组件。出于说明性目的,所加载的数据的一个或多个部分在本文中也称为“所遍历的二进制大对象数据”或“所选择的数据”。

[0038] 特定于迭代器控制器操作类型的填充命令,二进制大对象迭代器可以从协作存储器读取多个二进制大对象容量宽度的数据行,并且对于每个容量行,它可以将从一个或多个神经元处理器映射的值分布到相应神经元处理器FIFO。可以读取相同容量行上的数据,然后将其分布到多个神经元处理器FIFO。

[0039] 特定于具有宽度不是容量宽度的倍数的二进制大对象,可操作地,迭代器控制器可以提供指令来处理可以跨越容量宽度行中的多个行的容量。在一个说明性操作中,由迭代器控制器提供的指令允许二进制大对象迭代器使用多个容量宽度行跨越操作来遍历二进制大对象。

[0040] 在一个说明性操作中,二进制大对象迭代器可以允许在运行时向二进制大对象动态地添加填充容量。操作控制器可以指定在二进制大对象的每个维度上使用多少填充,并且类似于二进制大对象迭代器存在于协作物理存储器中,二进制大对象迭代器可以跟踪该填充容量。然后填充可以用于下一操作。

[0041] 在一个说明性操作中,示例性迭代器控制器可以提供执行输出迭代器操作类型的指令(例如,输出迭代器)。输出迭代器可以将由一个或多个神经元处理器产生的数据存储到输出二进制大对象中。输出迭代器可以被配置有输出二进制大对象的维度和产生数据的n个神经元处理器,以生成用于将经处理的数据存储到目标存储器中的一个或多个指令。说明性地,输出迭代器可以写入完整二进制大对象或仅写入二进制大对象的某些部分。

[0042] 神经网络背景:

[0043] 在人工神经网络中,神经元是用于对大脑中的生物神经元进行模拟的基本单元。人工神经元的模型可以包括输入向量和权重向量的内积,权重向量被添加到偏置并且应用了非线性。相比之下,示例性DNN模块中的神经元(例如,图1中的105)紧密地映射到人工神经元。

[0044] 说明性地,DNN模块可以被认为是超标量处理器。可操作地,它可以将一个或多个指令分派给被称为神经元的多个执行单元。执行单元可以是“同时分派同时完成”,其中每个执行单元与所有其他执行单元同步。DNN模块可以被分类为SIMD(单指令流、多数据流)架构。

[0045] 转向图1的示例性DNN环境100,DNN模块105具有存储器子系统,存储器子系统具有唯一的L1和L2高速缓存结构。这些高速缓存结构不是常规的高速缓存,而是专门针对神经处理而设计。为了方便,这些高速缓存结构采用了反映其预期目的的名称。作为示例,L2高速缓存125(A)可以说明性地维持1兆字节(1MB)的存储容量,具有高速专用接口以16GBps操作。L1高速缓存可以在内核和激活数据之间维持8KB的存储容量分配。L1高速缓存可以被称为行缓冲器,且L2高速缓存被称为BaSRAM。

[0046] DNN模块可以是仅召回(recall-only)的神经网络,并且以编程方式支持各种网络结构。网络训练可以在服务器场或数据中心中被离线执行,DNN模块不执行任何训练功能。训练的结果是可以被称为权重或内核的参数集合。这些参数表示可应用于输入的转换函数,其结果是分类或语义标记的输出。

[0047] 在一个说明性操作中,DNN模块可以接受平面数据作为输入。输入不仅限于图像数据,只要所呈现的数据是均匀的平面格式,DNN就可以对其进行操作。

[0048] DNN模块在对应于神经网络的层的层描述符列表上操作。说明性地,层描述符列表

可以由DNN模块处理为指令。这些描述符可以从存储器预获取到DNN模块中并按顺序执行。

[0049] 通常地,可以存在两个主要种类的层描述符:1)存储器到存储器移动描述符,以及2)操作描述符。存储器到存储器移动描述符可用于将数据从主存储器移出到本地高速缓存或从本地高速缓存移入到主存储器,以供操作描述符使用。存储器到存储器移动描述符遵循与操作描述符不同的执行管线。存储器到存储器移动描述符的目标管线可以是内部DMA引擎,而操作描述符的目标管线可以是神经元处理元件。操作描述符能够进行许多不同的层操作。

[0050] DNN的输出也是数据二进制大对象。输出可以可选地流式传输到本地高速缓存或流式传输到主存储器。DNN模块可以在软件允许的范围内预获取数据。软件可以通过在描述符之间使用屏蔽和设置相关性来控制预获取。具有相关性集合的描述符将被阻止前进,直到满足相关性。

[0051] 现在转向图1,示例性神经网络环境100可以包括各种协作组件,协作组件包括DNN模块105、高速缓冲存储器125(A)、低带宽结构110、桥接组件115、高带宽结构120、SOC 130、PCIE“端点”135、泰思立达(Tensilica)节点140、存储器控制器145、LPDDR4存储器105和输入数据源102。此外,如图所示,DNN模块105还可以包括若干组件,若干组件包括预获取105(A)、DMA 105(B)、寄存器接口105(D)、加载/存储单元105(C)、层控制器105(D)、保存/恢复组件105(E)和神经元105(F)。可操作地,示例性DNN环境100可以根据所选择的规范来处理数据,其中DNN模块执行如本文所述的一个或多个功能。

[0052] 图2图示了可操作以采用一个或多个虚拟化硬件迭代器作为数据处理的一部分的示例性神经网络环境200。如图所示,示例性神经网络环境200(这里也称为计算设备或计算设备环境)包括一个或多个操作控制器230,一个或多个操作控制器230与一个或多个迭代器控制器220协作来提供用于执行的一个或多个命令。一个或多个迭代器控制器220说明性地操作来生成借助示例性结构215通信到协作存储器组件210以及一个或多个神经元处理器205的指令。神经网络环境结构可以是能够传递各种数据的数据总线。此外,如图所示,一个或多个迭代器控制器220可以与迭代器状态存储器225可操作地协作,以保存和恢复迭代器状态存储器数据(未示出)。

[0053] 在说明性操作中,示例性神经网络环境200可以根据图6和图7中描述的过程可操作地处理数据。具体到图2中描述的组件,这些组件仅仅是说明性的,因为本领域一般技术人员将理解图6和图7中描述的、将由图2中所示的组件之外的其他组件执行的处理。

[0054] 此外,如图2所示,示例性神经网络环境可以可选地包括一个或多个硬件迭代器(如虚线所示),一个或多个硬件迭代器可以说明性地操作,以对输入数据(未示出)进行迭代来供一个或多个神经元处理器205处理。本领域技术人员可以理解,示例性的一个或多个硬件迭代器的这样的可选包括技术仅仅是说明性的,因为本文所公开的系统和方法描述的发明构思在没有任何硬件迭代器的情况下可以在示例性神经网络环境200中操作。

[0055] 图3图示了用于示例性输入数据的示例逻辑数据映射300。如图所示,数据305可以被表示为具有特定维度和容量340的数据,维度和容量340包括信道计数310、高度315和宽度320。根据本文描述的系统和方法,数据305可以被协作的n个神经元330分配和准备用于处理,使得第一部分a可以被通信到第一神经元、第二部分b可以被通信到第二神经元等,直到n个部分被通信到n个神经元。

[0056] 在一个说明性操作中，数据305的各部分可以基于由示例性神经网络环境(例如，图2的200)的协作控制器组件提供的一个或多个指令，使用n个滑动窗口/内核325被确定。进一步如图所示，可以使用由示例性神经网络环境(例如，图2的200)的协作操作控制器组件提供的一个或多个初始化参数，将输入数据部分a、b、c和d寻址到物理存储器325。

[0057] 图4图示了示例性输入数据(未示出)的示例性逻辑数据映射400。示例性逻辑数据映射400包括第一行410(利用对角线标记图示)和第二行420(由虚线图示)。每个映射行可以包括若干滑动窗口(例如，针对第一行410的430、440和450以及针对第二行420的460、470和480)。附加地，如图所示，逻辑数据映射400示出了滑动窗口跨越输入数据的数据维度边界(例如，跨越第一行410和第二行420)的能力。由于可以更有效地准备更多数据以用于协作神经网络处理组件(例如，图2的205)的后续处理，这样的能力允许提高性能。

[0058] 图5类似于图4，并且呈现图5来描述本文所描述的系统和方法允许使用填充来进一步增强示例性神经网络环境(例如，图1的100和图2的200)的性能特性的能力。如图所示，(未示出的示例性输入数据的)逻辑数据映射500可以包括跨越一个或多个行(例如，510和520)的各种滑动窗口(530、540、550、560、570和580)。附加地，逻辑数据映射500还可以包括填充580。

[0059] 在一个说明性操作中，在示例性神经网络环境(图1的100或图2的200)的运行时间，填充580可以被动态地添加。图2的操作控制器235可以指定将在图3中所示的输入数据(例如，二进制大对象)的每个维度(例如，整体称为容量340)上使用的填充量，并且神经网络环境(例如，迭代器控制器指令)可以可操作地构造数据量，就像填充物理地存在于存储器中。在添加填充的情况下，还可以在迭代器输出位置中通过示例性神经网络环境(例如，迭代器控制器指令)来生成默认值。

[0060] 图6是利用虚拟化硬件迭代器来增强神经网络环境的性能的说明性过程600的流程图。如图所示，处理开始于框605，在框605处，从神经网络计算环境的一个或多个协作组件接收一个或多个初始化参数。然后处理进行到框610，在框610处，从神经网络计算环境的一个或多个协作组件接收一个或多个数据处理命令。说明性地，示例性处理命令可以包括迭代器操作类型。在框615处，根据所选择的逻辑数据映射协议(LDMP)在所选择的相关联的存储器地址处读取输入数据，并且生成内部状态数据，内部状态数据表示读入数据的一个或多个特性。从那里，处理进行到框620，在框620处，用于关联读入数据的一个或多个部分以进行处理的一个或多个指令被生成。然后，在框625处，读入数据的一个或多个相关联的部分被通信到神经网络计算环境的一个或多个协作处理神经元。然后在框630处，内部状态被更新来指示读入数据的一个或多个部分已经被通信到神经元进行处理。然后在框635处检查被执行，以确定是否存在神经网络环境处理所需的附加输入数据。如果检查指示不存在附加数据，则处理在框640处终止。如果附加数据需要神经网络环境处理，则处理返回到框610并从那里继续。

[0061] 在一个说明性操作中，可以通过向存储器内数据提供逻辑映射来将硬件迭代器虚拟化。使用描述数据物理存储器寻址、维度和容量的参数来计算映射。附加地，关于如何遍历逻辑数据映射的指令被生成，以允许基于正在由神经网络神经元执行的操作以及数据的维度和容量描述对将由神经网络神经元准备进行处理的数据进行有效加载(即，从/到存储器)。

[0062] 图7是利用一个或多个虚拟化硬件迭代器来增强NN/DNN环境的性能的说明性过程700的流程图。如图所示，处理开始于框705，在框705处，从神经网络环境的协作组件(例如，操作控制器)接收一个或多个初始化参数，其中一个或多个初始化参数可以包括表示输入数据的维度的数据。处理然后进行到框710，在框710处，使用所接收的一个或多个初始化参数来标识输入数据的一个或多个物理地址。然后在框715处，使用所选择的逻辑数据映射协议(LDMP)读入(例如，加载)输入数据。说明性地，在框715，逻辑数据映射协议可以利用滑动窗口(如图3所描绘的)将输入数据的一个或多个部分与所接收的初始化参数相关联(例如，读入)，以将输入数据整理为一个或多个部分用于后续处理。此外，LDMP可以允许跨越的窗口和逻辑数据填充作为准备用于处理的输入数据的一部分。

[0063] 然后，处理进行到框720，在框720处，表示输入数据的经整理的一个或多个部分的内部状态被生成。从那里，一个或多个指令被生成来指示神经网络环境的协作存储器控制器组件对输入数据准备用于在框725处进行处理的一个或多个部分的存储器存储进行处理。然后处理进行到框730，在框730处，将输入数据的所准备的一个或多个部分通信到神经网络环境的一个或多个协作处理组件(例如，神经元处理器)来进行处理。

[0064] 说明性地，并且如图3所描绘的，示例性逻辑数据映射协议可操作地将二维平面文件类型数据渲染为多维表征数据，使得输入数据可以在视觉上表示为具有特定高度、宽度、深度(例如，信道数目)、深度切片(例如，内核数目)的二进制大对象。该逻辑数据映射允许使用滑动窗口来将用于迭代操作的输入数据的一个或多个数据元素相关联。说明性地，滑动窗口可以跨越两个逻辑映射的输入数据宽度行。

[0065] 然后在框735处执行检查来确定是否存在将处理的附加输入数据(即，作为迭代操作的一部分)。如果不存在附加输入数据，则处理在框740处终止。然而，如果附加输入数据需要迭代操作，则处理然后返回到框710并从那里继续。

[0066] 图8中图示的计算机架构800包括：中央处理单元802(“CPU”)、系统存储器804(包括随机存取存储器806(“RAM”)和只读存储器(“ROM”)808)以及将存储器804耦合到CPU 802的系统总线810。例如在启动期间，包括有助于在计算机架构800内的元件之间传递信息的基本例程的基本输入/输出系统可以存储在ROM 808中。计算机800还包括用于存储操作系统814、其他数据和一个或多个应用程序的大容量存储设备812。

[0067] 大容量存储设备812利用连接到总线810的大容量存储控制器(未示出)连接到CPU 802。大容量存储设备812及其相关联的计算机可读介质为计算机架构800提供非易失性存储。虽然本文中包括的计算机可读介质的描述指代大容量存储设备(例如，固态驱动、硬盘、CD-ROM驱动)，但是本领域技术人员应理解，计算机可读介质可以是可由计算机架构800访问的任何可用计算机存储介质或通信介质。

[0068] 通信介质包括计算机可读指令、数据结构、程序模块或经调制的数据信号(例如，载波或其他传输机制)中的其他数据，并且包括任何传递介质。术语“经调制的数据信号”表示以对信号中的信息进行编码的方式来改变或设置其一个或多个特性的信号。作为示例而非限制，通信介质包括诸如有线网络或直接有线连接的有线介质，以及诸如声学、射频、红外和其他无线介质的无线介质。上述任何组合也应包括在计算机可读介质的范围内。

[0069] 作为示例而非限制，计算机存储介质可包括以用于存储诸如计算机可读指令、数据结构、程序模块或其他数据的信息的任何方法或技术实现的易失性和非易失性、可移动

和不可移动介质。例如，计算机存储介质包括但不限于RAM、ROM、EPROM、EEPROM、闪存或其他固态存储器技术、CD-ROM、数字通用盘(“DVD”)、HD-DVD、BLU-RAY、或其他光学存储装置、磁带盒、磁带、磁盘存储装置或其他磁存储设备或可用于存储期望信息并且可由计算机架构800访问的任何其他介质。处于保护的目的，短语“计算机存储介质”、“计算机可读存储介质”及其变型不包括波、信号和/或其他暂时性和/或无形通信介质本身。

[0070] 根据各种技术，计算机架构800可以使用借助网络820和/或另一网络(未示出)到远程计算机的逻辑连接在联网环境中操作。计算机架构800可以借助连接到总线810的网络接口单元816连接到网络820。应当理解，网络接口单元816也可以用于连接到其他类型的网络和远程计算机系统。计算机架构800还可以包括输入/输出控制器818，用于接收和处理来自若干其他设备(包括键盘、鼠标或电子笔(图8中未示出))的输入。类似地，输入/输出控制器818可以向显示屏、打印机或其他类型的输出设备(在图8中也未示出)提供输出。还应当理解，经由借助网络接口单元816到网络820的连接，计算架构可以使得DNN模块105能够与计算环境100通信。

[0071] 应当理解，本文所描述的软件组件可以在被加载到CPU 802和/或DNN模块105中并被执行时，可以将CPU 802和/或DNN模块105和整个计算机架构800从通用计算设备变换为旨在促进本文所呈现的功能的专用计算系统。CPU 802和/或DNN模块105可以由任何数量的晶体管或其他分立电路元件和/或芯片组(可以单独地或共同地呈现任何数量的状态)构造。更具体地，响应于包含在本文所公开的软件模块内的可执行指令，CPU 802和/或DNN模块105可以作为有限状态机进行操作。这些计算机可执行指令可以通过指定CPU 802如何在状态之间转换来对CPU 802进行转换，从而对构成CPU 802的晶体管或其他分立硬件元件进行变换。

[0072] 对本文所呈现的软件模块进行编码还可以对本文所呈现的计算机可读介质的物理结构进行变换。在本说明书的不同实现中，物理结构的特定变换取决于各种因素。这样的因素的示例包括但不限于：用于实现计算机可读介质的技术、计算机可读介质是否被表征为主存储装置或辅助存储装置等。例如，如果计算机可读介质被实现为基于半导体的存储器，则可以通过对半导体存储器的物理状态进行变换而将本文所公开的软件编码在计算机可读介质上。例如，软件可以对构成半导体存储器的晶体管、电容器或其他分立电路元件的状态进行变换。软件还可以对这样的组件的物理状态进行变换，以在其上存储数据。

[0073] 作为另一示例，本文所公开的计算机可读介质可以使用磁或光技术来实现。在这样的实现中，当在其中对软件进行编码时，本文所呈现的软件可以对磁或光介质的物理状态进行变换。这些变换可以包括改变给定磁介质内特定位置的磁特性。这些变换还可以包括改变给定光介质内特定位置的物理特征或特性，以改变那些位置的光特性。在不脱离本说明书的范围和精神的情况下，前述示例仅用于促进该讨论，物理介质的其他变换是可能的。

[0074] 鉴于以上所述，应当理解，在计算机架构800中发生许多类型的物理变换，以存储并执行本文所呈现的软件组件。还应理解，计算机架构800可以包括其他类型的计算设备(包括手持式计算机、嵌入式计算机系统、个人数字助手以及本领域技术人员已知的其他类型的计算设备)。还预期计算机架构800可以不包括图8中所示的所有组件、可以包括未在图8中明确示出的其他组件或可以使用与图8中所示的架构完全不同的架构。

[0075] 如上所述，计算系统800可以被部署为计算机网络的一部分。通常，以上对计算环境的描述适用于在网络环境中部署的服务器计算机和客户端计算机。

[0076] 图9图示了具有服务器的示例性说明性联网计算环境900，服务器经由通信网络与客户端计算机通信，其中可以采用本文所描述的装置和方法。如图9所示，(多个)服务器905可以经由通信网络820(可以是固定有线或无线LAN、WAN、内联网、外联网、对等网络、虚拟专用网络、因特网、蓝牙通信网络、专有低压通信网络或其他通信网络中的任一个或组合)与若干客户端计算设备(例如，平板个人计算机910、移动电话915、电话920、(多个)个人计算机801、个人数字助理925、智能电话手表/个人目标跟踪器(例如，苹果手表、三星、FitBit等)930和智能手机935)互连。在通信网络820是因特网的网络环境中，例如，(多个)服务器905可以是可操作以经由若干已知协议(例如，超文本传输协议(HTTP)、文件传输协议(FTP)、简单对象访问协议(SOAP)或无线应用协议(WAP))中的任一个来处理数据并从客户端计算环境801、910、915、920、925、930和935通信数据且向客户端计算环境801、910、915、920、925、930和935通信数据的专用计算环境服务器。附加地，联网计算环境900可以利用各种数据安全协议(例如，安全套接层协议(“SSL”)或加密软体协议(“PGP”))。客户端计算环境801、810、815、820、825、830和835中的每一个可以配备有计算环境805，计算环境805可操作来支持一个或多个计算应用程序或终端会话(例如，web浏览器(未示出)或其他图形用户界面(未示出)或移动桌面环境(未示出))，以获得对服务器计算环境(一个或多个)905的访问。

[0077] (多个)服务器905可以通信地耦合到其他计算环境(未示出)并且接收关于参与用户的交互/资源网络的数据。在说明性操作中，用户(未示出)可以与在客户端计算环境上运行的计算应用程序交互来获得期望的数据和/或计算应用程序。数据和/或计算应用可以存储在服务器计算环境(一个或多个)905上，并借助客户端计算环境901、910、915、920、925、930和935通过示例性通信网络820通信到协作用户。参与用户(未示出)可以请求访问整体地或部分地容纳在(多个)服务器计算环境905上的特定数据和应用程序。这些数据可以在客户端计算环境801、910、915、920、925、930、935和(多个)服务器计算环境905之间通信，以进行处理和存储。(多个)服务器计算环境905可以托管用于数据和应用程序的生成、认证、加密和通信的计算应用程序、进程和小应用程序，并且可以与其他服务器计算环境(未示出)、第三方服务提供商(未示出)、网络附加存储(NAS)和存储区域网络(SAN)协作来实现应用程序/数据交易。

[0078] 示例条款

[0079] 鉴于以下条款，可以考虑本文所呈现的公开内容。

[0080] 示例条款A，用于在神经网络环境中使用一个或多个虚拟化硬件迭代器来增强数据处理的系统，该系统包括：至少一个处理器；以及与至少一个处理器通信的至少一个存储器(210)，该至少一个存储器(210)具有存储在其上的计算机可读指令，当计算机可读指令由至少一个处理器执行时，使得至少一个处理器：从神经网络环境的协作控制器组件接收一个或多个初始化参数，初始化参数包括表示将由所述神经网络环境处理的数据的维度的数据；从神经网络环境的协作存储器组件加载数据；从神经网络环境的协作控制器组件接收一个或多个指令，以根据一个或多个迭代器操作类型和一个或多个初始化参数来遍历所加载的数据；以及将所遍历的数据作为一个或多个部分通信到神经网络环境的一个或多个

处理组件。

[0081] 示例条款B,根据示例条款A所述的系统,其中计算机可读指令还使得至少一个处理器将所遍历的数据分成相等的部分,以通信到一个或多个处理组件。

[0082] 示例条款C,根据示例条款A和B所述的系统,其中计算机可读指令还使得至少一个处理器根据由所接收的一个或多个初始化参数限定的选择数据量来遍历数据,包括数据的所述参数包括数据高度、数据宽度、信道数目和内核数目。

[0083] 示例条款D,根据示例条款A至C所述的系统,其中计算机可读指令还所述至少一个处理器利用一个或多个滑动窗口来遍历所述数据,窗口可操作来选择数据量的一个或多个数据元素,作为通信到所述一个或多个处理组件的所述一个或多个部分。

[0084] 示例条款E,根据示例条款A至D所述的系统,其中计算机可读指令还使得至少一个处理器使用跨越数据量的两个或更多个行宽度的一个或多个滑动窗口来遍历所加载的数据。

[0085] 示例条款F,根据示例条款A至E所述的系统,其中计算机可读指令还使得至少一个处理器将一个或多个数据填充插入到所加载的数据。

[0086] 示例条款G,根据示例条款A至F所述的系统,其中计算机可读指令还使得至少一个处理器生成表示由一个或多个神经处理器处理的数据的数据输出量,根据由所述神经网络环境的协作控制器组件接收的一个或多个指令来生成输出量。

[0087] 示例条款H,计算机实现的方法,包括:从神经网络环境的协作控制器组件接收一个或多个初始化参数,初始化参数包括表示数据维度的数据作为数据量,数据维度包括高度、宽度、信道数目和内核数目中的任一个;从神经网络环境的协作存储器组件加载数据;从神经网络环境的协作控制器组件接收一个或多个指令,以根据一个或多个迭代器操作类型和一个或多个初始化参数来遍历所加载的数据,加载数据的遍历利用在数据量上操作的一个或多个滑动窗口;以及将所遍历的数据作为一个或多个部分通信到所述神经网络环境的一个或多个处理组件。

[0088] 示例条款I,根据条款H所述的计算机实现的方法,还包括:遍历所加载的数据来生成所遍历的数据的相等部分,用于与一个或多个处理组件进行通信。

[0089] 示例条款J,根据条款H和I所述的计算机实现的方法,其中滑动窗口可操作,以跨越所述数据量的两个或更多个宽度行。

[0090] 示例条款K,根据条款H至J所述的计算机实现的方法,还包括:将填充子容量插入到所加载的数据中。

[0091] 示例条款L,根据条款H至K所述的计算机实现的方法,其中填充子容量由从协作控制器组件接收的一个或多个指令以及由所接收的一个或多个初始化参数来限定。

[0092] 示例条款M,根据条款H至L所述的计算机实现的方法,还包括:针对由神经环境的一个或多个处理组件处理的数据来生成并输出数据量,输出数据量由从协作控制器组件接收的一个或多个指令限定。

[0093] 示例条款N,根据条款H至M所述的计算机实现的方法,还包括:丢弃所生成的输出数据量的一个或多个部分。

[0094] 示例条款O,根据条款H至N所述的计算机实现的方法,将所加载的数据的遍历位置作为内部状态数据存储在所述神经网络环境的协作存储器组件中。

[0095] 示例条款P,计算机可读存储介质,具有存储在其上的计算机可执行指令,当计算机可执行指令由计算设备的一个或多个处理器执行时,使得计算设备的一个或多个处理器:从神经网络环境的协作控制器组件接收一个或多个初始化参数,一个或多个初始化参数包括表示将由所述神经网络环境处理的数据的维度的数据,数据的所述维度包括表示数据量的数据,一个或多个初始化参数包括待处理的数据的物理存储器寻址;利用一个或多个初始化参数,从神经网络环境的协作存储器组件加载数据,协作存储器组件可操作地耦合到神经网络环境的一个或多个处理组件;从神经网络环境的协作控制器组件接收一个或多个指令,以根据一个或多个迭代器操作类型和所述一个或多个初始化参数遍历所加载的数据,指令包括存储表示遍历所加载的数据的所述内部状态的数据的一个或多个指令;以及将所遍历的数据作为一个或多个部分通信到所述神经网络环境的所述一个或多个处理组件。

[0096] 示例条款Q,根据条款P所述的计算机可读存储介质,其中指令还使得计算设备的一个或多个处理器:将内部状态数据存储在神经网络环境的协作存储器组件中。

[0097] 示例条款R,根据条款P和Q所述的计算机可读存储介质,其中指令还使得计算设备的一个或多个处理器将附加数据量插入到所加载的数据。

[0098] 示例条款S,根据条款P至R所述的计算机可读存储介质,其中指令还使得计算设备的一个或多个处理器生成表示由神经网络环境的一个或多个处理组件处理的数据的输出数据量。

[0099] 示例条款T,根据条款P至S所述的计算机可读存储介质,其中指令还使得计算设备的一个或多个处理器:利用所加载的数据的逻辑数据映射来遍历所加载的数据,遍历所加载的数据包括将一个或多个滑动窗口应用于所述逻辑数据映射,以将所加载的数据的一部分与一个或多个物理存储器地址相关联。

[0100] 结论

[0101] 最后,尽管以结构特征和/或方法动作专用的语言描述了各种技术,但是应理解,所附表示中限定的主题不必限于所描述的特定特征或动作。相反,公开了特定特征和动作作为实现所要求保护的主题的示例形式。

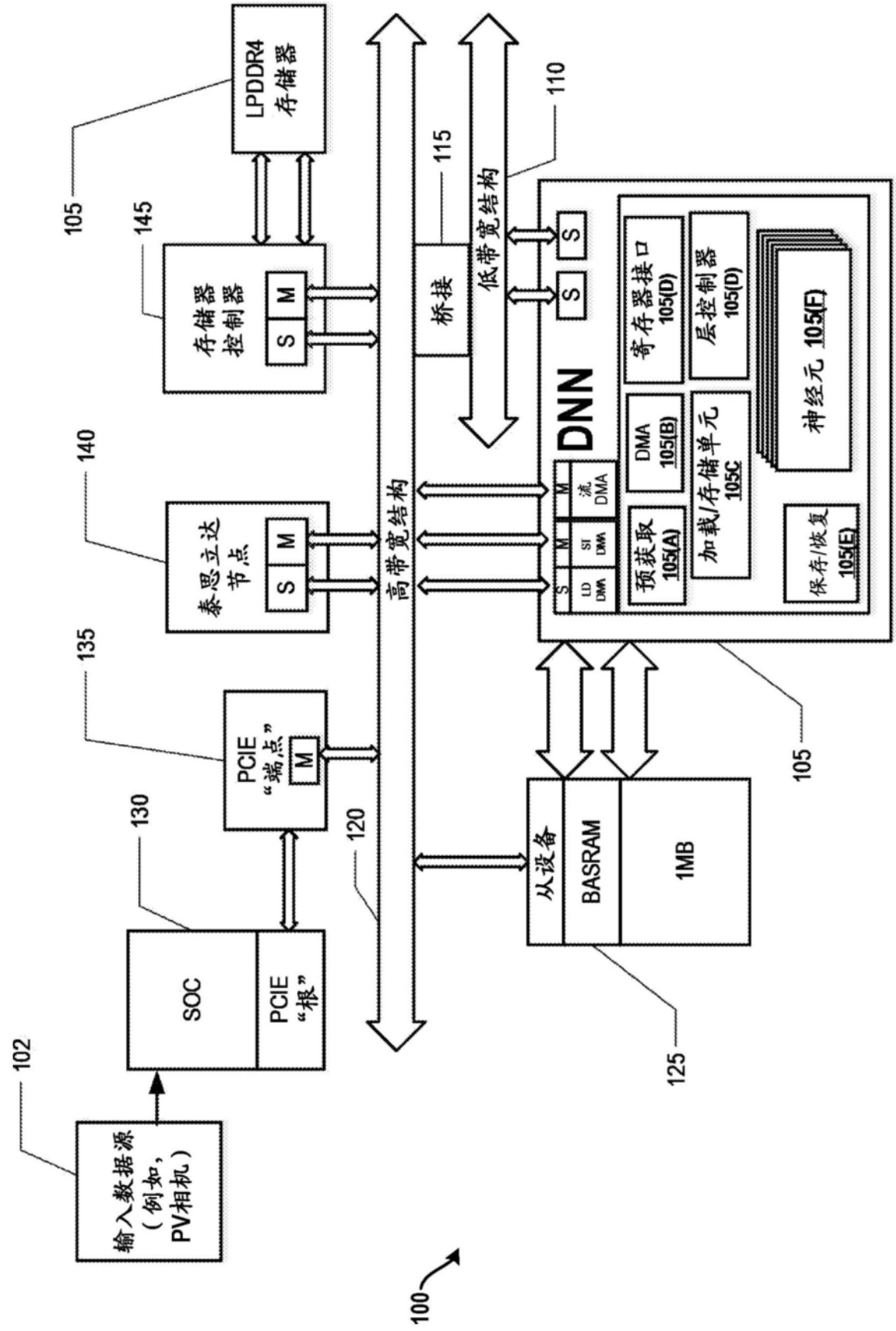

LPDDR4
存储器
105
145
存储器
控制器
S
M
S
M
115
桥接
低带宽结构
110
S
S
DNN
寄存器接口
105(D)
层控制器
105(D)
神经元 105(F)
DMA
105(B)
加载/存储单元
105C
预获取
105(A)
保存/恢复
105(E)
M
流
DMA
M
SI
DMA
S
LD
DMA
140
泰思立达
节点
S
M
高带宽结构
135
PCIE
“端点”
M
130
SOC
PCIE
“根”
120
从设备
BASRAM
1MB
125
105
102
输入数据源
（例如，
PV相机）
100

图1

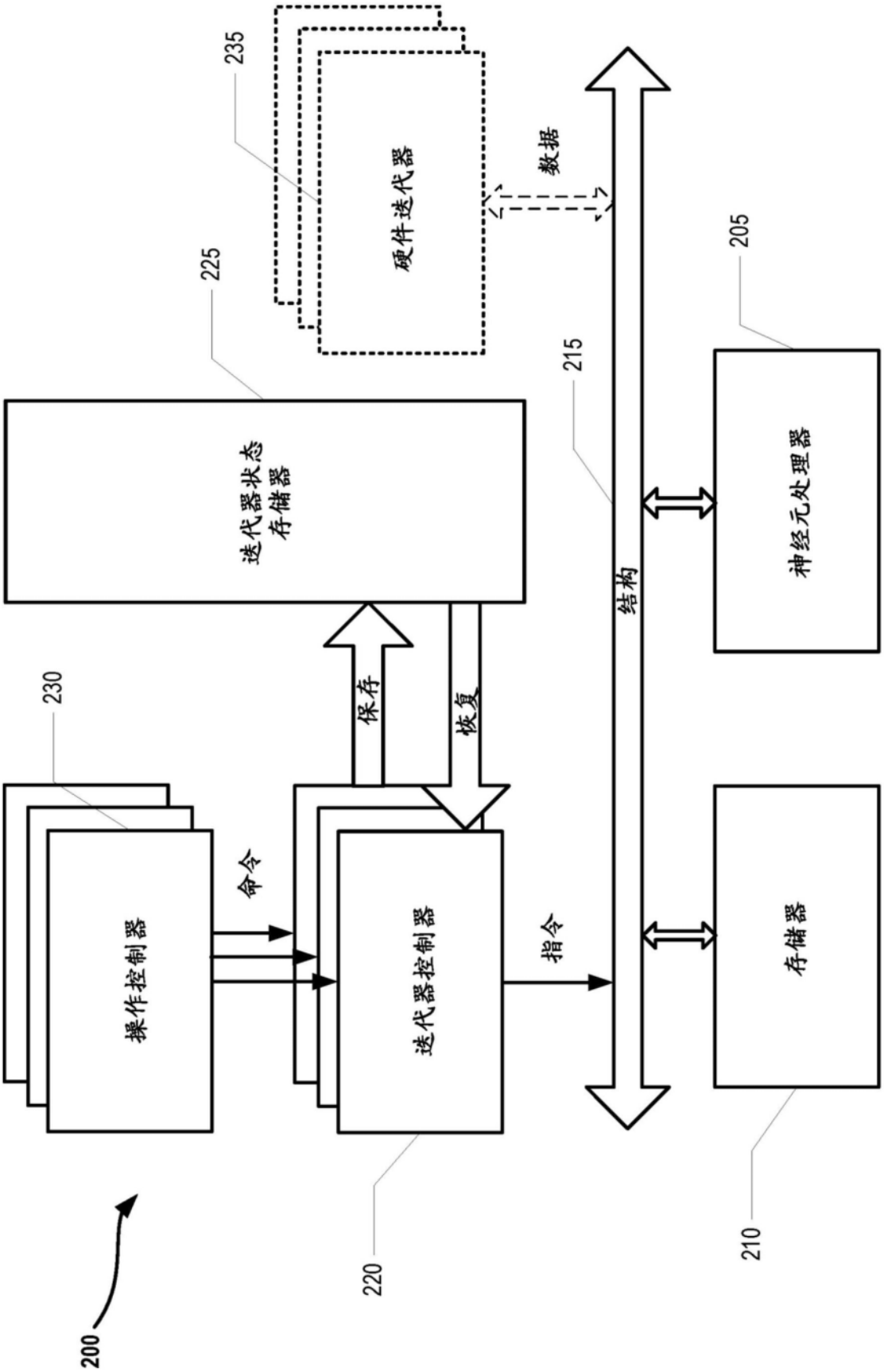

235
硬件迭代器
数据
225
迭代器状态存储器
205
神经元处理器
215
结构
230
保存
恢复
操作控制器
命令
迭代器控制器
指令
存储器
220
210
200

图2

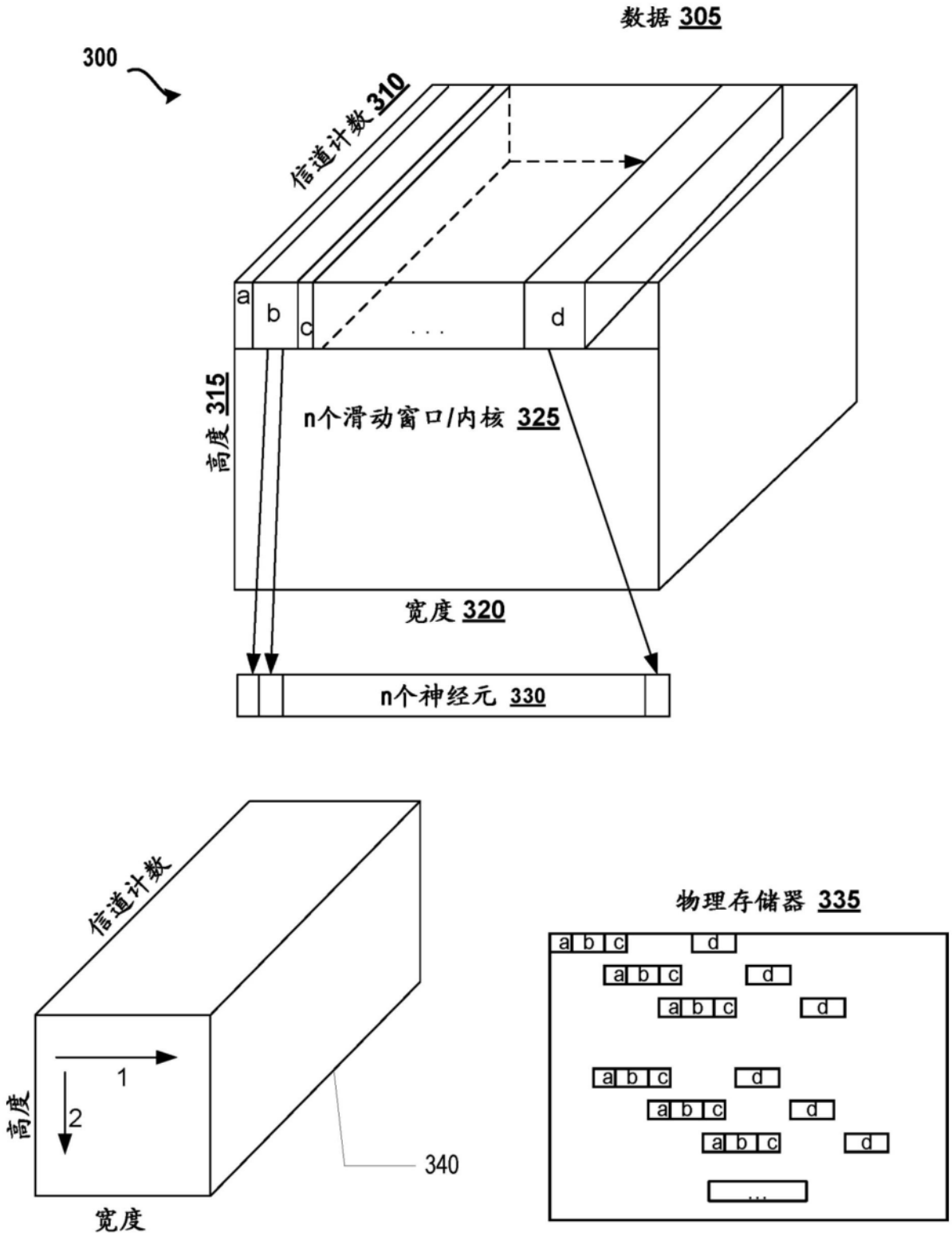

数据 305
300
信道计数 310
a
b
c
...
d
高度 315
n个滑动窗口/内核 325
宽度 320
n个神经元 330
信道计数
1
2
高度
宽度
340
物理存储器 335
a b c
d
...

图3

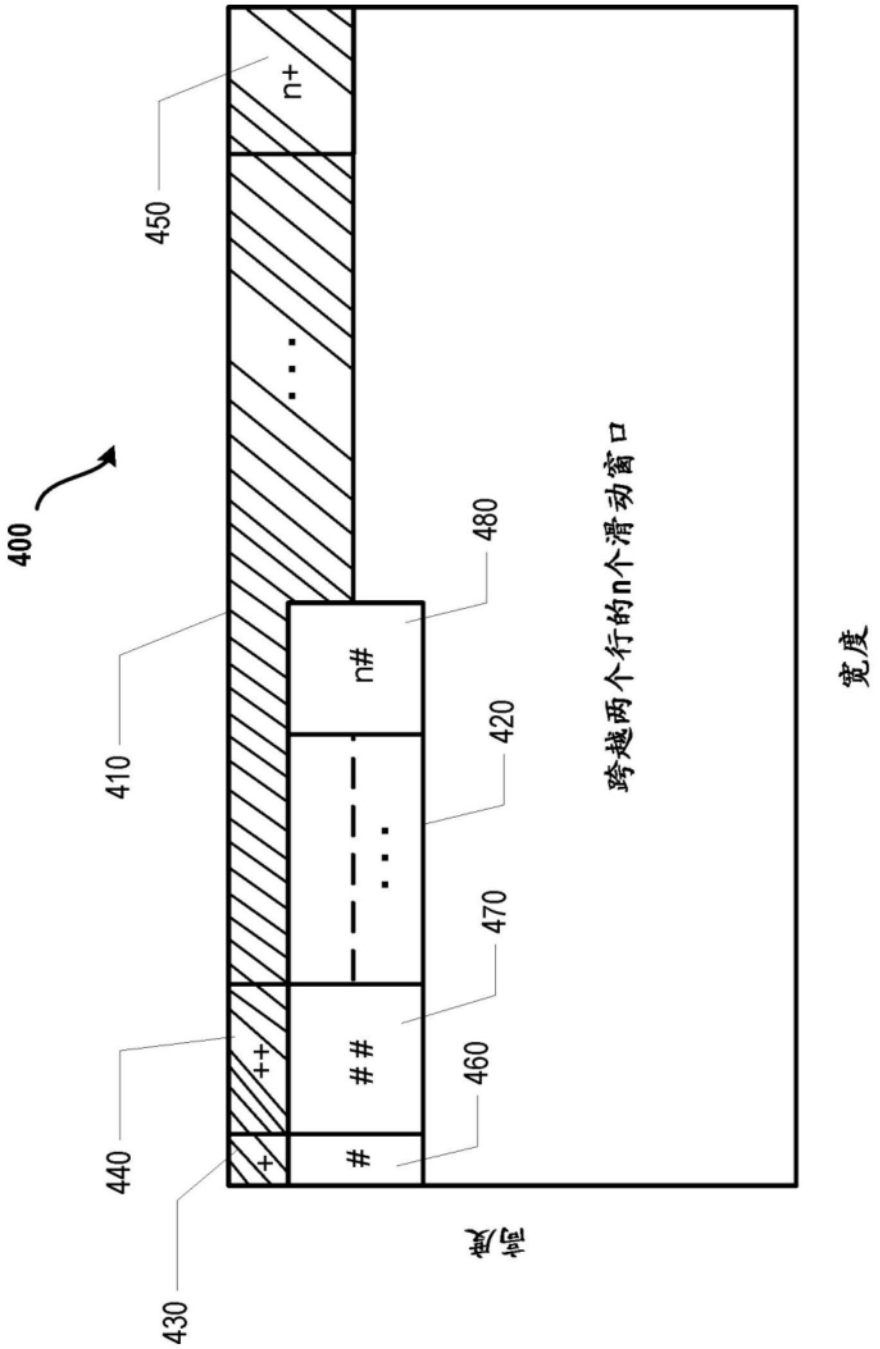


图4

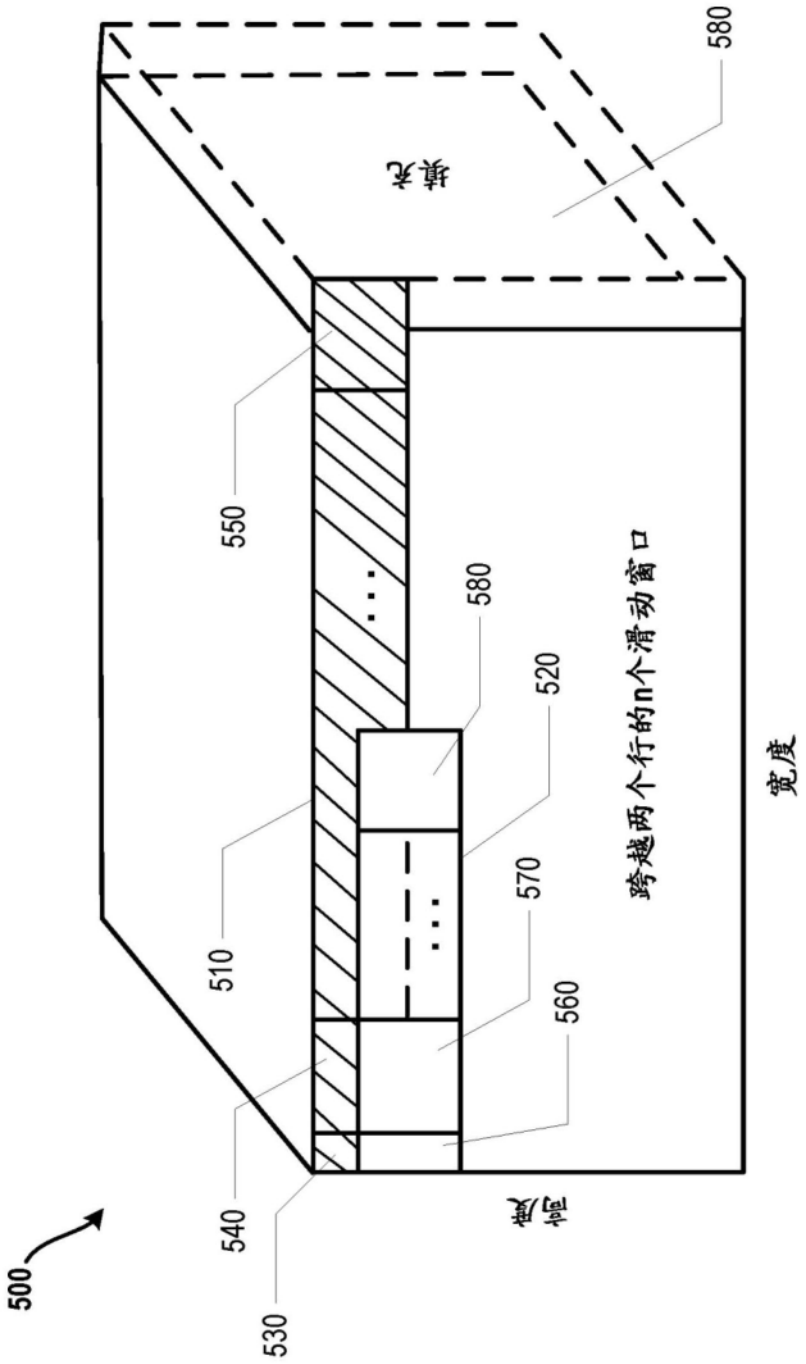


图5

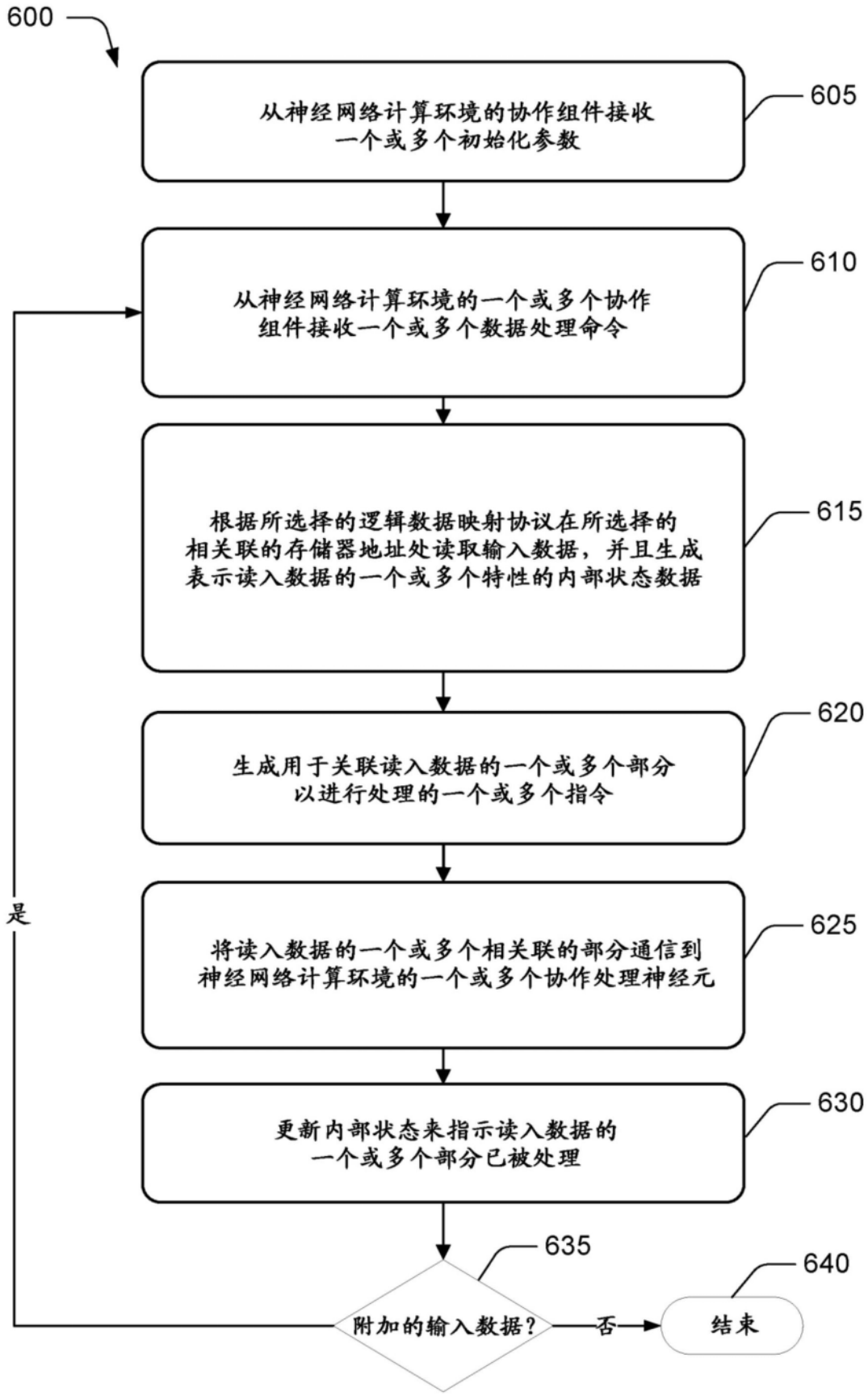

600
从神经网络计算环境的协作组件接收一个或多个初始化参数
605
从神经网络计算环境的一个或多个协作组件接收一个或多个数据处理命令
610
根据所选择的逻辑数据映射协议在所选择的相关联的存储器地址处读取输入数据，并且生成表示读入数据的一个或多个特性的内部状态数据
615
生成用于关联读入数据的一个或多个部分以进行处理的一个或多个指令
620
将读入数据的一个或多个相关联的部分通信到神经网络计算环境的一个或多个协作处理神经元
625
更新内部状态来指示读入数据的一个或多个部分已被处理
630
635
附加的输入数据?
是
否
640
结束

图6

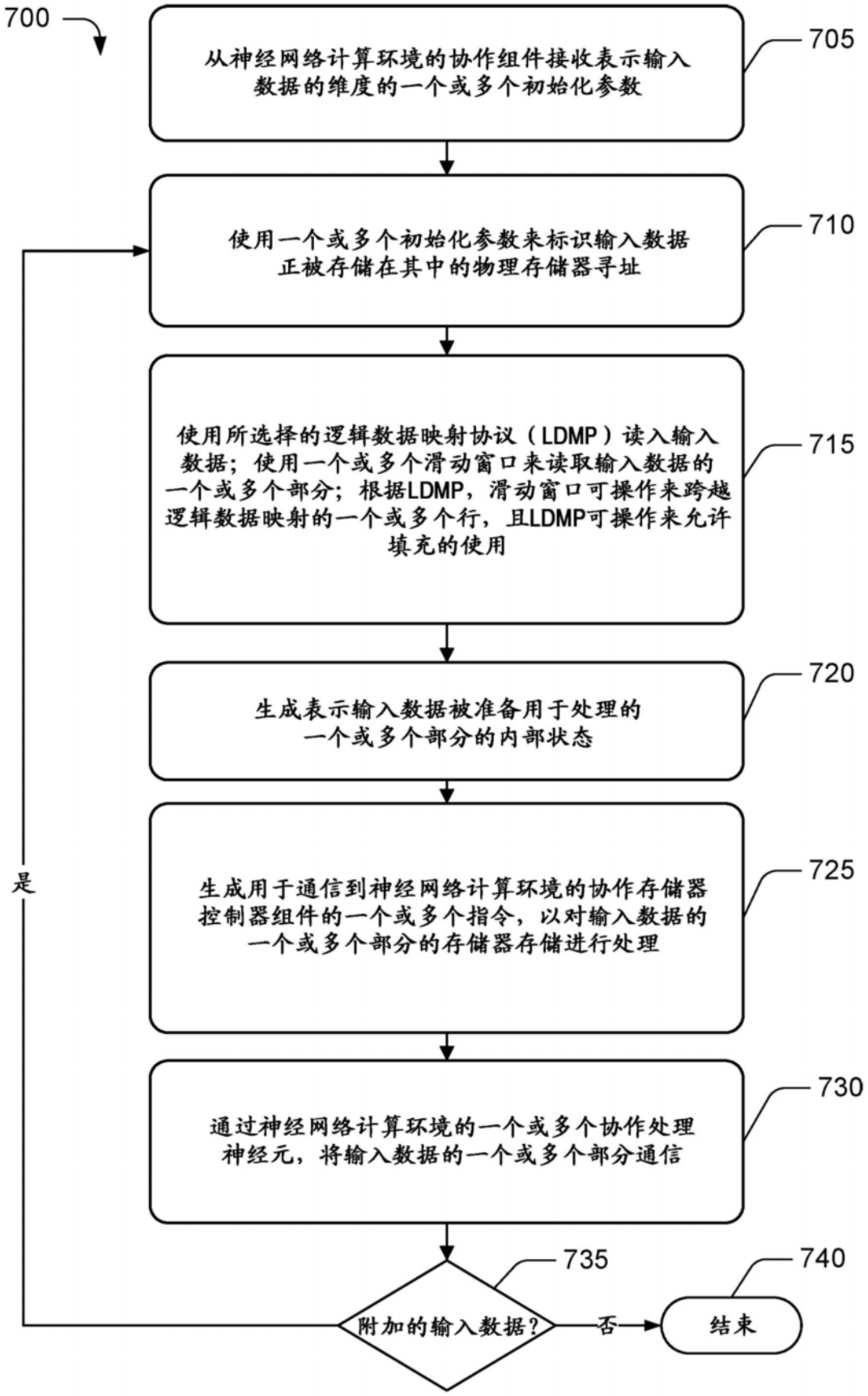

700
从神经网络计算环境的协作组件接收表示输入数据的维度的一个或多个初始化参数
705
使用一个或多个初始化参数来标识输入数据正被存储在其中的物理存储器寻址
710
使用所选择的逻辑数据映射协议（LDMP）读入输入数据；使用一个或多个滑动窗口来读取输入数据的一个或多个部分；根据LDMP，滑动窗口可操作来跨越逻辑数据映射的一个或多个行，且LDMP可操作来允许填充的使用
715
生成表示输入数据被准备用于处理的一个或多个部分的内部状态
720
生成用于通信到神经网络计算环境的协作存储器控制器组件的一个或多个指令，以对输入数据的一个或多个部分的存储器存储进行处理
725
通过神经网络计算环境的一个或多个协作处理神经元，将输入数据的一个或多个部分通信
730
附加的输入数据?
735
是
否
结束
740

图7

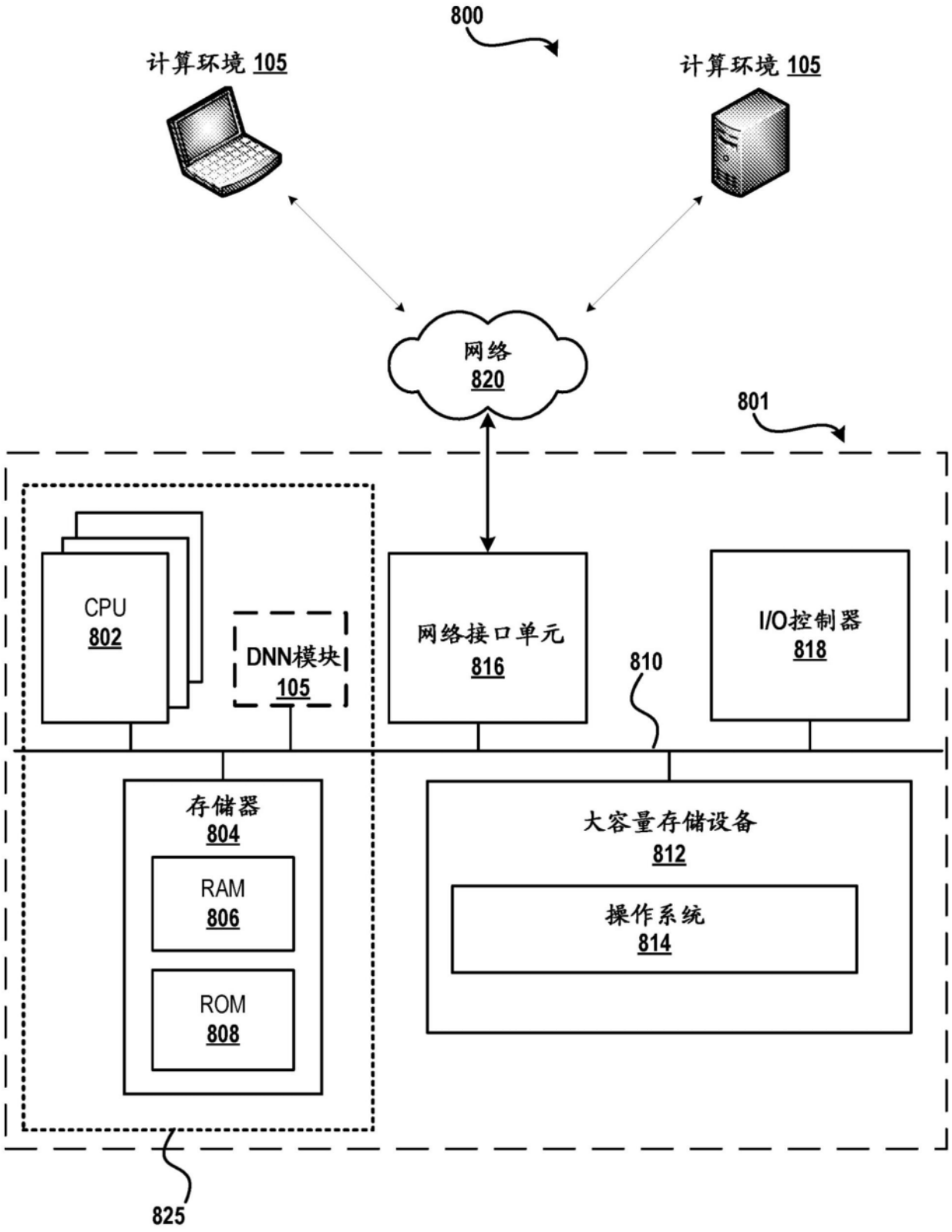

800
计算环境 105
计算环境 105
网络
820
801
CPU
802
DNN模块
105
网络接口单元
816
I/O控制器
818
810
存储器
804
RAM
806
ROM
808
大容量存储设备
812
操作系统
814
825

图8

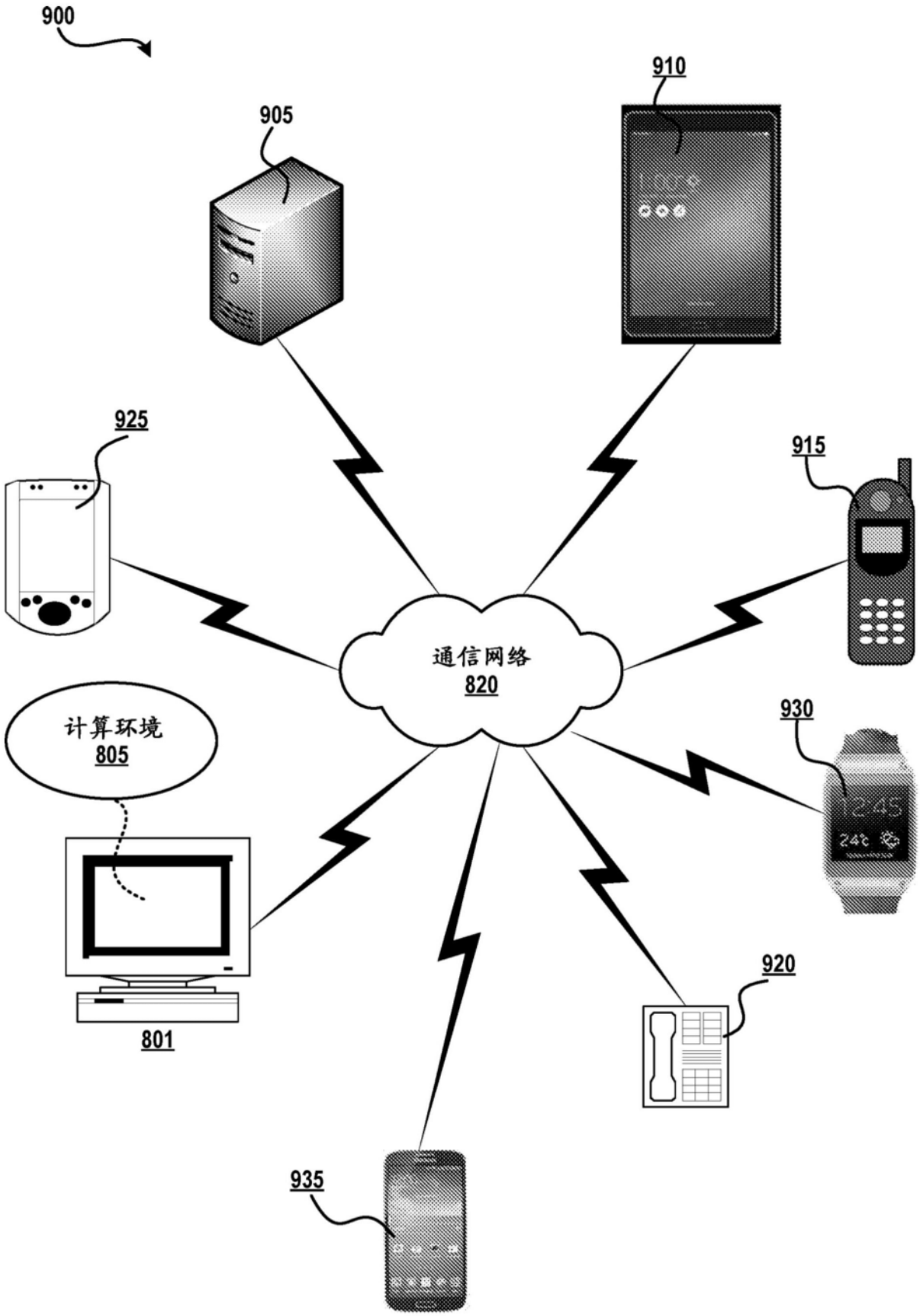


图9