(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 110520846 A
(43)申请公布日 2019.11.29

(21)申请号 201880025417.X

(22)申请日 2018.04.06

(30)优先权数据
62/486,432 2017.04.17 US
15/719,351 2017.09.28 US

(85)PCT国际申请进入国家阶段日
2019.10.16

(86)PCT国际申请的申请数据
PCT/US2018/026354 2018.04.06

(87)PCT国际申请的公布数据
WO2018/194847 EN 2018.10.25

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

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 辛鸣

(51)Int.Cl.
G06F 9/50(2006.01)

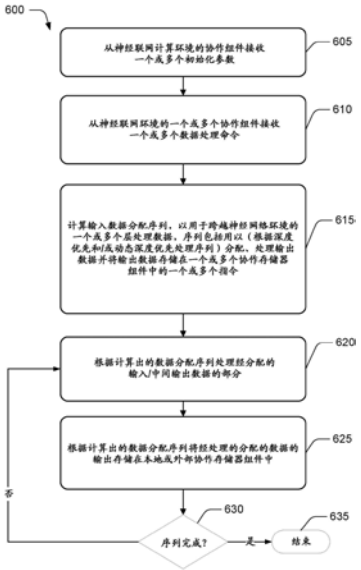
权利要求书2页 说明书15页 附图10页

(54)发明名称

对用于优化存储器利用和神经网络的性能的数据分区的动态排序

(57)摘要

经优化的存储器使用和管理对于神经网络(NN)或深度神经网络(DNN)计算环境的整体性能至关重要。使用输入数据维度的各种特性,分配序列针对将由NN或DNN处理的输入数据而被计算,分配序列优化本地和外部存储器组件的有效使用。分配序列可以描述如何将输入数据(及其相关联的处理参数(例如,处理权重))拆分为一个或多个部分以及输入数据的这样的部分(及其相关联的处理参数)如何在本地存储器、外部存储器和NN或DNN的处理单元组件之间被传递。附加地,分配序列可以包括将所生成的输出数据存储在本地和/或外部存储器组件中以优化本地和/或外部存储器组件的有效使用的指令。



CN 110520846 A

1.一种用于神经网络环境中的增强数据处理的系统,所述系统包括:

至少一个处理器;

至少一个第一存储器组件;以及

与所述至少一个处理器通信的至少一个第二存储器组件,所述至少一个第一存储器组件和/或所述至少一个第二存储器组件具有存储在其上的计算机可读指令,所述计算机可读指令当由所述至少一个处理器执行时,使得所述至少一个处理器:

从所述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,所述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;

计算针对所述数据的分配序列,所述分配序列包括用以在所述至少一个第一存储器组件和/或所述至少一个第二存储器组件中存储输出数据的一个或多个指令,所述分配序列包括仅广度处理序列、深度优先处理序列和动态深度优先处理序列;

从所述神经网络环境的所述协作存储器组件加载数据;

从所述神经网络环境的所述协作控制器组件接收用以根据计算出的所述分配序列处理所述数据的所选择的部分的一个或多个指令;

由一个或多个协作处理单元处理数据的所述部分以生成所述输出数据,以用于存储在所述至少一个第一存储器组件或所述至少一个第二存储器组件上;以及

将所生成的所述输出数据存储在所述至少一个第一存储器组件或所述至少一个第二存储器组件上。

2.根据权利要求1所述的系统,其中计算出的所述分配序列基于所述神经网络环境的层的数目。

3.根据权利要求2所述的系统,其中所述计算机可读指令还使得所述至少一个处理器:加载表示一个或多个层权重的数据,以用于由所述神经网络环境的一个或多个处理单元使用以生成所述输出数据。

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至少一个处理器将所生成的所述输出数据存储在所述至少一个第一存储器组件中。

5.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至少一个处理器将所生成的所述输出数据存储在所述至少一个第二存储器组件中。

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至少一个处理器基于所述至少一个第一存储器组件的大小和/或所述至少一个第二存储器组件的大小来计算所述分配序列。

7.根据权利要求6所述的系统,其中所述计算机可读指令还使得所述至少一个处理器将所生成的所述输出数据存储在所述至少一个第一存储器组件中。

8.根据权利要求6所述的系统,其中所述计算机可读指令还使得所述至少一个处理器将所生成的所述输出数据存储在所述至少一个第二存储器组件中。

9.一种计算机实现的方法,包括:

从协作控制器组件接收一个或多个初始化参数,所述初始化参数包括表示输入数据的维度的数据以及表示与所述输入数据相关联的处理层的数量的数据;

计算针对所述输入数据的分配序列,所述分配序列包括用以将输出数据存储在至少一个第一存储器组件和/或至少一个第二存储器组件中的一个或多个指令,并且包括用以加

载处理所述输入数据所需的相关联数据的一个或多个指令；

从协作存储器组件加载表示一个或多个处理权重的数据；

从所述协作控制器组件接收用以根据计算出的所述分配序列处理所述数据的所选择的部分的一个或多个指令，所述分配序列包括仅广度处理序列、深度优先处理序列和动态深度优先处理序列；

由一个或多个协作处理单元根据计算出的所述分配序列处理数据的所述部分以生成输出数据，以用于存储在所述至少一个第一存储器组件或所述至少一个第二存储器组件上；以及

将所生成的所述输出数据存储在所述至少一个第一存储器组件或所述至少一个第二存储器组件上。

10.根据权利要求9所述的计算机实现的方法，还包括：接收表示所述至少一个第一存储器组件的大小和所述至少一个第二存储器组件的大小的数据，以用于在计算所述分配序列时使用。

11.根据权利要求10所述的计算机实现的方法，还包括：将所生成的所述输出数据存储在包括第一存储器组件和第二存储器组件中的任何存储器组件的至少一个存储器组件上。

12.根据权利要求9所述的计算机实现的方法，还包括：将所生成的所述输出数据的一部分存储在所述至少一个第一存储器组件上。

13.根据权利要求9所述的计算机实现的方法，还包括：加载与正沿计算出的所述分配序列而被处理的数据的所述部分相关联的处理权重数据。

14.根据权利要求9所述的计算机实现的方法，还包括：针对每个处理层，加载与整个所述输入数据相关联的处理权重数据。

15.根据权利要求9所述的计算机实现的方法，还包括：将先前处理层的所述输出数据作为后续处理层的输入数据处理。

对用于优化存储器利用和神经网络的性能的数据分区的动态排序

背景技术

[0001] 神经网络(NN)或深度神经网络(DNN)处理通常要求运行时间在网络的多个层上发生,其中激活在每个层处被计算。第一层的输出充当后续层的输入。例如,第一NN/DNN层的经处理的激活充当第二NN/DNN层的输入。然后,第二层的经处理的激活充当第三层的输入。经处理的激活继续发生,直到到达网络的最后一层,其中最后一层的输出被NN/DNN计算环境用于呈现或存储。

[0002] 当前部署的NN/DNN计算环境(具有诸如本地存储器(即,本地高速缓存)的有限资源)通常将中间层激活存储到主存储器。在操作上,即,当需要数据来处理下一层时,正在处理的数据从本地存储器/处理单元传送到主存储器,然后返回到本地存储器/处理单元。这样的实践通常是低效的并且需要经历可避免的处理周期以及关键存储器管理资源的使用,这导致NN/DNN计算环境的时延和加压的性能。

[0003] 此外,由于所有中间激活通常都被存储到主存储器,使用小的本地存储器作为其输入/激活和权重的分级区域的NNN/DNN计算环境可能是低效的。例如,如果给定层的输入和权重的大小大于本地存储器的大小,则在不复写正在处理的输入和权重的情况下,当前层的输出不能存储到本地存储器。在操作上,该给定层的输出将需要存储到主存储器中。如果网络中的下一层是消耗该数据的层(大部分时间都是这种情况),则系统通过将数据存储到主存储器然后将其复制回本地分级存储器用于网络的下一层所需的处理来利用双倍带宽。

[0004] 更有利的NN/DNN架构/数据管理方案将本地数据(即,处理单元的本地)的使用最大化并将主存储器的数据读取/写入最小化,这将导致在处理速度和功耗方面的净益处。

[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] 本文描述的技术提供了使用“深度优先”和/或动态“深度优先”方法在示例性神经网络(NN)和/或深度神经网络(DNN)环境中利用来进行数据处理,其中“深度优先”和/或“动态深度优先”处理协议(例如,表示为由示例性NN和/或DNN环境的控制器组件提供的一个或多个指令)可操作地计算并执行允许数据处理的分配序列,分配序列改进整体性能并优化存储器管理。在其他例示性实现中,数据分配序列可以由示例性神经网络(NN)和/或深度神经网络(DNN)环境的其他协作组件(包括但不限于在线或离线编译器以及其他相关联组件)来计算。

[0007] 在一个例示性实现中,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)、操作控制器和DNN模块。在例示性实现中,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、负载控制器和存储控制器。

[0008] 在一个例示性操作中，NN/DNN环境的操作控制器可操作地处理大量数据，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)。在例示性操作中，参与用户可以指定由NN/DNN环境处理的数据的维度。例示性地，使用数据维度，NN/DNN环境中的可用处理层的数量以及表示NN/DNN环境的协作存储器组件的一个或多个特性的数据(例如，存储器大小、位置、时延、效率等)、数据分配序列可以由NN/DNN环境组件来计算，NN/DNN环境组件指定每个层的输入数据(以及任何相关联的处理参数)将被分配并在协作的NN/DNN存储器组件和NN/DNN处理器之间通信来实现最佳处理。

[0009] 在一个例示性实现中，示例性NN/DD环境可以包括本地存储器组件和外部存储器组件。在该实现中，本地存储器组件相对于外部存储器组件以降低的时延、以较高的速率可操作地传送数据。在操作上，相对于外部存储器组件，本地存储器组件可以包括存储较小量数据的存储器大小。

[0010] 在一个例示性操作中，可以接收输入数据(例如，数据blob)以供NN/DNN环境处理，NN/DNN环境具有特定限定的数据维度、相关联的数据处理参数(例如，层权重)、进行处理所需的限定数量的处理层、以及表示NN/DNN环境的协作存储器组件的一个或多个特性的数据。在操作上，操作控制器提供使用所接收的数据维度、协作存储器特性和层的数目来计算数据分配序列的指令。计算出的数据分配序列生成表示在每个可用处理层上拆分输入数据的部分数量的数据以及从外部存储器组件将数据部分(及其相关联的处理参数)加载到内部存储器组件、到NN/DNN环境的可用处理单元的定时。

[0011] 附加地，计算出的数据分配序列可以包括用于操作控制器根据计算出的序列，将数据部分(一个或多个)从外部存储器通信到本地存储器、到可用处理单元(一个或多个)的指令。例示性地，可以在处理序列中利用本地存储器组件来存储用于每个处理层处理的数据部分的输出数据。在例示性操作中，在处理给定处理层的所有部分并将这样的输出部分数据存储在本地存储器中时，协作处理单元可以将针对给定处理层所生成的输出部分数据(存储在本地存储器中)聚合以生成针对给定处理层的完整输出数据。在一个例示性操作中，然后可以将所生成的完整层输出数据存储在外部存储器组件中，这导致使得更多本地存储器组件的存储器可用于后续层处理。

[0012] 本文所呈现的技术提供可以将本地数据(即，处理单元(一个或多个)本地)的使用最大化并将主存储器的数据读取/写入最小化的有利的NN/DNN架构/数据管理方案，这导致在处理速度和功耗方面的净益处。

[0013] 应当理解，尽管关于系统进行了描述，但是上述主题还可以被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。

[0014] 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

[0015] 参考附图描述了具体实施方式。在附图中，附图标记的最左边数字(一个或多个)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。

[0016] 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环境的框图。

[0017] 图2图示了根据本文所描述的系统和方法部署数据分配序列的示例性神经网络环境的框图。

[0018] 图3图示了根据本文所描述的系统和方法的例示性逻辑数据映射中表示的示例性输入数据的框图。

[0019] 图4图示了根据本文所描述的系统和方法的、根据例示性分配序列处理的示例性输入数据的示例性块序列图的框图。

[0020] 图5A图示了根据本文所描述的系统和方法的、根据另一例示性分配序列处理的示例性输入数据的示例性逐步处理序列的块序列图。

[0021] 图5B图示了根据本文所描述的系统和方法的各种处理序列的示例性处理序列。

[0022] 图6是允许在示例性神经网络环境中进行数据的深度优先处理的例示性过程的流程图。

[0023] 图7是示例性神经网络环境的层的例示性深度优先处理的流程图。

[0024] 图8示出了能够执行本文所描述的方法的计算机的例示性计算机架构的附加细节。

[0025] 图9示出了根据本文所描述的系统和方法协作的例示性计算设备的附加细节。

具体实施方式

[0026] 以下具体实施方式描述了用于对示例性神经网络(NN)和/或深度神经网络(DNN)环境中使用的处理和存储器资源进行优化的技术。通常，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)可操作地允许处理数据，这改进了整体性能并优化了存储器管理。在一个例示性实现中，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。在例示性实现中，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、负载控制器和存储控制器。

[0027] 在一个例示性操作中，NN/DNN环境的操作控制器可操作地处理大量数据，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)。在例示性操作中，参与用户可以指定正在处理的数据的维度以及NN/DNN环境的配置数据处理。这样的数据可以包括可用于处理NN/DNN环境中的数据的处理层的数量以及NN/DNN环境的协作存储器组件的一个或多个操作特性。

[0028] 在一个例示性实现中，将由NN/DNN环境处理的数据可以表示为blob。通常，blob表示存储器中需要进行处理的数据。每个blob可以维持由各种维度(例如，宽度、高度、信道数、内核数和其他可用维度单位)限定的逻辑映射形状。在一个例示性操作中，NN/DNN的组

件可遍历多维blob(例如,如由逻辑数据映射限定)或这样的blob的较小的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 blob,N＝3)。在遍历blob时,NN/DNN的一个或多个组件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处理单元(一个或多个)(例如,神经元处理器)的加载指令、或用于将由处理单元(一个或多个)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-本地或外部存储器)的存储指令。在例示性操作中,操作控制器能够同时产生读取/写入多个数据的指令。

[0029] 通常,在处理神经网络(NN)中的数据时,在运行时可以存在多个处理层,其中可以在每一层处计算激活,然后将激活作为输入通信到后续层,直到到达网络的最后一层。当实现具有有限资源(例如,本地存储器(例如,本地高速缓存))的系统时,可能需要将中间层激活存储到协作的较大存储器组件(例如,主存储器、外部存储器)。当前,由于数据从本地存储器/处理单元传送到主存储器然后返回到本地存储器/处理单元,当需要数据以用于由下一层进行处理时,发生这样的数据传送消耗可避免的处理。

[0030] 本文描述的系统和方法提供了机制,借助该机制,将本地数据(即,处理单元(一个或多个)本地的数据并存储在本地存储器组件上)的使用优化,并且将需要发生的进出主存储器的数据传送最小化。这可以在提高处理速度和降低功耗方面产生净益处。

[0031] 通常,NN/DNN可操作在处理网络中的下一层之前,执行每个完整层的处理执行。此外,通常,中间层激活(即,中间层生成的输出数据)通常存储在主存储器中,直到被下一处理层消耗。

[0032] 在一个例示性实现中,可以计算分配序列,分配序列将输入数据拆分为跨越可用处理层的处理部分,输入数据包括将中间激活存储在协作本地存储器组件中并将完整输出数据存储在协作主存储器组件中的指令。可以使用输入数据维度、表示协作存储器组件的操作特性(例如,大小、速度、存储器组件的位置等)的数据、可用于处理数据的神经元处理器的数量、处理单元的时钟速度以及NN/DD环境中可用处理层的数量来计算分配序列。例示性地,计算出的分配序列提供对数据(以及相关联的处理参数-例如,层权重数据)的哪些部分进行逐步处理以及在何处逐步加载和/或存储数据(即,本地存储器组件或主存储器组件)的特定序列。

[0033] 本文所描述的技术提供了对示例性神经网络(NN)和/或深度神经网络(DNN)环境中使用的数据处理的“深度优先”和/或动态“深度优先”方法的使用,其中“深度优先”和/或“动态深度优先”处理协议(例如,表示为由示例性NN和/或DNN环境的控制器组件提供的一个或多个指令)可操作地计算并执行允许处理数据的分配序列,分配序列改进整体性能并优化存储器管理。在其他例示性实现中,数据分配序列可以由示例性神经网络(NN)和/或深度神经网络(DNN)环境的其他协作组件(包括但不限于在线或离线编译器和其他相关组件)来计算。

[0034] 在一个例示性实现中,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)、操作控制器和DNN模块)。在例示性实现中,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流式激活(DSA)、操作控制器、加载控制器和存储控制器。

[0035] 在一个例示性操作中，NN/DNN环境的操作控制器可操作地处理大量数据，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)。参与用户可以指定由NN/DNN环境正在处理的数据的维度。然后，使用所指定的数据维度以及NN/DNN环境中的可用处理层的数量、以及表示NN/DNN环境的协作存储器组件的一个或多个特性的数据(例如，存储器大小、位置、时延、效率等)以及关于NN/DNN环境的处理单元的特性，可以由NN/DNN环境组件来计算分配序列，NN/DNN环境组件指定每个层的输入数据(以及任何相关联的处理参数)将被分配以及在协作NN/DNN存储器组件和NN/DNN处理器之间进行通信来实现最佳处理。例示性地，数据分配序列可以包括仅“宽度”处理序列、“深度优先”处理序列和/或“动态深度优先”处理序列。

[0036] 例示性地，仅“宽度”处理序列描述其中从第一处理层到后续处理层顺序地处理每个层的分区的处理序列。“深度优先”处理序列描述其中以优选顺序来处理来自每个可用处理层的分区的处理序列。“动态深度”优先处理序列描述其中根据各种NN/DNN特性以及数据特性、根据仅“宽度”处理和/或“深度优先”处理的组合处理层的数目据的处理序列。当环境和/或数据特性改变时，“动态深度”处理序列可以在“仅广度”和“深度优先”处理之间可操作地跳转。

[0037] 在一个例示性实现中，示例性NN/DD环境可以包括本地存储器组件和外部存储器组件。相对于外部存储器组件，本地存储器组件以较高的速率、降低的时延可操作地传送数据。相对于外部存储器组件，本地存储器组件可以包括存储较小量数据的存储器大小。

[0038] 在一个例示性操作中，可以接收输入数据(例如，数据blob)以供NN/DNN环境处理，NN/DNN环境具有特定限定的数据维度、相关联的数据处理参数(例如，层权重)、进行处理所需的限定数量的处理层、以及表示NN/DNN环境的协作存储器组件的一个或多个特性的数据。在操作上，示例性操作控制器或其他协作NN/DNN组件提供使用所接收的数据维度、协作存储器特性和层的数目来计算数据分配序列的指令。在例示性实现中，计算出的数据分配序列生成表示在每个可用处理层上拆分输入数据的部分数量的数据并指定从外部存储器组件将数据部分(及其相关联的处理参数)加载到内部存储器组件、到NN/DNN环境的可用处理单元的定时。

[0039] 附加地，计算出的数据分配序列可以包括用于操作控制器根据计算出的序列，将数据部分(一个或多个)从外部存储器通信到本地存储器、到可用处理单元(一个或多个)的指令。例示性地，可以在处理序列中利用示例性本地存储器组件来存储用于每个层处理的数据部分的输出数据。在例示性操作中，在处理给定处理层的所有部分并将这样的输出数据存储在本地存储器中时，协作处理单元可以将针对给定处理层所生成的输出部分数据(存储在示例性本地存储器中)聚合以生成针对给定处理层的完整输出数据。在一个例示性操作中，然后可以将所生成的完整层输出数据存储在外部存储器组件中，这导致使得更多示例性本地存储器组件的存储器可用于后续层处理。

[0040] 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。

[0041] 神经网络背景：

[0042] 在人工神经网络中，神经元是用于模拟大脑中的生物神经元的基本单元。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。

[0043] 例示性地,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。

[0044] 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)可以例示性地维持1兆字节(1MB)的存储容量,其中高速专用接口以16GBps操作。L1高速缓存可以在内核和激活数据之间维持128KB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。

[0045] 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。可以在服务器场或数据中心中离线执行网络训练,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变换函数,其结果是分类或经语义标记的输出。

[0046] 在一个例示性操作中,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。

[0047] DNN模块在对应于神经网络的层的层描述符列表上操作。例示性地,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取读到DNN模块中并按顺序执行。

[0048] 通常,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。

[0049] DNN的输出也是数据blob。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取读数据。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取读。具有相关性集合的描述符将被阻止前进,直到满足相关性。

[0050] 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和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、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,若干组件包括预取读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。

[0051] 图2图示了根据本文所描述的系统和方法操作以计算和执行示例性数据分配序列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包括提供一个或多个用于执行的命令的一个或多个操作控制器225。一个或多个操作控制器225可以操作以生成指令

(例如,数据分配序列230),指令借助结构215通信到协作存储器组件本地存储器210以及一个或多个处理单元205(例如,神经元处理器)。此外,如图2所示,数据可以由处理单元205借助结构215可操作地检索并存储在主存储器组件220中。在一些实施例中,数据可以借助本地结构235从处理单元205存储到本地存储器210。在一个例示性实现中,神经网络环境结构可以是能够接收各种数据并将各种数据通信到一个或多个协作组件的数据总线。

[0052] 在例示性操作中,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,这些组件仅仅是例示性的,因为本领域普通技术人员将理解图6和图7中描述的处理将由图2中所示的组件之外的其他组件执行。

[0053] 图3图示了用于示例性输入数据的一个示例逻辑数据映射300。如图所示,数据305可以被表示为具有特定维度和体积340的数据,维度和体积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划分为若干部分并准备用于处理,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。

[0054] 在一个例示性操作中,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来确定数据305的各部分。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。

[0055] 图4图示了示例性神经网络环境400的示例性数据状态模型和示例性数据处理序列。如图所示,神经网络环境400可以包括主存储器组件220、本地存储器组件210和处理单元205(如图2所示)。在操作上,数据可以从主存储器220直接加载到本地存储器210,使得示例性第一层输入数据层1的输入410从主存储器220加载到本地存储器210。附加地,层1的权重420(例如,与层1的输入410相关联的处理参数)也可以从主存储器220加载到本地存储器210。如图1所示,然后可以将层1的输入和层1的权重通信到处理单元205以供处理层1 430处理。在操作上,处理单元205可以处理层1的输入数据410和层1的权重420以生成输出数据(未示出),输出数据然后可以作为层1的输出440存储到主存储器220。该序列包括完整的层1处理。

[0056] 此外,如图4所示,层2的输入数据450及其相关联的层2的权重460可以从主存储器220加载到本地存储器210。可操作地,层2的输入数据450和层2的权重460然后可以从本地存储器210通信到处理单元205,以用于处理。在一个例示性操作中,处理单元205可以在逐步处理序列步骤470处,处理层2的输入数据450和层2的权重460,以生成用于存储在主存储器220上的层2的输出数据480。

[0057] 在一个例示性操作中,通过序列步骤440在主存储器上生成的层1的输出数据存储可以是层2的输入数据450。这样的顺序层处理(其中先前处理层的输出可以充当后续处理层的输入)是常规深度神经网络环境处理协议的典型。图4的示例性数据处理序列依赖于在由示例性神经网络环境400的处理单元205处理后续层输入数据之前,被完全处理的每个层输入数据。

[0058] 仅出于示例性目的,图4描绘了示例性神经网络环境400,示例性神经网络环境400包括具有处于特定协作配置的主存储器组件220、本地存储器组件210和处理单元205的两个处理层。本领域普通技术人员将理解,这样的描述仅仅是例示性的,因为由本文描述的系

统和方法描述的发明构思考虑了具有各种处理层并支持主存储器220组件、本地存储器组件和处理单元205之间的各种协作配置的神经网络环境的操作。

[0059] 图5A图示了示例性网络环境500的另一示例性数据逐步序列处理模型。如图所示，示例性网络环境500可包括主存储器组件220、本地存储器组件210和处理单元205(如图2所示)。仅作为图示，示例性神经网络环境被描绘为支持两个处理层。在该图示中，一个或多个神经网络环境500的组件(未示出)(例如，操作控制器(一个或多个)225)可以针对层1的完整的输入数据(未示出)来计算数据分配序列。可以使用层1的输入数据的各种特性(例如，数据维度)以及主存储器和本地存储器组件的各种特性(例如，存储器大小、存储器时延、存储器位置等)、以及神经网络环境中可用的处理层的数量(例如，在该图示中为两个层)来计算数据分配序列。

[0060] 可操作地，计算出的分配序列可以包括可由示例性神经网络环境500的一个或多个组件(例如，图1的DNN模块105)执行的指令，以针对每个输入数据层部分，将每个层的输入数据以及相关联的处理参数划分为最优数据部分大小，当最优数据部分大小由示例性神经网络环境500处理时，将在处理周期期间对一个或多个协作存储器组件(例如，220、210)的效用进行优化。附加地，在一个例示性操作中，计算出的分配序列可以包括逐步处理序列，其中经分配的输入层的数目据在离散步骤中，在协作存储器组件(220和210)之间进行通信/由处理单元205处理。此外，示例性计算的分配序列可以包括根据所提供的逐步处理序列，将中间输出层的数目据存储在本地存储器中的指令。

[0061] 图5A示出了具有两个处理层的示例性神经网络环境中的输入数据的处理，两个处理层支持本地存储器组件210、主存储器组件220以及一个或多个处理单元205。例示性地，输入数据可以根据示例性计算的数据分配序列被分配，以包括层1部分1的部分(以及相关联的层1的权重)、层1部分2的部分以及层2的权重。附加地，在例示性实现中，示例性计算的分配序列可以包括关于何时将已分配的这些数据部分加载在可用存储器组件(本地存储器210和主存储器220)与处理单元205(例如，示例性逐步处理序列)之间的指令。根据示例性逐步处理序列在本地存储器210中存储的中间层输出数据也可以由计算出的分配序列中找到的一个或多个指令来管理。

[0062] 根据一个示例性计算的分配序列(未示出)，层1的输入部分1(即，根据示例性分配序列分配的)可以在逐步处理序列步骤505处，从主存储器组件220加载到本地存储器组件210。附加地，层1的权重(例如，层1部分1的输入数据505的处理参数)可以在逐步处理序列步骤510处，从主存储器组件220加载到本地存储器组件220。处理单元205可以根据逐步处理序列步骤515，处理层1部分1的输入数据和层1的权重，以在逐步处理序列步骤520处，生成并存储层1部分1的输出数据。处理沿逐步处理序列前进到步骤525，其中从主存储器220加载到本地存储器210的层2的权重允许在逐步处理序列步骤530处，由处理单元205处理层2部分1的输入数据(即，如本文所述，给定处理层的输出可以充当后续层的输入，其中在该例示性实现中，层1的输出是层2的输入数据)。在逐步处理序列步骤535处，层2部分1的输出数据然后可以存储在主存储器220中。

[0063] 可操作地，然后在逐步处理序列步骤540处，将层1输入数据的剩余部分(即，层1部分2的输入数据)从主存储器220加载到本地存储器210，并且在逐步处理序列步骤545处处理其以生成层1部分2的输出数据，然后在逐步处理步骤550处，将层1部分2的输出数据存储

在本地存储器210中。然后在逐步处理序列步骤555处，从本地存储器加载层2的剩余输入数据部分(即，层2部分2也是层1部分2的输出数据)，以由处理单元205处理。在逐步处理序列步骤560处，生成层2部分2的输出数据以存储到主存储器220中，并完成由示例性神经网络环境500针对例示性实现的最初接收的输入数据生成完整输出数据集(即，由于不再存在处理层，存储在主存储器中的层2部分1的输出数据和层2部分2的输出数据可以表示由示例性神经网络环境500针对最初接收的输入数据生成的输出数据的总和)。

[0064] 仅出于示例性目的，图5A描绘了示例性神经网络环境500，示例性神经网络环境500包括具有处于特定协作配置的主存储器组件220、本地存储器组件210和处理单元205的两个处理层。本领域普通技术人员将理解，该描述仅仅是例示性的，因为由本文所描述的系统和方法描述的发明构思考虑了具有各种处理层并支持主存储器220组件、本地存储器组件210和处理单元205之间的各种协作配置以及各种逐步处理序列的神经网络环境的操作，各种逐步处理序列可以基于示例性神经网络环境500的组件的特性而不同。

[0065] 图5B是示出示例性神经网络环境500A的示例性处理序列570、575和580的框图。如图所示，处理序列570可以处理来自可用处理层(层0、层1和层2)的数据分区(570(a)、570(b)、570(c)、570(d)、570(e)、570(f)、570(g)、570(h)、570(i)、570(j)、570(k)和570(l)-为了简单起见，在下文中称为570(a)-570(a))。在该处理序列中，按顺序(如由处理序列步骤1-12所指示的)处理数据分区(570(a)-570(a))(例如，层0;1-4，层1;5-8和层2;9-12)，使得层0的数据分区(即，570(a)、570(b)、570(c)和570(d))首先被处理，然后是层1的数据分区(即，570(e)、570(f)、570(g)和570(h))，然后是层2的数据分区(即，570(i)、570(j)、570(k)和570(l))。该类型的处理序列可以被认为是仅“广度”的处理序列。

[0066] 此外，如图5B所示，示例性NN/DNN环境也可以执行处理序列575。在该处理序列中，按顺序处理来自每个层的数据分区(即，1-575(a)、2-575(e)、3-575(i))，然后按顺序处理来自每个层的第二数据分区(即，4-575(b)、5-575(f)、6-575(j))，直到跨越各层，所有数据分区都被处理。该类型的处理序列可以被认为是“深度优先”处理序列。

[0067] 进一步如图5B所示，示例性NN/DNN环境也可以执行处理序列580。在该处理序列中，可以根据仅“广度”和/或“深度优先”处理顺序来处理数据分区(570(a)、570(b)、570(c)、570(d)、570(e)、570(f)、570(g)、570(h)、570(i)、570(j)、570(k)和570(l))，使得在一个例示性操作中，可以根据“深度优先”处理序列顺序地处理来自每个层的数据分区(例如，1-580(a)、2-580(e)、3-580(i))，并且然后可以根据仅“广度”处理序列来处理来自单个分区的数据(4-580(b)、5-580(c)，然后是6-580(f)、7-580(g)，然后是8-580(j)、9-580(k))，然后可以根据“深度优先”处理序列来处理来自每个分区的数据(10-580(d)、11-580(h)、12-580(l))。因此，“动态深度优先”处理序列可以可操作地允许在数据和环境特征指示时，通过示例性交织操作，从仅“广度”切换到“深度优先”处理序列来处理层分区。“动态深度优先”处理序列可以提供附加的鲁棒性，以根据环境和/或数据特性对性能和存储器使用进行优化。

[0068] 应理解，图5B中所图示的示例性处理序列仅是例示性的，因为本文描述的发明构思可以利用可操作地优化示例性NN/DNN环境的性能和/或存储器利用的其他处理序列。附加地，尽管分区的大小被示出为具有相同的大小，但是本文所描述的发明构思考虑了不同大小的数据分区。

[0069] 图6是使用深度优先方法处理数据以增强神经网络环境的性能的例示性过程600的流程图。应理解，本文所公开的过程的操作不一定以任何特定的顺序呈现，并且以可替换的顺序执行一些或所有操作是可能的并且是可预期的。为了便于描述和图示，已经以示出的顺序呈现了操作。在不脱离所附权利要求的范围的情况下，可以添加、省略和/或同时执行操作。

[0070] 还应理解，所图示的过程(也称为“方法”或“例程”)可以在任何时间结束，并且不需要完整地执行。如下所述，可以通过执行包括在计算机存储介质上的计算机可读指令来执行方法的一些或所有操作和/或基本等同的操作。如在说明书和权利要求中使用的术语“计算机可读指令”及其变型在本文中被广泛使用，以包括例程、应用程序、应用程序模块、程序模块、程序、组件、数据结构、算法等。计算机可读指令可以在各种系统配置(包括单处理器或多处理器系统、小型计算机、大型计算机、个人计算机、手持计算设备、基于微处理器的可编程消费电子产品及其组合等)上实现。

[0071] 因此，应当理解，本文所描述的逻辑操作被实现为(1)在计算系统上运行的计算机实现的动作或程序模块的序列和/或(2)计算系统内的互连机器逻辑电路或电路模块。实现是取决于计算系统的性能和其他要求的选择问题。因此，本文所描述的逻辑操作被不同地称为状态、操作、结构设备、动作或模块。这些操作、结构设备、动作和模块可以在软件、固件、专用数字逻辑及其任何组合中被实现。

[0072] 例如，过程600的操作在本文中被描述为至少部分地由本文描述的组件和/或远程系统的组件而被实现。在一些配置中，本文所描述的组件或运行本文所公开的特征的另一模块可以是动态链接库(DLL)、静态链接库、由应用程序编程接口(API)产生的功能、编译程序、解释程序、微代码、机器代码、脚本或任何其他可执行指令集。数据可以存储在一个或多个存储器组件中的数据结构中。可以通过对链接寻址或参考数据结构来从数据结构中检索数据。

[0073] 尽管以下图示涉及附图的组件，但是可以理解，例程的操作也可以按照许多其他方式而被实现。例如，过程600可以至少部分地由另一远程电路或本地电路的处理器而被实现。附加地，过程600的一个或多个操作可以备选地或附加地至少部分地由单独工作或与其他软件模块结合工作的芯片组实现。适合于提供本文所公开的技术的任何服务、电路或应用程序可以在所描述的操作中使用。

[0074] 如图所示，处理开始于框605，在框605处，从神经联网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，从神经联网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。在框615处，可以计算输入数据分配序列，输入数据分配序列提供逐步处理序列和相关指令，用于根据“仅广度”、“深度优先”和“动态深度优先”处理序列中的一个或多个，在示例性神经网络环境的一个或多个协作存储器组件中，针对每个层，分配、处理并存储一个或多个分配数据(例如，输入或所生成的输出数据)。处理然后进行到框620，在框620处，根据框615的计算出的数据分配序列的逐步处理序列，处理经分配的输入数据或所生成的中间输出数据的部分。处理然后进行到框625，在框625处，根据计算出的数据分配序列，经处理的分配数据可以存储在示例性神经网络环境的一个或多个协作存储器组件中(例如，在本地或外部/主存储器组件中)。然后在框630处，执行检查来确定逐步处理序列的所有步骤是否都已完成。

[0075] 如果框630处的检查指示不存在待逐步处理序列处理的附加步骤，则处理在框635处终止。但是，如果框630处的检查指示存在待逐步处理序列处理的附加步骤(例如，存在待处理的更多数据层部分)，处理返回到框620并从那里继续。

[0076] 图7是利用深度优先处理以增强示例性神经网络环境的整体处理性能的例示性过程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包括表示输入数据的维度、示例性神经网络环境的一个或多个存储器组件的一个或多个特性、和/或示例性神经网络环境的处理层的数量的数据。然后处理进行到框710，在框710处，可以使用输入数据维度数据来计算输入数据分配序列，输入数据维度数据在示例性神经网络环境的一个或多个协作存储器组件中，针对每个层，提供逐步处理序列以及用于分配、处理和存储一个或多个分配数据(例如，输入或所生成的输出)的相关指令。处理然后进行到框715，在框715处，一个或多个处理指令由示例性神经网络环境的一个或多个协作组件(例如，处理单元/神经处理器)接收，以根据示例性逐步处理序列(例如，根据“仅广度”、“深度优先”和“动态深度优先”处理序列中的一个或多个)来处理分配数据的指定部分。处理然后进行到框720，在框720处，使用所接收的一个或多个处理指令、根据逐步处理序列来处理分配数据的指定部分。在框725处，根据示例性逐步处理序列，将经处理的分配数据的输出存储在一个或多个协作存储器组件(例如，本地存储器或外部/主存储器)中。

[0077] 然后在框730处执行检查来确定逐步处理序列的所有步骤是否已完成。如果框730处的检查指示不存在待逐步处理序列处理的附加步骤，则处理在框735处终止。但是，如果框730处的检查指示存在待逐步处理序列处理的附加步骤(例如，存在待处理的更多数据层部分)，则处理返回到框715并从那里继续。

[0078] 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内的元件之间传送信息的基本例程的基本输入/输出系统可以存储在ROM 808中。计算机架构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。

[0079] 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。

[0080] 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。

[0081] 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。

[0082] 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。

[0083] 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进行变换,从而对构成CPU 802的晶体管或其他分立硬件元件进行变换。

[0084] 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。

[0085] 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而被实现。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。

[0086] 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。

[0087] 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。

[0088] 图9图示了具有服务器的示例性例示性联网计算环境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，服务器(一个或多个)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)与若干客户端计算环境(例如，平板个人计算机910、移动电话915、电话920、个人计算机(一个或多个)801、个人数字助理925、智能电话手表/个人目标跟踪器(例如，Apple Watch、Samsung、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，服务器(一个或多个)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、910、915、920、925、930和935中的每一个可以配备有计算环境905，计算环境905操作以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。

[0089] 服务器(一个或多个)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。在一个例示性操作中，用户(未示出)可以与在客户端计算环境(一个或多个)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905、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服务器计算环境(一个或多个)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和服务器计算环境(一个或多个)905之间通信，以进行处理和存储。服务器计算环境(一个或多个)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)、网络附接存储(NAS)和存储区域网络(SAN)协作来实现应用程序/数据交易。

[0090] 示例条款

[0091] 鉴于以下条款，可以考虑本文所呈现的公开内容。

[0092] 示例条款A，一种用于神经网络环境中的增强数据处理的系统，系统包括至少一个处理器；至少一个第一存储器组件；以及与至少一个处理器通信的至少一个第二存储器组件，至少一个第一存储器组件和/或第二存储器组件具有存储在其上的计算机可读指令，计算机可读指令当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神经网络环境处理的数据的维度的数据；计算针对数据的分配序列，分配序列包括用以在至少一个第一存储器组件和/或至少一个第二存储器组件中存储输出数据的一个或多个指令，分配序列包括仅广度处理序列、深度优先处理序列和动态深度优先处理序列；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用以根据计算出的分配序列处理数

据的所选择的部分的一个或多个指令;由一个或多个协作处理单元处理数据的部分以生成输出数据,以用于存储在至少一个第一存储器组件或至少一个第二存储器组件上;以及将所生成的输出数据存储在至少一个第一存储器组件或至少一个第二存储器组件上。

[0093] 示例条款B,根据示例条款A所述的系统,其中计算出的分配序列基于神经网络环境的层的数目。

[0094] 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加载表示一个或多个层权重的数据,以用于由神经网络环境的一个或多个处理单元(205)使用以生成输出数据。

[0095] 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使得至少一个处理器将所生成的输出数据存储在至少一个第一存储器组件中。

[0096] 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将所生成的输出数据存储在至少一个第二存储器组件中。

[0097] 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基于至少第一存储器组件的大小和/或至少一个第二存储器组件的大小来计算分配序列。

[0098] 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将所生成的输出数据存储在至少一个第一存储器组件中。

[0099] 示例条款H,根据示例条款A至G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生成的输出数据存储在至少一个第二存储器组件中。

[0100] 示例条款I,一种计算机实现的方法,包括:从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输入数据的维度的数据以及表示与输入数据相关联的处理层的数量的数据;计算针对输入数据的分配序列,分配序列包括用以将输出数据存储在至少一个第一存储器组件和/或至少一个第二存储器组件中的一个或多个指令,并且包括用以加载处理输入数据所需的相关联数据的一个或多个指令,分配序列包括仅广度处理序列、深度优先处理序列和动态深度优先处理序列;从协作存储器组件加载表示一个或多个处理权重的数据;从协作控制器组件接收用以根据计算出的分配序列处理数据的所选择的部分的一个或多个指令;由一个或多个协作处理单元根据计算出的分配序列处理数据部分以生成输出数据,以用于存储在至少一个第一存储器组件或至少一个第二存储器组件上;以及将所生成的输出数据存储在至少一个第一存储器组件或至少一个第二存储器组件上。

[0101] 示例条款J,根据条款I所述的计算机实现的方法,还包括接收表示至少一个第一存储器组件的大小和至少一个第二存储器组件的大小的数据,以用于在计算分配序列时使用。

[0102] 示例条款K,根据条款I和J所述的计算机实现的方法,还包括将所生成的输出数据存储在至少一个第一存储器组件上。

[0103] 示例条款L,根据条款I至K所述的计算机实现的方法,还包括将所生成的输出数据存储在至少一个第二存储器组件上。

[0104] 示例条款M,根据条款I至L所述的计算机实现的方法,还包括将所生成的输出数据的一部分存储在至少一个第一存储器组件上。

[0105] 示例条款N,根据条款I至M所述的计算机实现的方法,还包括加载与正沿计算出的

分配序列而被处理的数据部分相关联的处理权重数据。

[0106] 示例条款O,根据条款I至N所述的计算机实现的方法,还包括针对每个处理层加载整个输入数据相关联的处理权重数据。

[0107] 示例条款P,根据条款I至O所述的计算机实现的方法,还包括将先前处理层的输出数据作为后续处理层的输入数据处理。

[0108] 示例条款Q,一种具有存储在其上的计算机可执行指令的计算机可读存储介质,计算机可执行指令当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,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,计算针对数据的分配序列,分配序列包括用以在至少一个第一存储器组件和/或至少一个第二存储器组件中存储输出数据以及在至少一个第一存储器组件和至少一个第二存储器组件之间加载包括输入数据以及相关联的输入数据参数的数据的一个或多个指令,分配序列包括仅广度处理序列、深度优先处理序列和动态深度优先处理序列,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用以根据计算出的分配序列处理数据的所选择的部分,使用相关联的输入数据参数由一个或多个协作处理单元处理数据部分以生成输出数据以用于存储在至少一个第一存储器组件或至少一个第二存储器组件上的一个或多个指令;以及根据计算出的分配序列将所生成的输出数据存储在至少一个第一存储器组件或至少一个第二存储器组件上。

[0109] 示例条款R,根据条款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:根据计算出的分配序列将表示针对数据的部分的处理权重的数据从外部存储器组件加载到神经网络环境的本地存储器组件。

[0110] 示例条款S,根据条款Q和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:根据计算出的分配序列将针对数据的部分的所生成的输出数据存储在本地存储器组件中。

[0111] 示例条款T,根据条款Q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:生成用于神经网络环境的最终处理层的输出数据。

[0112] 示例条款U,根据条款Q至T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:将所生成的最终处理层输出数据存储在外部存储器组件中。

[0113] 示例条款V,根据条款Q至U所述的计算机可读存储介质,其中存储器组件与能够产生输入数据的物理传感器协作,输入数据包括音频数据、视频数据、触觉感测数据以及用于由一个或多个协作处理单元后续处理的其他数据。

[0114] 示例条款W,根据条款Q至V所述的计算机可读存储介质,其中协作处理单元与一个或多个输出物理组件电子地协作,一个或多个输出物理组件操作以接收用于人类交互的经处理的输入数据,经处理的输入数据包括音频数据、视频数据、触觉感测数据和其他数据。

[0115] 结论

[0116] 最后,尽管已利用结构特征和/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。

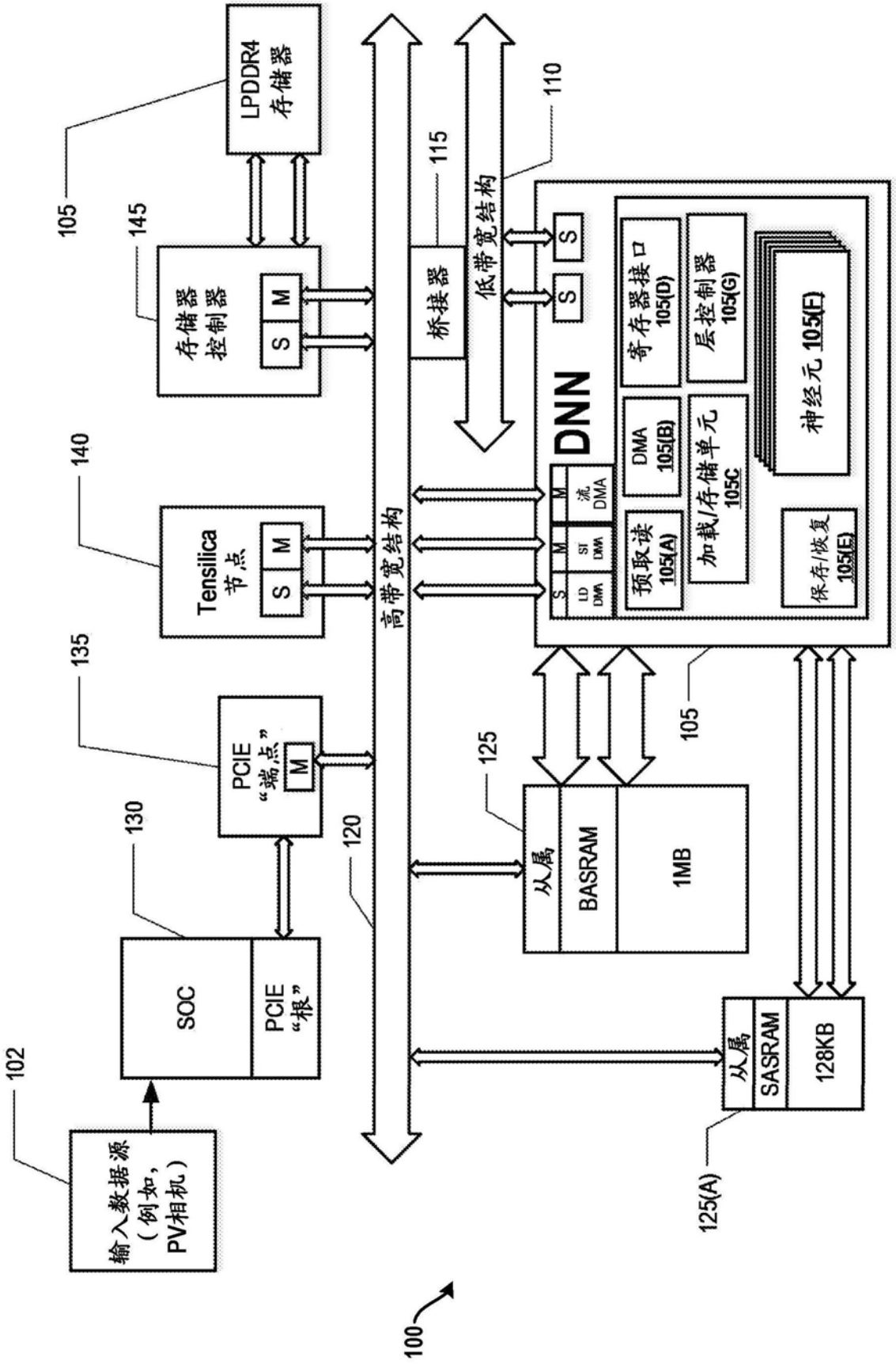

LPDDR4 存储器
105
145
存储器控制器
S M
115
桥接器
低带宽结构
110
S
S
DNN
寄存器接口 105(D)
层控制器 105(G)
神经元 105(F)
DMA 105(B)
加载/存储单元 105C
预取读 105(A)
保存/恢复 105(E)
M 流 DMA
M SI DMA
S LD DMA
140
Tensilica 节点
S M
高带宽结构
135
PCIE “端点”
M
130
SOC
PCIE “根”
120
125
从属
BASRAM
1MB
105
从属
SASRAM
128KB
125(A)
102
输入数据源（例如，PV相机）
100

图1

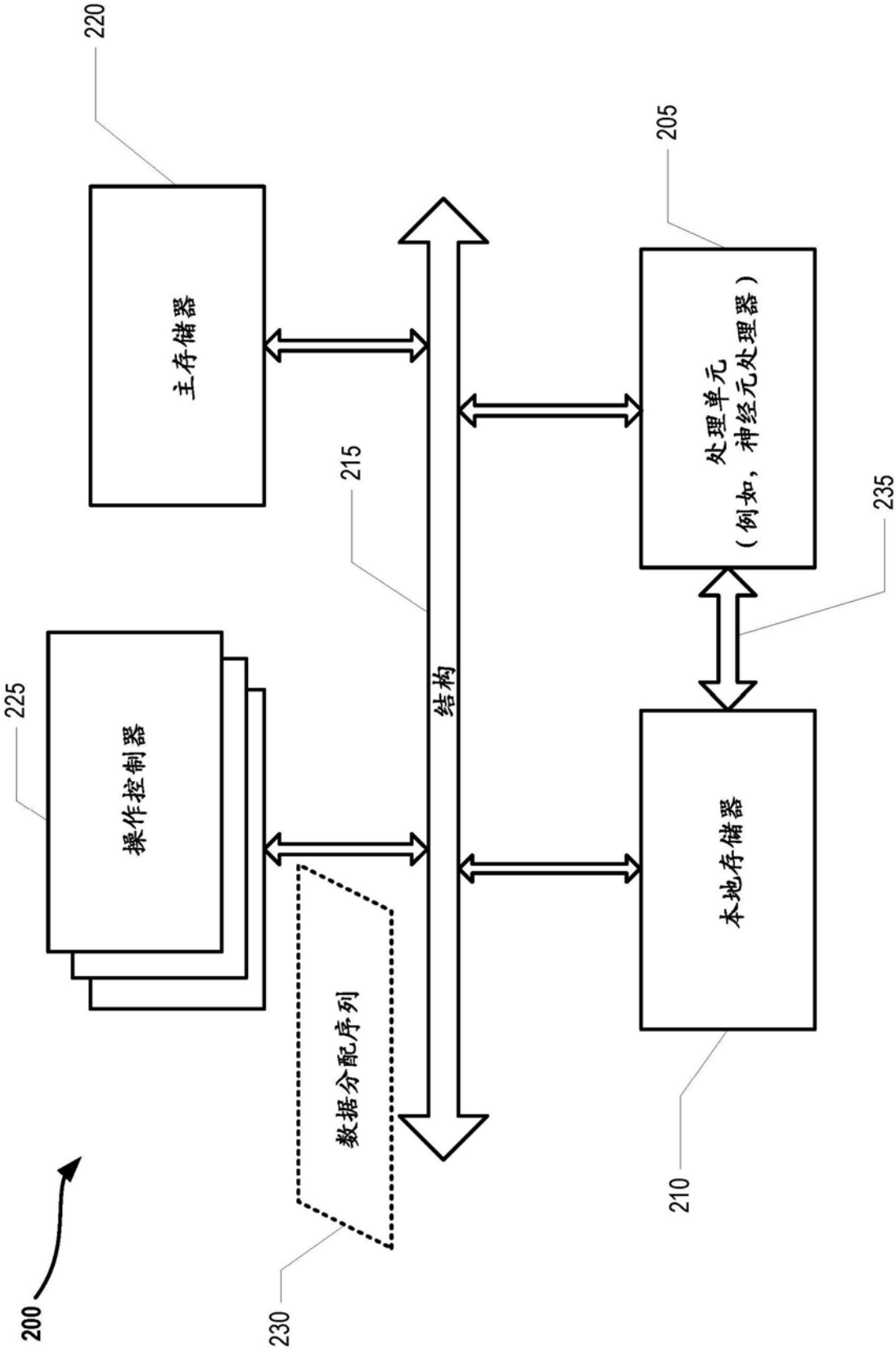

220
主存储器
205
处理单元
（例如，神经元处理器）
215
结构
235
225
操作控制器
本地存储器
数据分配序列
230
210
200

图2

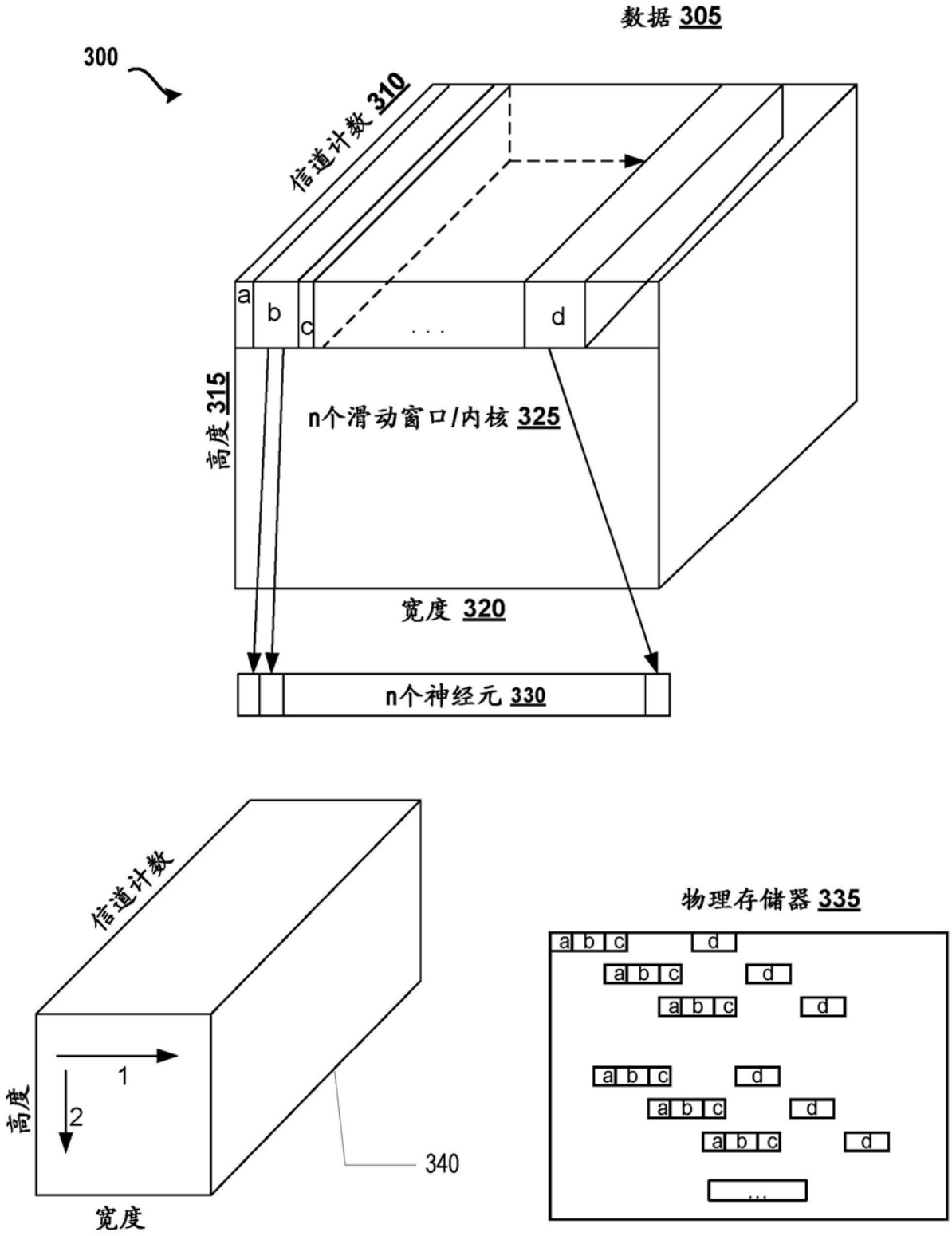

数据 305
300
信道计数 310
a
b
c
. . .
d
高度 315
n个滑动窗口/内核 325
宽度 320
n个神经元 330
信道计数
高度
1
2
340
宽度
物理存储器 335
a b c d
a b c d
a b c d
a b c d
a b c d
a b c d
...

图3

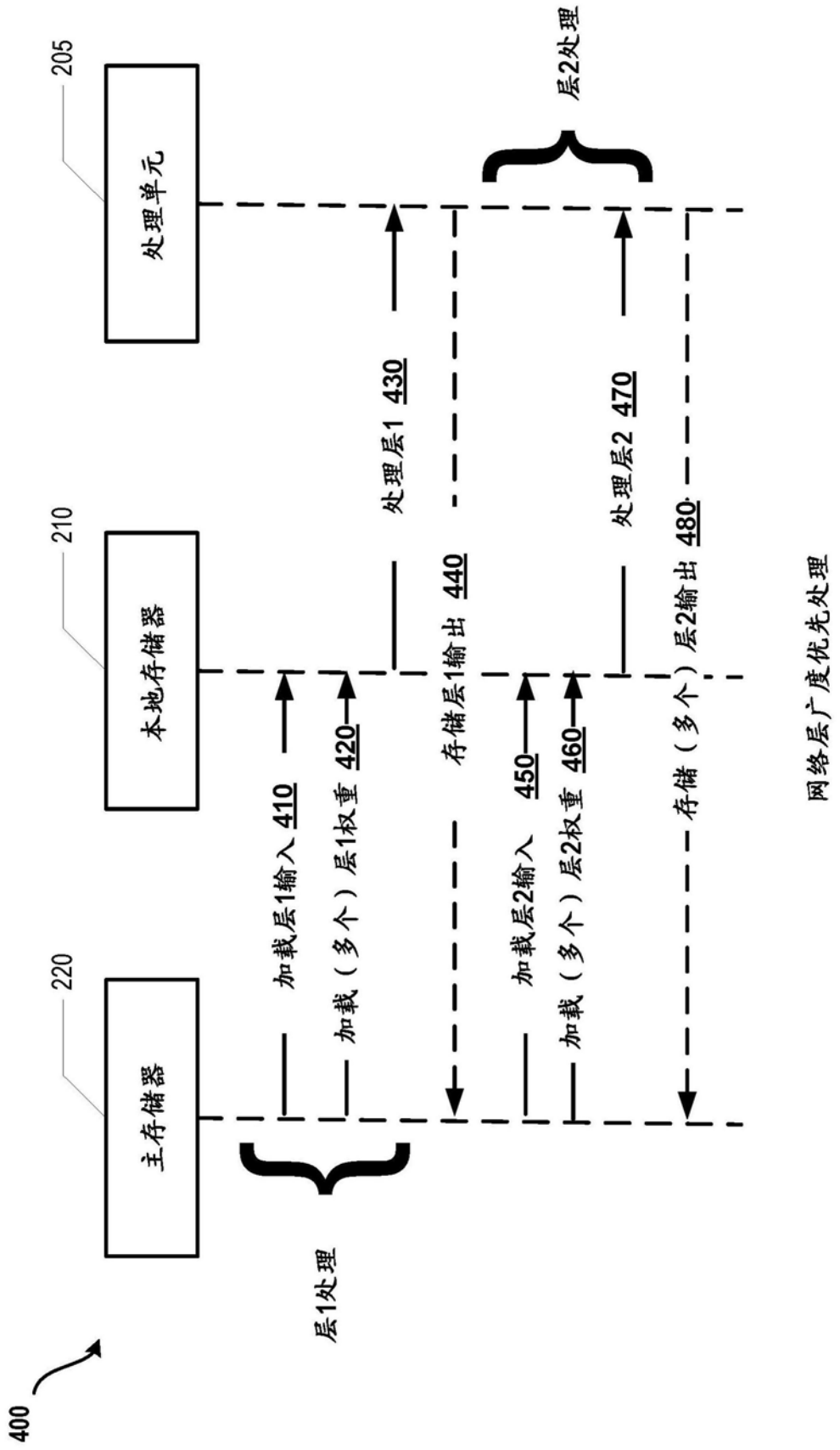

400
205
处理单元
210
本地存储器
220
主存储器
加载层1输入 410
加载（多个）层1权重 420
处理层1 430
存储层1输出 440
加载层2输入 450
加载（多个）层2权重 460
处理层2 470
存储（多个）层2输出 480
层1处理
层2处理
网络层广度优先处理

图4

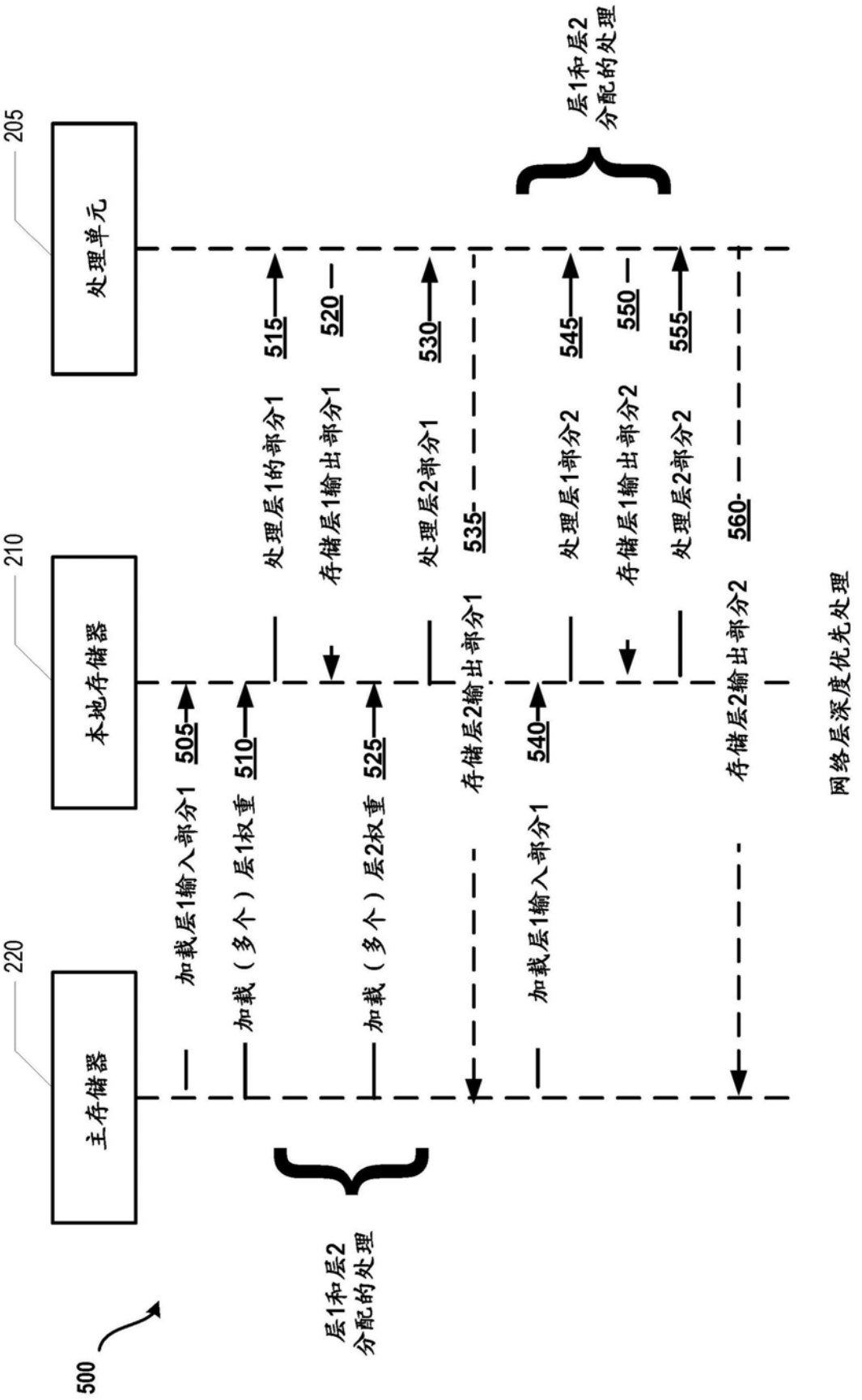

205
处理单元
210
本地存储器
220
主存储器
500
加载层1输入部分1 505
加载（多个）层1权重 510
处理层1的部分1 515
存储层1输出部分1 520
加载（多个）层2权重 525
处理层2部分1 530
存储层2输出部分1 535
加载层1输入部分1 540
处理层1部分2 545
存储层1输出部分2 550
处理层2部分2 555
存储层2输出部分2 560
层1和层2分配的处理
层1和层2分配的处理
网络层深度优先处理

图5A

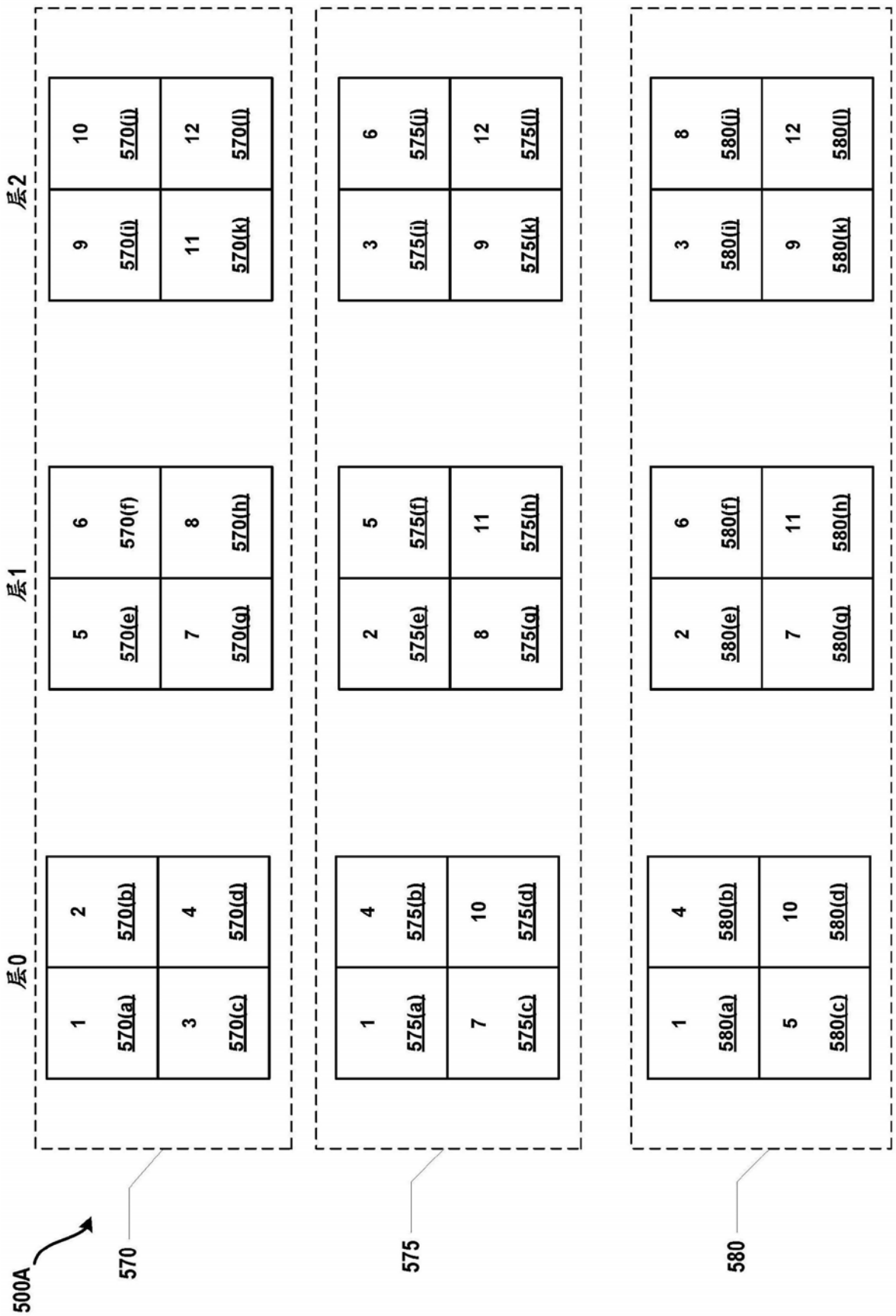

层2
层1
层0
1 570(a)
2 570(b)
3 570(c)
4 570(d)
5 570(e)
6 570(f)
7 570(g)
8 570(h)
9 570(i)
10 570(j)
11 570(k)
12 570(l)
1 575(a)
4 575(b)
7 575(c)
10 575(d)
2 575(e)
5 575(f)
8 575(g)
11 575(h)
3 575(i)
6 575(j)
9 575(k)
12 575(l)
1 580(a)
4 580(b)
5 580(c)
10 580(d)
2 580(e)
6 580(f)
7 580(g)
11 580(h)
3 580(i)
8 580(j)
9 580(k)
12 580(l)
500A
570
575
580

图5B

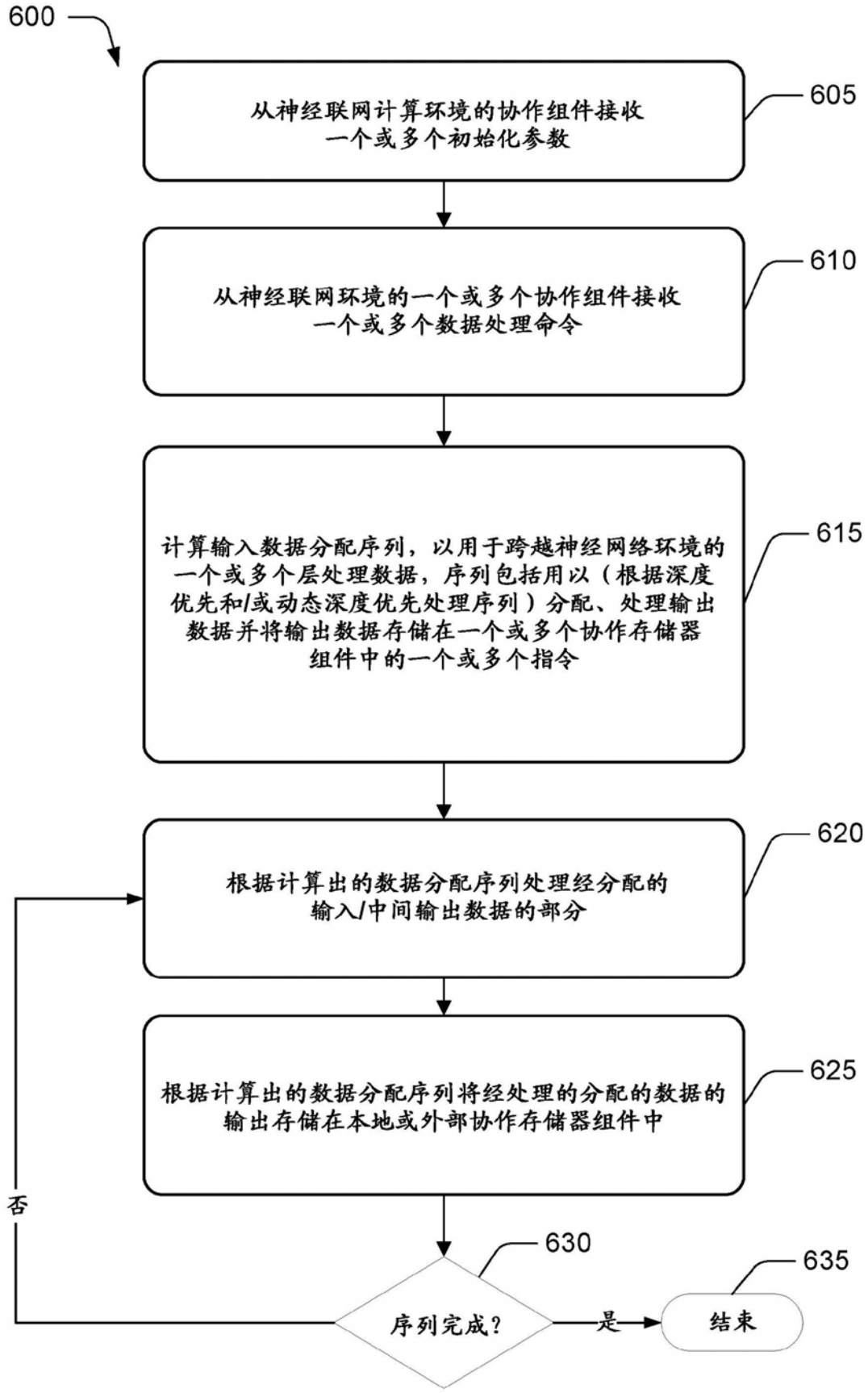

600
从神经联网计算环境的协作组件接收一个或多个初始化参数
605
从神经联网环境的一个或多个协作组件接收一个或多个数据处理命令
610
计算输入数据分配序列，以用于跨越神经网络环境的一个或多个层处理数据，序列包括用以（根据深度优先和/或动态深度优先处理序列）分配、处理输出数据并将输出数据存储在一个或多个协作存储器组件中的一个或多个指令
615
根据计算出的数据分配序列处理经分配的输入/中间输出数据的部分
620
根据计算出的数据分配序列将经处理的分配的数据的输出存储在本地或外部协作存储器组件中
625
序列完成?
630
否
是
结束
635

图6

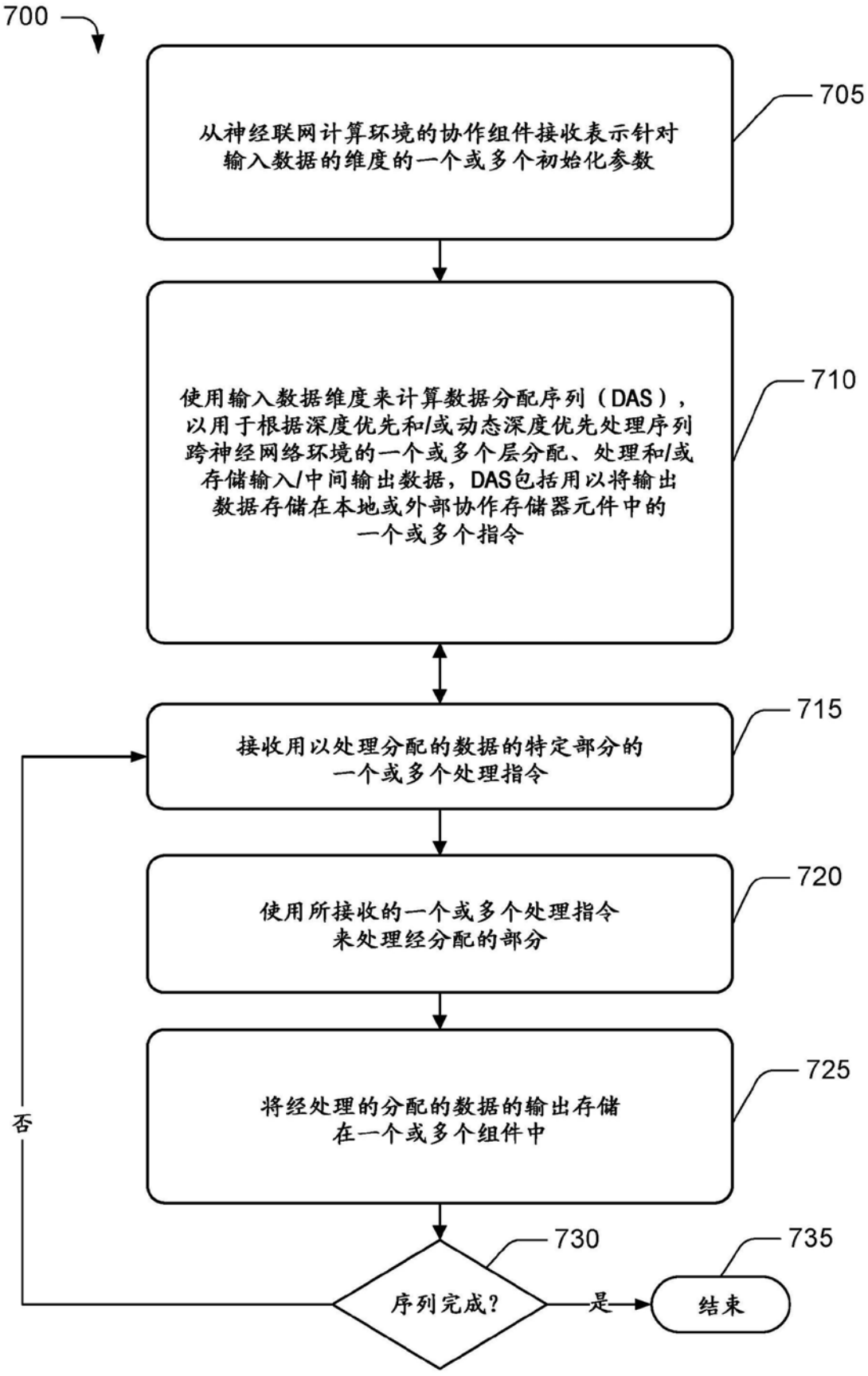

700
从神经联网计算环境的协作组件接收表示针对输入数据的维度的一个或多个初始化参数
705
使用输入数据维度来计算数据分配序列（DAS），以用于根据深度优先和/或动态深度优先处理序列跨神经网络环境的一个或多个层分配、处理和/或存储输入/中间输出数据，DAS包括用以将输出数据存储在本地或外部协作存储器元件中的一个或多个指令
710
接收用以处理分配的数据的特定部分的一个或多个处理指令
715
使用所接收的一个或多个处理指令来处理经分配的部分
720
将经处理的分配的数据的输出存储在一个或多个组件中
725
730
序列完成?
否
是
735
结束

图7

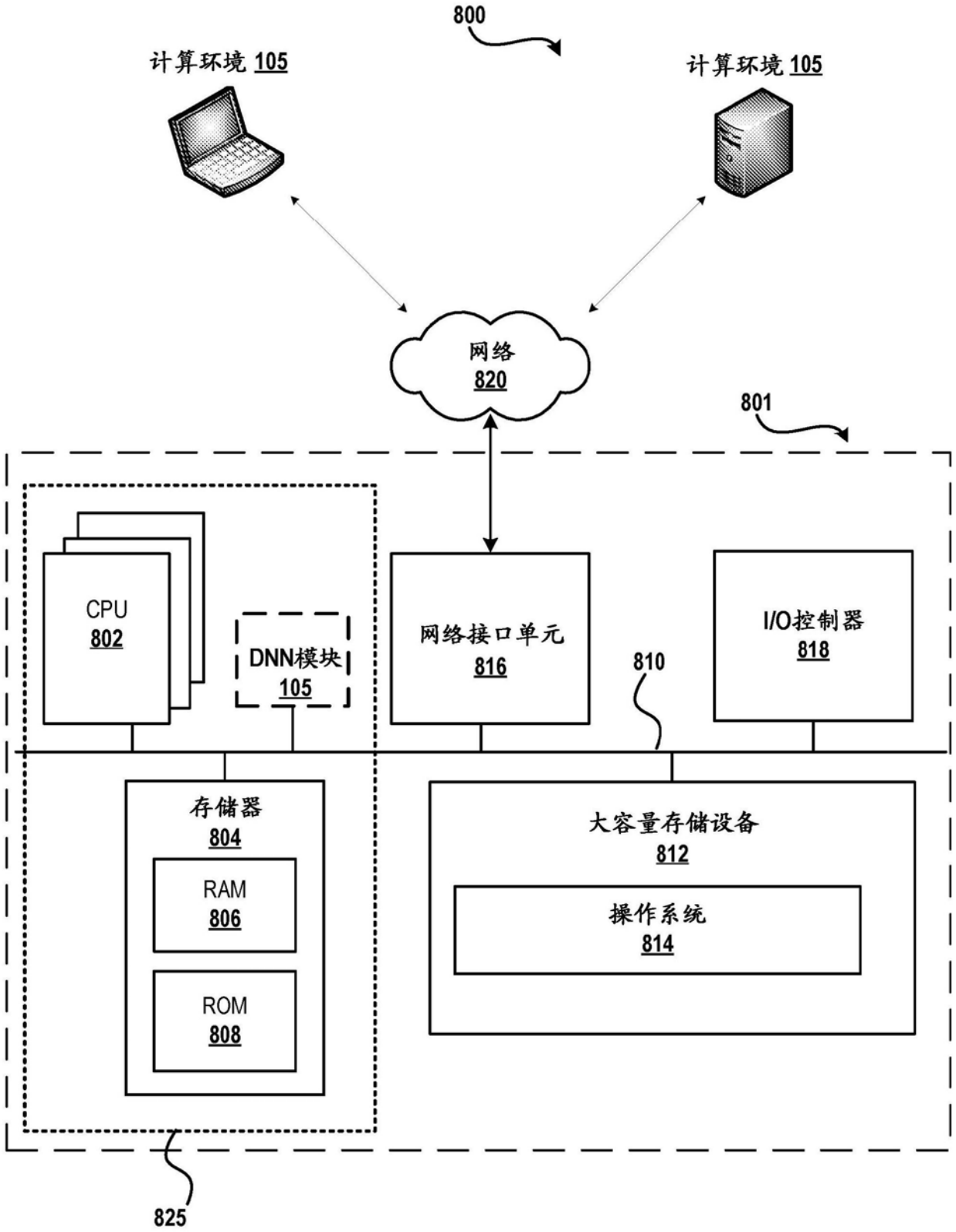

800
计算环境 105
计算环境 105
网络
820
801
CPU
802
DNN模块
105
网络接口单元
816
I/O控制器
818
810
存储器
804
RAM
806
ROM
808
大容量存储设备
812
操作系统
814
825

图8

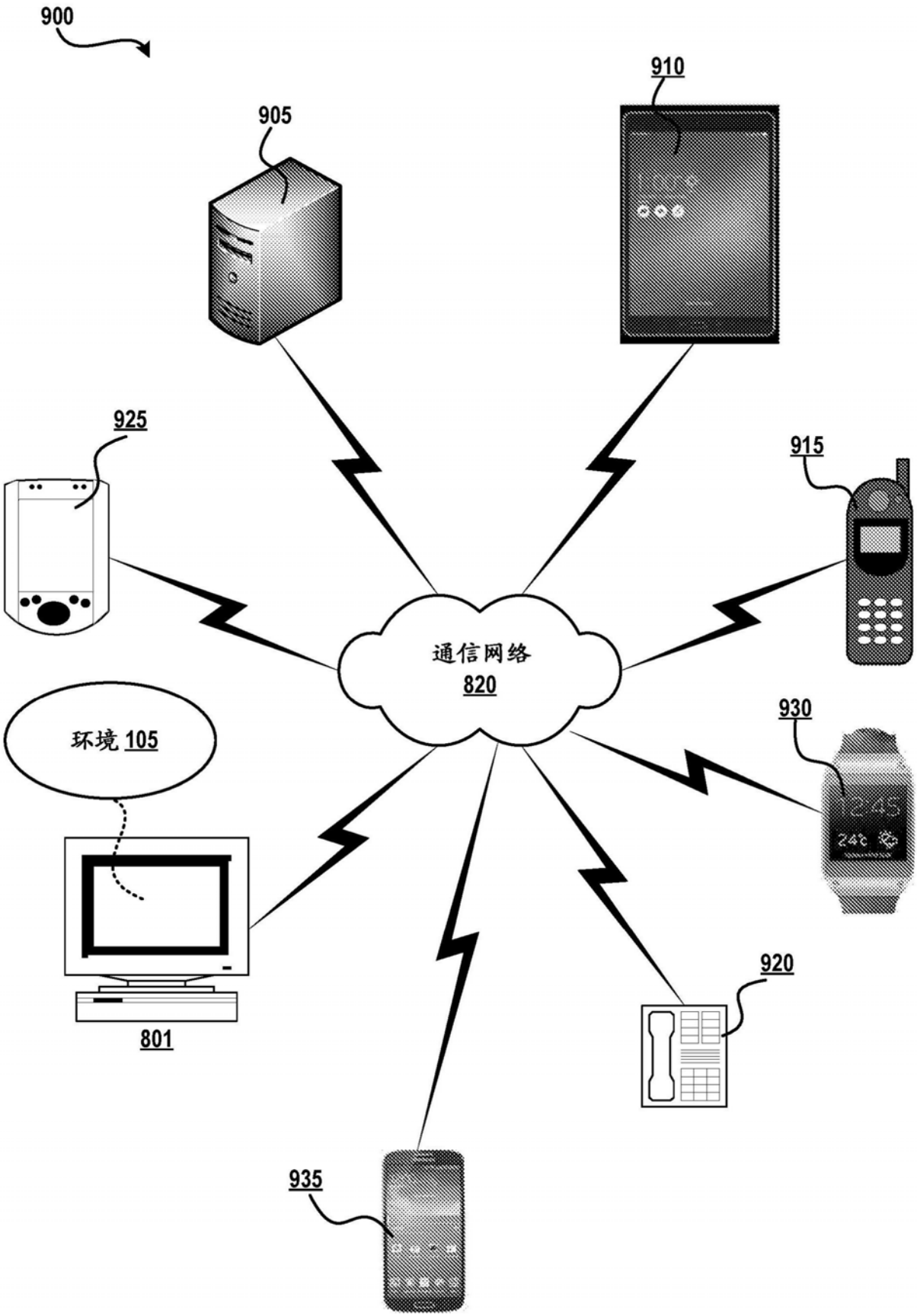


图9