(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 110520870 A
(43)申请公布日 2019.11.29

(21)申请号 201880025227.8

(22)申请日 2018.04.06

(30)优先权数据
62/486,432 2017.04.17 US
15/881,519 2018.01.26 US

(85)PCT国际申请进入国家阶段日
2019.10.15

(86)PCT国际申请的申请数据
PCT/US2018/026358 2018.04.06

(87)PCT国际申请的公布数据
WO2018/194851 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 A·A·安巴德卡 A·托米克 C·B·麦克布赖德 G·彼得 K·D·塞多拉 L·M·瓦尔 B·博布罗夫

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 丁君军

(51)Int.Cl.
G06N 3/06(2006.01)
G06F 9/50(2006.01)

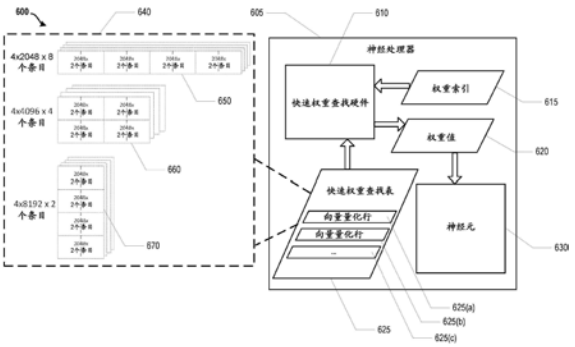
权利要求书2页 说明书13页 附图9页

(54)发明名称

用于具有动态向量长度和码本大小的高吞吐量向量去量化的灵活硬件

(57)摘要

神经网络(NN)和/或深度神经网络(DNN)的性能可以由正执行的操作数目以及NN/DNN的存储器数据管理来限制。使用神经元权重值的向量量化,神经元的数据的处理可以优化操作的数目以及存储器利用以便增强NN/DNN的总体性能。操作地,权重值的一个或多个连续段可以被转换为任意长度的一个或多个向量,并且一个或多个向量中的每个向量可以被分配有索引。所生成的索引可以被存储在示例性向量量化查找表中并且在飞行中在运行时由示例性快速权重查找硬件来取回,作为NN的示例性数据处理功能的一部分,作为内联去量化操作的一部分,以获得所需要的一个或多个神经元权重值。



CN 110520870 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;

至少一个存储器部件;以及

至少一个存储器,其与所述至少一个处理器通信,所述至少一个存储器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部件接收一个或多个初始化参数,所述初始化参数包括代表待由所述神经网络环境处理的数据的维度的数据以及代表一个或多个向量量化索引值的数据,所述一个或多个索引值代表被存储在所述至少一个存储器部件上的一个或多个向量,所述一个或多个向量包括代表一个或多个神经元权重值的一个或多个连续段的数据;

利用所述一个或多个向量量化索引值从所述至少一个存储器部件取回代表一个或多个神经元权重值的所述一个或多个向量;

将所取回的所述一个或多个向量进行去量化以取回底层的一个或多个神经元权重值;以及

传递所述一个或多个神经元权重值以用于由所述神经网络环境的所述一个或多个处理部件处理。

2.根据权利要求1所述的系统,其中所述一个或多个向量被存储在驻留在所述至少一个存储器部件上的快速查找表中。

3.根据权利要求2所述的系统,其中所述一个或多个向量具有任意长度。

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从所述快速查找表的一个或多个行取回所述一个或多个向量。

5.根据权利要求4所述的系统,其中所述一个或多个向量的向量长度对于所述神经网络环境的所述神经元层中的每个神经元层是可选择的。

6.根据权利要求5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执行用于所述神经网络环境的所述神经元层中的所选择的一个或多个神经元层的一个或多个神经元权重值的向量去量化。

7.根据权利要求2所述的系统,其中所述计算机可读指令还包括可操作以执行存储在所述快速查找表上的所述向量的快速查找的一个或多个硬件部件。

8.一种计算机实现的方法,包括:

从所述神经网络环境的协作控制器部件接收一个或多个初始化参数,所述初始化参数包括代表待由所述神经网络环境处理的数据的维度的数据以及代表一个或多个向量量化索引值的数据,所述一个或多个索引值代表被存储在所述至少一个存储器部件上的一个或多个向量,所述一个或多个向量包括代表一个或多个神经元权重值的一个或多个连续段的数据,所述一个或多个向量由所述神经网络环境的处理器生成;

利用所述一个或多个向量量化索引值从所述至少一个存储器部件取回代表一个或多个神经元权重值的所述一个或多个向量,所述一个或多个向量被操作地存储在快速查找表上;

将所取回的所述一个或多个向量进行去量化以取回底层的一个或多个神经元权重值;

以及

传递所述一个或多个神经元权重值以用于由所述神经网络环境的所述一个或多个处理部件处理。

9.根据权利要求8所述的计算机实现的方法，还包括通过所述神经网络环境的一个或多个协作硬件部件对所取回的所述一个或多个向量进行内联去量化，以获得所述一个或多个神经元权重值。

10.根据权利要求8所述的计算机实现的方法，还包括利用用于所生成的所述一个或多个向量的协作存储器部件中的基索引来生成虚拟化的一个或多个快速查找表。

11.根据权利要求8所述的计算机实现的方法，还包括生成用于所述神经网络环境的一个或多个神经元层的一个或多个向量。

12.根据权利要求11所述的计算机实现的方法，还包括将所述一个或多个向量存储在快速查找表的一个或多个行中。

13.根据权利要求12所述的计算机实现的方法，还包括生成任意长度的一个或多个向量。

14.根据权利要求8所述的计算机实现的方法，还包括选择向量长度以用于所述神经网络环境的所述神经元层中的每个神经元层的所述一个或多个向量的所述生成。

15.根据权利要求8所述的计算机实现的方法，还包括将所生成的所述一个或多个向量存储在本地存储器部件中。

用于具有动态向量长度和码本大小的高吞吐量向量去量化的灵活硬件

背景技术

[0001] 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与添加到具有应用的非线性的偏置的权重向量的内积。对于深度神经网络(DNN)(例如,如由示例性DNN模块所表达),神经元可以紧密地映射到人工神经元。

[0002] 在跨NN或DNN处理数据中,执行示例性处理操作的示例性神经元被要求以处理大量的数据以便应用各种数据处理/操纵操作,其可能影响导致对期望的状态处理目标不利的关键潜在因素的总体NN或DNN性能(例如,标识示例性输入数据中的对象和/或对象特性——图像、声音、地理坐标等)。通常,现有NN和DNN在执行这些各种操作时花费可避免的处理时间(例如,每秒浮动/固定点操作(每秒所执行的浮点运算次数,GFlops/s))和存储器空间(例如,每秒传送的字节数(每秒G字节数,GBytes/s))。特别地,当前实践要求在由人工神经元处理之前从协作存储器部件读取神经元权重值。通常,权重值可以被存储在通用存储器(诸如DRAM)中或者被高速缓存在快速本地存储器(诸如SRAM)中。利用通用存储器,要求时间和功率以读取权重值。利用本地存储器,高性能高速缓存存储器是昂贵的,并且通常是大小有限的。由于可避免的时间/功率被要求或者直接地从通用存储器或者间接地从本地高速缓存存储器读取权重值,因而当前实践缺乏完全优化NN/DNN的处理能力。

[0003] 克服当前实践的低效率的传统方法是降低权重数据的精度以降低所要求的存储器量。例如,32位浮点权重值可以减少到16位半精度值,其导致权重存储器要求中的50%节省。具有权重值的精度的大降低的问题是结果准确度的降低。

[0004] 更有利的NN/DNN将部署操作地允许更多权重值表示在给定量的本地存储器中的神经元权重值的向量量化的使用,其进而减少将权重值从主存储器加载到本地存储器高速缓存中的开销和/或降低所要求的本地存储器量。特别地,向量量化过程可以利用查找表将权重编码转换为权重数据。操作地,通过利用向量量化,整个权重团块可以操作地解译为可以在运行时期间解码的权重编码。

[0005] 更特别地,权重值的向量量化操作地可以将权重值的连续段转换为任意长度的向量(例如,2个权重值、4个权重值等)并且每个向量可以分配有索引值。在要求权重值的神经元计算操作的执行期间,索引被用于引用查找表中被用于计算的特定向量。由于单个索引被用于引用多个权重值,因而实现存储器空间的降低,而不必降低权重值的精度。

[0006] 所呈现的本文中做出的公开内容关于这些考虑和其他考虑。

发明内容

[0007] 本文所描述的技术提供使用神经元权重值的向量量化来减少示例性神经网络(NN)和/或深度神经网络(DNN)环境的存储器要求和处理周期。本文所描述的系统和方法的方面涉及机器/人工智能(MI)硬件架构。这样的架构和其实现可以被称为“NN”。在说明性实现中,示例性NN中的向量量化(VQ)的使用可能导致读取权重值的有效神经元性能的增加。

在说明性操作中，一个或多个索引可以被存储到可以利用快速查找表(物理或虚拟)表示权重值的一个或多个向量行。“权重”可以被认为是当处理一个或多个数据元素时由神经元处理器消耗的数值。权重值的可能格式可以为可以是有符号或无符号、字节、整数和/或浮点的任意位长度。由于索引而不是全部权重数据存储，因而存储器传送中的降低能够通过使用向量量化实现。

[0008] 在说明性实现中，权重值的一个或多个连续段可以操作地转换为任意长度的一个或多个向量，并且一个或多个向量中的每个向量可以分配索引。所生成的索引可以被存储在示例性VQ查找表中。在说明性操作中，在可以要求权重值的神经元计算操作的示例性执行期间，索引可以从可以表示包含一个或多个权重值的特定向量的所生成的查找表被取回。神经元计算操作可以被认为是由一个或多个神经元执行以根据所选择的操作(诸如卷积)处理输入数据或者完全连接以生成输出数据的一个或多个计算步骤。

[0009] VQ查找表的一个或多个行可以从协作存储器部件(诸如通用或者本地存储器部件)被读取。VQ查找表可以包括N行和M宽，并且可以由用于将索引快速转换为VQ行的协作快速权重查找硬件(FWLH)操作地使用。FWLH可以被认为是存在于NN中的硬件逻辑，其可操作以迅速地执行将权重值转换为VQ查找表的VQ行。在说明性实现中，行数N可以表示索引范围。例如，对于4096个向量行，可以要求12位的索引。作为每行权重值的数目的宽M可以包括任意值，其可以包括但不限于2的倍数，诸如2、4、8和16等。如果这样要求，则也可以操作地部署VQ查找表的较大宽度。

[0010] 在说明性操作中，在使用来自VQ查找表的说明性索引值执行取回向量时，如由向量表示的对应的权重值可以由神经元消耗作为NN的示例性数据处理功能的一部分。

[0011] 在说明性实现中，VQ查找表可以存储在NN的一个或多个协作硬件部件中，诸如寄存器、SRAM和DRAM。这样的硬件部件可以利用包含具有单个基本索引值的多个VQ表的固定存储器块或者虚拟存储器块实现以选择当前VQ查找表。

[0012] 在另一说明性实现中，一个或多个虚拟VQ查找表可以定义在具有基础索引值的单个物理VQ表定义。在说明性实现中，与本文所描述的向量量化过程有关的可选择的向量长度可以根据神经元层功能被利用，使得一个神经元层功能可以使用具有第一VQ长度(例如，2)的向量，其中另一神经元层功能可以使用具有第二VQ长度(例如，4)的向量，其中又一神经元层功能可以使用具有第三VQ长度(例如，16)的向量。说明性地，神经元层功能可以被认为是由示例性神经网络环境的一个或多个层执行的一个或多个操作。

[0013] 在说明性实现中，本文所描述的系统和方法可以被部署为“片上系统”，其中一个或多个NN被实例化，使得NN可以包含用于权重值的VQ查找表。

[0014] 在说明性操作中，在示例性运行时间处，可以执行内联向量去量化处理以确定可能导致神经元吞吐量的维护并且维持NN的优化性能的底层的(underlying)神经元权重值。在说明性操作中，如本文所描述的向量量化可以根据神经网络层而被启用/禁用。

[0015] 操作地，向量量化的使用可能导致NN的许多优化性能操作，包括但不限于：对于存储与神经元网络计算有关的神经元权重值的存储器存储要求的减少；当执行神经元层功能时所要求的存储器带宽的减少；当执行神经元层功能时所要求的时间的减少；以及利用神经元权重值数据的较高准确度实现传统神经元权重值存储器管理技术的期望性能水平所要求的本地高速缓存存储器量的减少。

[0016] 应当理解，虽然相对于系统进行描述，但是上文所描述的主题也可以实现为计算机控制的装置、计算机过程、计算系统、或者制品(诸如计算机可读介质和/或专用芯片集)。这些和各种其他特征从以下具体实施方式的读取和相关联的附图的查阅将是明显的。提供本发明内容以引入以在具体实施方式中下面进一步描述的简化形式的概念的选择。

[0017] 本发明内容不旨标识要求保护的主题的关键特征或基本特征，本发明内容也不旨在用于要求保护的主题的范围。此外，要求保护的主题不限于解决本公开的任何部分中陈述的任何或所有缺点的实现。

附图说明

[0018] 参考附图描述具体实施方式。在附图中，附图标记中的最左边的(一个或多个)数字标识附图标记首次出现的附图。不同的附图中的相同附图标记指示类似或者相同项。对多个项的个体项做出的参考可以使用具有字母序列的字母的附图标记以指代每个个体项。对项的通用参考可以使用没有字母序列的特定附图标记。

[0019] 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块图。

[0020] 图2图示了根据本文所描述的系统和方法的具有协作部件的示例性神经网络环境的块图。

[0021] 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块图。

[0022] 图4图示了示出可操作以跨说明性逻辑数据映射的一个或多个线的说明性n个滑动窗口的使用的说明性逻辑数据映射中表示的示例性输入数据的框图。

[0023] 图5图示了根据本文所描述的系统和方法的示出可操作以跨可操作以允许数据填充作为处理增强的说明性逻辑数据映射的一个或多个线的说明性n个滑动窗口的使用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6是根据本文所描述的系统和方法的示出可操作以执行神经元权重值的向量量化/去量化的示例性神经网络环境的各种部件的交互的框图。

[0025] 图7是说明性神经网络计算环境中的用于根据所要求的神经元权重值的向量量化/去量化处理数据的说明性过程的流程图。

[0026] 图8示出了用于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。

[0027] 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

[0028] 以下具体实施方式描述了用于使用神经元权重值的向量量化来减少示例性神经网络(NN)和/或深度神经网络(DNN)环境的存储器要求和处理周期的技术。本文所描述的系统和方法的方面涉及机器/人工智能(MI)硬件架构。这样的架构和其实现可以被称为“NN”。在说明性实现中，示例性NN中的向量量化(VQ)的使用可能导致读取权重值的有效神经元性能的增加。在说明性操作中，一个或多个索引可以被存储到可以利用快速查找表(物理或虚拟)表示权重值的一个或多个向量行。“权重”可以被认为是当处理一个或多个数据元素时由神经元处理器消耗的数值。权重值的可能格式可以为可以是有符号或无符号、字节、整数

和/或浮点的任意位长度。通过使用向量量化，与全部权重数据相比，索引被存储，可以实现存储器传送中的减少。

[0029] 在说明性实现中，权重值的一个或多个连续段可以操作地转换为任意长度的一个或多个向量，并且一个或多个向量中的每个向量可以分配索引。所生成的索引可以被存储在示例性VQ查找表中。在说明性操作中，在可以要求权重值的神经元计算操作的示例性执行期间，可以从可以表示包含一个或多个权重值的特定向量的所生成的查找表取回索引。神经元计算操作可以被认为是由一个或多个神经元执行以根据所选择的操作(诸如卷积)处理输入数据或者完全连接以生成输出数据的一个或多个计算步骤。

[0030] VQ查找表的一个或多个行可以从协作存储器部件(诸如通用或者本地存储器部件)读取。VQ查找表可以包括N行和M宽并且可以由用于将索引快速转换为VQ行的协作快速权重查找硬件(FWLH)操作地使用。FWLH可以被认为是存在于NN中的硬件逻辑，其可操作以迅速地执行将权重值转换为VQ查找表的VQ行。在说明性实现中，行数N可以表示索引范围。例如，对于4096向量行，可以要求12位的索引。作为每行权重值的数目的宽M可以包括任意值，其可以包括但不限于2的倍数，诸如2、4、8和16等。如果这样要求，则也可以操作地部署VQ查找表的较大宽度。

[0031] 在说明性操作中，在使用来自VQ查找表的说明性索引值执行取回向量时，如由向量表示的对应的权重值可以由神经元消耗作为NN的示例性数据处理功能的一部分。

[0032] 在说明性实现中，VQ查找表可以存储在NN的一个或多个协作硬件部件中，诸如寄存器、SRAM和DRAM。这样的硬件部件可以利用包含具有单个基本索引值的多个VQ表的固定存储器块或者虚拟存储器块实现以选择当前VQ查找表。

[0033] 在另一说明性实现中，一个或多个虚拟VQ查找表可以定义在具有基础索引值的单个物理VQ表定义。在说明性实现中，与本文所描述的向量量化过程有关的可选择的向量长度可以根据神经元层利用，使得一个神经元层功能可以使用具有第一VQ长度(例如，2)的向量，其中另一神经元层功能可以使用具有第二VQ长度(例如，4)的向量，其中又一神经元层功能可以使用具有第三VQ长度(例如，16)的向量。在说明性实现中，本文所描述的系统和方法可以被部署为“片上系统”，其中一个或多个NN被实例化，使得NN可以包含用于权重值的VQ查找表。说明性地，神经元层功能可以被认为是由示例性神经网络环境的一个或多个层执行的一个或多个操作。

[0034] 在说明性操作中，在示例性运行时，可以执行内联向量去量化处理以确定可能导致神经元吞吐量的维护并且维持NN的优化性能的底层神经元权重值。在说明性操作中，如本文所描述的向量量化可以根据神经网络层被启用/禁用。

[0035] 操作地，向量量化的使用可能导致NN的许多优化性能操作，包括但不限于：对于存储与神经元网络计算有关的神经元权重值的存储器存储要求的减少；当执行神经元层功能时所要求的存储器带宽的减少；当执行神经元层功能时所要求的时间的减少；以及利用神经元权重值数据的较高准确度实现传统神经元权重值存储器管理技术的期望性能水平要求的本地高速缓存存储器量的减少。

[0036] 应当理解，上文所描述的主题可以实现为计算机控制的装置、计算机过程、计算系统、或者制品(诸如计算机可读介质)。除了许多其他益处，本文中的技术改进相对于各种各样的计算资源的效率。例如，移位步长的确定可以降低对于执行许多复杂任务需要的许多

计算周期，诸如脸部识别、目标识别、图像生成等。

[0037] 另外，经改进的人类交互可以通过这样的任务的更准确并且更快速的完成的引入来实现。另外，移位步长的使用可以降低网络流量、降低功率消耗和存储器的使用。除本文所提到的技术效果之外的其他技术效果还可以从本文所公开的技术的实施例来实现。

[0038] 应当理解，虽然相对于系统描述，但是上文所描述的主题也可以实现为计算机控制的装置、计算机过程、计算系统、或者制品(诸如计算机可读介质和/或专用芯片集)。这些和各种其他特征从以下具体实施方式的读取和相关联的附图的查阅将是明显的。提供本发明内容以引入以在具体实施方式中下面进一步描述的简化形式的概念的选择。

[0039] 在人工神经网络中，神经元是被用于对大脑中的生物神经元建模的基本单元。人工神经元的模型可以包括输入向量与添加到具有应用的非线性的偏置的权重向量的内积。比较地，在示例性DDN模块(例如，图1的105)中，神经元紧密地映射到人工神经元。

[0040] 说明性地，DNN模块可以被认为是超标量处理器。操作地，其可以将一个或多个指令分派到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同步。DNN模块可以被分类为SIMD(单指令流、多数据流)架构。

[0041] 转到图1的示例性DNN环境100，DNN模块105具有存储器子系统，该存储器子系统具有唯一的L1和L2高速缓存结构。这些不是传统高速缓存，而是特别地被设计用于神经处理。为了方便起见，这些高速缓存结构已经采取反映其预期目的的名字。以示例的方式，L2高速缓存150可以说明性地维持具有以选择频率(例如，十六吉比特每秒(16Gbps))操作的高速私有接口的选择的存储容量(例如，一个兆字节(1MB))。L1高速缓存可以维持可以在内核与激活数据之间分割的选择的存储容量(例如，八千字节(8KB))。L1高速缓存可以被称为线缓冲器，并且L2高速缓存被称为BaSRAM。

[0042] DNN模块可以是仅回忆的神经网络并且以编程方式支持各种各样的网络结构。对于网络的训练可以在服务器群或者数据中心中离线执行。训练的结果是可以被称为或者权重或者内核的参数集。这些参数表示可以应用到输入的传递函数，其中，结果是分类或者语义标记的输出。

[0043] 在说明性操作中，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，只要呈现的数据以DNN可以在其上操作的均匀平面格式。

[0044] DNN模块在对应于神经网络的层的层描述符的列表上操作。说明性地，层描述符的列表可以由DNN模块视为指令。这些描述符可以从存储器预取到DNN模块中并且按次序执行。

[0045] 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；和2)操作描述符。存储器到存储器移动描述符可以被用于将至/自主存储器的数据移动至/自用于由操作描述符消耗的本地高速缓存。存储器到存储器移动描述符跟随与操作描述符不同的执行管线。对于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，而对于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层操作。

[0046] DNN的输出也是数据大块。输出可以可选地流动到本地高速缓存或者流动到主存储器。由于软件将允许，因而DNN模块可以提前预取数据。软件可以通过使用描述符之间的保护和设置依存性控制预取。防止具有依存性集的描述符进步直到已经满足相关性。

[0047] 现在转到图1，示例性神经网络环境100可以包括各种协作部件，包括DNN模块105、

高速缓存存储器125或150、低带宽结构110、桥接器部件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155和输入数据源102。进一步地，如所示出的，DNN模块105还可以包括许多部件，包括预取105(A)、DMA 105(B)、寄存器接口105(D)、负载/存储单元105(C)、层控制器105(D)、保存/恢复部件105(E)和神经元105(F)。操作地，示例性DNN环境100可以根据选择的规格处理数据，其中DNN模块执行如本文所描述的一个或多个功能。

[0048] 图2图示了可操作以采用直接线缓冲器220作为数据处理的一部分的示例性神经网络环境200。如所示出的，示例性神经网络环境200(在本文中还被称为计算设备或者计算设备环境)包括一个或多个操作控制器235，其与线缓冲器220协作以提供用于数据处理的一个或多个指令。线缓冲器220可以操作以通过外部结构230和结构215从协作外部存储器部件225接收数据以及操作以从(一个或多个)迭代器240(例如，基于硬件和/或虚拟化迭代器)接收一个或多个指令/命令(例如，从协作存储器部件读取数据的指令/命令和/或将从协作存储器部件加载的数据写入在线缓冲器中的指令)。此外，如图2中所示，示例性神经网络环境还可以包括快速权重查找硬件245(FWLH)，其可以操作地接收去量化被接收作为示例性码本的索引的列表的一个或多个神经元权重的请求。在说明性操作中，FWLH 245可以通过结构215从一个或多个协作存储器部件(210、225)接收神经元权重数据。FWLH 245可以处理神经元权重索引数据、去量化接收到的数据以产生可以操作地写入到线缓冲器220的相同数目的向量(即，码本条目)。

[0049] 操作地，线缓冲器220可以根据从一个或多个操作控制器235(在本文中也被称为“协作控制器部件235”)接收到的一个或多个指令根据选择的步长宽度移动数据。此外，线缓冲器220可以与(一个或多个)处理单元(例如，(一个或多个)神经元)协作以提供写入位移位数据以用于直接地或者间接地通过结构215进一步处理。神经网络环境结构可以是能够穿过各种数据的数据总线。直接线缓冲器可以被认为是能够根据一个或多个接收到的指令读取和写入数据和数据元素的存储器部件。

[0050] 在说明性操作中，示例性神经网络环境200可以根据图7中所描述的过程操作地处理数据。特定于图2中所描述的部件，这些部件仅是说明性的，因为本领域普通技术人员将理解到，图6和7中所描述的处理也将由除了图2中所图示的部件之外的其他部件执行。

[0051] 此外，如图2中所示出的，示例性神经网络环境可以可选地包括一个或多个迭代器(例如，基于硬件和/或虚拟化迭代器)(如由虚线指示的)，其可以说明性地操作以迭代输入数据(未示出)以用于由一个或多个神经元处理器205处理。本领域技术人员将理解到，由于由本文所公开的系统和方法描述的发明构思在没有任何迭代器的情况下操作的示例性神经网络环境200中操作，因而示例性一个或多个迭代器的这样的可选包括仅是说明性的。

[0052] 图3图示了用于示例性输入数据的示例逻辑数据映射300。如所示，数据305可以被表示为具有包括通道数310、高度315和宽度320的某个维度340(例如，使得总体上采取的数据维度可以定义数据量)的数据。根据本文所描述的系统和方法，数据305可以被分配并且准备用于由协作n个神经元330处理，使得第一部分a可以传递到第一神经元，第二部分b可以传递到第二神经元等，直到n个部分被传递到n个神经元。

[0053] 在说明性操作中，可以基于由示例性神经网络环境(例如，图2的200)的协作控制器部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的部分。进一步

地，如所示出的，输入数据部分a、b、c和d可以使用由示例性神经网络环境(例如，图2的200)的协作操作控制器部件(235)提供的一个或多个初始化参数寻址到物理存储器325。

[0054] 图4图示了示例性输入数据(未示出)的示例性逻辑数据图400。示例性逻辑数据图400包括第一线410(利用斜线标记图示的)和第二线420(通过虚线图示的)。每个地图线可以包括若干滑动窗口(例如，用于第一线410的430、440和450和用于第二线420的460、470和480)。此外，如所示出的，逻辑数据图400示出滑动窗口的能力以跨输入数据的数据维度边界(例如，跨第一线410和第二线420)。由于更多数据可以更高效地准备用于由协作神经网络处理部件(例如，图2的205)随后处理，因而这样的能力允许增加的性能。

[0055] 图5类似于图4并且被呈现以描述本文所描述的系统和方法的能力以允许使用填充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。如所示出的，逻辑数据图500(未示出的示例性输入数据的)可以包括跨一个或多个线(例如，510和520)的各种滑动窗口(530、540、550、560、570和580)。此外，逻辑数据图500还可以包括填充580。

[0056] 在说明性操作中，在示例性神经网络环境(图1的100或者图2的200)的运行时处，填充580可以被动态地添加。图2的操作控制器235可以指定待使用在输入数据(例如，大块)的图3中所示的维度340中的每个维度340上的填充量(例如，使得共同取得的维度可以被认为是数据量)，并且神经网络环境(例如，迭代器控制器指令)可以操作地构建数据量，好像填充物理地存在于存储器中。默认值还可以在其中添加填充的迭代器输出位置中由示例性神经网络环境(例如，迭代器控制器指令)生成。

[0057] 图6是示出可操作以执行神经元权重值的向量量化/去量化的示例性神经网络环境600的各种部件的交互的示图。如图6中所示，示例性神经网络环境600可以包括示例性神经处理器605(例如，图1的100)。神经处理器605还可以包括快速权重查找硬件610，其操作地处理权重索引数据615以及来自示例性快速权重查找表625的数据以取回/去量化用于由示例性神经元630消耗的神经元权重值620。进一步地，如所示出的，快速权重查找表625可以包括若干行625(a)、625(b)和625(c)等。

[0058] 在说明性操作中，一个或多个索引可以被存储到可以利用快速查找表625表示权重值的一个或多个向量行(诸如625(a)、625(b)、625(c))，快速查找表可以说明性地是物理硬件表或者以软件创建的虚拟化表。在说明性实现中，权重值的一个或多个连续段可以操作地转换为任意长度的一个或多个向量并且一个或多个向量中的每个向量可以被分配索引。所生成的索引可以被存储在示例性VQ查找表625中。

[0059] VQ查找表625的一个或多个行可以从协作存储器部件(诸如通用或者本地存储器部件)读取。VQ查找表可以包括N行和M宽并且可以由用于将索引快速转换为VQ行的协作快速权重查找硬件(FWLH)610操作地使用。FWLH 610可以被认为是存在于NN中的硬件逻辑，其可操作以迅速地执行将权重值转换为VQ查找表的VQ行625(a)、625(b)和615(c)。

[0060] 在说明性实现中，行数N可以表示索引范围。例如，对于4096个向量行，可以要求12位的索引。作为每行权重值的数目的宽M可以包括任意值，其可以包括但不限于2的倍数，诸如2、4、8和16等。如果这样要求，则也可以操作地部署VQ查找表的较大宽度。

[0061] 在说明性操作中，在使用来自VQ查找表的说明性索引值执行向量的取回时，如由向量表示的对应的权重值620可以根据驻留在FWLH 610的物理存储器部件640上的一个或多个码本由FWLH 610去量化，并且可以由神经元630消耗作为NN的示例性数据处理功能的

一部分。

[0062] 如图6中所示，以图示的方式，快速权重查找表625可以存储由FWLH 610可操作并且具有用于在利用选择的码本的存储向量的去量化中使用的一个或多个说明性动态物理存储器配置650、660和670的物理存储器640的示例性阵列的配置数据。码本被表示为被用于量化数据的向量的列表。该列表中的每个向量的位置(索引)可以操作地表示量化向量。超过一个码本可以被用于实现期望的去量化速率。在示例性情况下，在如图6中所示的第二情况660中，采用4个码本和去量化4个索引同时地有效地16(每行4个条目*4个码本)权重项可以去量化以实现期望的去量化速率。

[0063] 在说明性操作中，如果这样的向量量化/去量化处理被激活用于这样的一个或多个处理层，则一个或多个动态物理存储器配置650、660和670可以由示例性神经网络环境的处理层中的一个或多个处理层来采用。

[0064] 在说明性操作中，说明性动态物理存储器配置可以通过设置示例性配置寄存器(未示出)来配置，其可以是FWLH 610的驻留部件以允许用于示例性神经网络环境的处理层中的一个或多个的这些示例性物理存储器配置650、660和670之一的使用。操作地，动态物理存储器可以由FWLH 610被用于在其上加载用作向量去量化处理的一部分的示例性码本。

[0065] 在说明性实现中，协作物理存储器中的每一个可以维持码本的副本。操作地，当示例性码字被加载到物理存储器中时，码本的单个副本可以从协作存储器部件(例如，DRAM)被复制到码本存储器，并且FWLH可以操作地自动地将单个码本存储器数据写入到其他协作物理存储器中。

[0066] 将理解到，虽然动态物理存储器地址在图6中被描述为具有示例性位数和条目，但是由于本文所描述的发明构思预期其他备选位数和条目计数的使用，因而这样的示例仅是说明性的。

[0067] 图7是利用神经元权重值的向量量化增强用于NN/DNN环境的性能的说明性过程700的流程图。如所示出的，处理在块705处开始，其中一个或多个初始化参数从神经网络环境的协作部件(例如，操作控制器)被接收，其中一个或多个初始化参数可以包括代表用于包括神经元权重值的一个或多个连续段的输入数据的维度的数据。处理然后转到块710，其中神经元权重值的一个或多个连续段被转换为任意长度的一个或多个向量并且被分配所生成的索引值。

[0068] 在块715处，经转换的一个或多个向量然后被存储在向量量化查找表的一个或多个行中。处理然后转到块720，其中向量的一个或多个行操作地使用步骤710的生成的索引值中的一个或多个被取回并且去量化以获得底层神经元权重值。在块725处，所取回的权重值然后说明性地由神经网络环境的示例性神经处理器部件的一个或多个神经元来消耗。

[0069] 然后在块730处执行检查以确定是否存在待处理的附加的输入数据(即，作为迭代操作的一部分)。如果不存在附加输入数据，则在块735处处理终止。然而，如果附加输入数据要求迭代操作，则处理然后返回块710并且从那里进行。

[0070] 图8中所图示的计算机架构800包括中央处理单元80(“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。基本输入/输出系统(包含帮助在计算机架构800内的元件之间传递信息(诸如在启动期间)的基本例程)被存储在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。

[0071] 大容量存储设备812通过被连接到总线810的大容量存储控制器(未示出)而被连接到CPU 802。大容量存储设备812和其相关联的计算机可读介质为架构800提供非易失性存储装置。虽然本文包含的计算机可读介质的描述指代大容量存储设备(诸如固态驱动器、硬盘或者CD-ROM驱动器),但是本领域技术人员应当理解,计算机可读介质可以是可以由计算机架构800访问的任何可用计算机存储介质或者通信介质。

[0072] 通信介质包括计算机可读指令、数据结构、程序模块或者调制数据信号(诸如载波或者其他传输机制)中的其他数据,并且包括任何递送介质。术语“调制数据信号”意味着具有以将信息编码在信号中的方式设定或改变的其特性中的一个或多个特性的信号。以示例而非限制的方式,通信介质包括有线介质(诸如有线网络或直接有线连接)和无线介质(诸如声学、RF、红外线和其他无线介质)。上文的任何组合还应当被包括在计算机可读介质的范围内。

[0073] 以示例而非限制的方式,计算机存储介质可以包括在用于信息(诸如计算机可读指令、数据结构、程序模块或者其他数据)的存储的任何方法或技术中实现的易失性和非易失性、可移除和不可移除的介质。例如,计算机介质包括但不限于RAM、ROM、EPROM、EEPROM、闪速存储器或者其他固态存储器技术、CD-ROM、数字通用光盘(“DVD”)、HD-DVD、BLU-RAY、或者其他光学存储装置、磁带盒、磁带、磁盘存储装置或者其他磁性存储设备或者可以被用于存储期望信息并且可以由计算机架构800访问的期望信息的其他介质。出于权利要求的目的,短语“计算机存储介质”、“计算机可读存储介质”和其变型不包括波、信号和/或其他暂态和/或无线通信介质本身。

[0074] 根据各种技术,计算机架构800可以使用通过网络820和/或另一网络(未示出)对远程计算机的逻辑连接在联网环境中操作。计算机架构800可以通过连接到总线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还可以被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还可以包括用于接收并且处理来自若干其他设备(包括键盘、鼠标或者电子笔(未示出在图8中))的输入的一个或多个输入/输出控制器818。类似地,输入/输出控制器818可以向显示屏、打印机、或者其他类型的输出设备(也未示出在图8中)提供输出。还应当理解,经由通过网络接口单元816对网络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。

[0075] 应当理解,当加载到CPU 802和/或DNN模块105中并且执行时,本文所描述的软件部件将CPU 802和/或DNN模块105和总体计算机架构800从通用计算系统转换为定制为促进本文中呈现的功能的专用计算系统。CPU 802和/或DNN模块105可以根据任何数目的晶体管或者其他分立电路元件和/或芯片集构建,其可以单独地或者共同地假定任何数目的状态。更特别地,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可以操作为有限状态机。这些计算机可执行指令可以通过指定CPU 802在状态之间如何转换来转换CPU 802,从而转换晶体管或者构成CPU 802的其他分立硬件元件。

[0076] 编码本文中呈现的软件模块可以转换本文中呈现的计算机可读介质的物理结构。在本描述的不同实现中,物理结构的特定转换可以取决于各种因素。这样的因素的示例可以包括但不限于用来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要或次要存储装置等。例如,如果计算机可读介质被实现为基于半导体的存储器,则本文中所公开

的软件可以通过转换半导体存储器的物理状态编码在计算机可读介质上。例如，软件可以转换晶体管、电容器、或者构成半导体存储器的其他分立电路元件的状态。软件还可以转换这样的部件的物理状态以便将数据存储在其上。

[0077] 作为另一示例，本文所公开的计算机可读介质可以使用磁性或光学技术来实现。在这样的实现中，当软件被编码在其中时，本文呈现的软件可以转换磁性或光学介质的物理状态。这些转换可以包括改变给定磁性介质内的特定位置的磁性特性。这些转换还可以包括改变给定光学介质内的特定位置的物理特征或特性以改变那些位置的光学特性。在不脱离本描述的范围和精神的情况下，物理介质的其他转换是可能的，其中前述示例仅被提供以促进本讨论。

[0078] 鉴于上文，应当理解，许多类型的物理转换在计算机架构800中发生以便存储并且执行本文呈现的软件组件。还应当理解，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、个人数字助理和本领域中已知的其他类型的计算设备。还应预期到，计算机架构800可以不包括图8中所示的所有部件，可以包括未明确示出在图8中的其他部件，或者可以利用与图8中所示的架构完全不同的架构。

[0079] 上文所描述的计算系统800可以被部署作为计算机网络的一部分。一般而言，对于计算环境的以上描述用于部署在网络环境中的服务器计算机和客户端计算机二者。

[0080] 图9图示了可以采用本文中所描述的装置和方法的具有经由通信网络与客户端计算机通信的服务器的示例性说明性联网计算环境900。如图9中所示，(一个或多个)服务器905可以经由通信网络820(其可以是固定线或无线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、或者其他通信网络中的任一项或组合)与若干客户端计算环境相互连接，诸如平板个人计算机910、移动电话915、电话920、(一个或多个)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，Apple Watch、Samsung、FitBit等)930和智能电话935。在通信网络820是因特网的网络环境中，例如，(一个或多个)服务器905可以是可操作以经由若干已知协议(诸如超文本传送协议(HTTP)、文件传送协议(FTP)、简单对象访问协议(SOAP)、或者无线应用协议(WAP))中的任一项处理并且传递至和自客户端计算设备801、910、915、920、925、930和935的数据的专用计算环境服务器。此外，联网计算环境900可以利用各种数据安全协议，诸如安全套接字层(SSL)或良好隐私(PGP)。客户端计算环境801、810、815、820、825、830和835中的每一个可以装备有可操作以支持一个或多个计算应用或者终端会话的计算环境805，诸如网络浏览器(未示出)、或其他图形用户接口(未示出)、或移动桌面环境(未示出)，以获得对(一个或多个)服务器计算环境905的访问。

[0081] (一个或多个)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(一个或多个)客户端计算环境上运行的计算应用交互以获得期望的数据和/或计算应用。数据和/或计算应用可以被存储在(一个或多个)服务器计算环境905上并且在过示例性通信网络820之上通过客户端计算环境901、910、915、920、925、930和935传递到协作用户。参与用户(未示出)可以请求全部或者部分安置在(一个或多个)服务器计算环境905上的特定用户和应用的访问。这些数据可以在客户端计算环境801、910、915、920、925、930、935与(一个或多个)服务器计算环境905之间传递以用于处理和存储。(一个或多个)服务器计算环境905可以托管用

于数据和应用的生成、认证、加密和通信的计算应用、过程和小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)协作以实现应用/数据事务。

[0082] 示例条款

[0083] 可以鉴于以下条款考虑本文中呈现的公开内容。

[0084] 示例条款A,一种用于神经网络环境(100)中的增强数据处理的系统,系统包括至少一个处理器、至少一个存储器部件、以及与至少一个处理器通信的至少一个存储器,至少一个存储器具有存储在其上的计算机可读指令,其当由至少一个处理器执行时,使得至少一个处理器:

[0085] 从神经网络环境的协作控制器部件接收一个或多个初始化参数,初始化参数包括代表待由神经网络环境处理的数据的维度的数据以及代表一个或多个向量量化索引值的数据,一个或多个索引值代表被存储在至少一个存储器部件上的一个或多个向量,一个或多个向量包括代表一个或多个神经元权重值的一个或多个连续段的数据,利用一个或多个向量量化索引值(615)从至少一个存储器部件取回代表一个或多个神经元权重值的一个或多个向量,去量化所取回的一个或多个向量以取回底层一个或多个神经元权重值,并且传递一个或多个神经元权重值(620)以用于由神经网络环境的一个或多个处理部件(630)处理。

[0086] 示例条款B,根据示例条款A的系统,其中一个或多个向量被存储在驻留在至少一个存储器部件上的快速查找表中。

[0087] 示例条款C,根据示例条款A和B所述的系统,其中一个或多个向量具有任意长度。

[0088] 示例条款D,根据示例条款A到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从快速查找表的一个或多个行取回一个或多个向量。

[0089] 示例条款E,根据示例条款A到D所述的系统,其中一个或多个向量的向量长度对于神经网络环境的神经元层中的每一个是可选择的。

[0090] 示例条款F,根据示例条款A到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执行用于神经网络环境的神经元层中的所选择的一个或多个神经元层的一个或多个神经元权重值的向量去量化。

[0091] 示例条款G,根据示例条款A到F所述的系统,其中计算机可读指令还包括可操作以执行对被存储在快速查找表上的向量的快速查找的一个或多个硬件部件。

[0092] 示例条款H,一种计算机实现的方法,包括:从神经网络环境的协作控制器部件接收一个或多个初始化参数,初始化参数包括代表待由神经网络环境处理的数据的维度的数据以及代表一个或多个向量量化索引值的数据,一个或多个索引值代表被存储在至少一个存储器部件上的一个或多个向量,一个或多个向量包括代表一个或多个神经元权重值的一个或多个连续段的数据,一个或多个向量由神经网络环境的处理器生成,利用一个或多个向量量化索引值从至少一个存储器部件取回代表一个或多个神经元权重值的一个或多个向量,一个或多个向量操作地被存储在快速查找表上,去量化所取回的一个或多个向量以取回底层一个或多个神经元权重值,并且传递一个或多个神经元权重值以用于由神经网络环境的一个或多个处理部件处理。

[0093] 示例条款H,根据示例条款G所述的计算机实现的方法,还包括对于由神经网络环

境的一个或多个协作硬件部件对所取回的一个或多个向量的进行内联去量化以获得一个或多个神经元权重值。

[0094] 示例条款I,根据示例条款G和H所述的计算机实现的方法,还包括利用用于所生成的一个或多个向量的协作存储器部件中的基索引以生成虚拟化的一个或多个快速查找表。

[0095] 示例条款J,根据示例G到I所述的计算机实现的方法,还包括生成用于神经网络环境的一个或多个神经元层的一个或多个向量。

[0096] 示例条款K,根据示例条款G到J所述的计算机实现的方法,还包括将一个或多个向量存储在快速查找表的一个或多个行中。

[0097] 示例条款L,根据示例条款G到K所述的计算机实现的方法,还包括生成任意长度的一个或多个向量。

[0098] 示例条款M,根据示例条款G到L所述的计算机实现的方法,还包括选择向量长度以用于神经网络环境的神经元层中的每个神经元层的一个或多个向量的生成。

[0099] 示例条款N,根据示例条款G到M所述的计算机实现的方法,还包括将所生成的一个或多个向量存储在本地存储器部件中。

[0100] 示例条款O,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部件接收一个或多个初始化参数,初始化参数包括代表待由神经网络环境处理的数据的维度的数据以及代表一个或多个向量量化索引值的数据,一个或多个索引值代表被存储在至少一个存储器部件上的一个或多个向量,一个或多个向量包括代表一个或多个神经元权重值的一个或多个连续段的数据,利用一个或多个向量量化索引值从至少一个存储器部件取回代表一个或多个神经元权重值的一个或多个向量,去量化取回的一个或多个向量以取回底层的一个或多个神经元权重值,并且传递一个或多个神经元权重值(620)以用于由神经网络环境的一个或多个处理部件(630)处理。

[0101] 示例条款P,根据示例条款O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:将一个或多个向量存储在一个或多个快速查找表中。

[0102] 示例条款Q,根据示例条款O和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:选择一个或多个向量的长度。

[0103] 示例条款R,根据示例条款O到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:不取回神经网络环境的神经元层的向量。

[0104] 示例条款S,根据示例条款O到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:执行一个或多个向量的内联去量化以取回底层一个或多个神经元权重值。

[0105] 示例条款T,根据示例条款O到S所述的计算机可读介质,其中存储器部件与物理传感器协作,物理传感器能够产生包括音频数据、视频数据、触觉感觉数据和其他数据的输入数据以用于由一个或多个协作处理单元后续处理。

[0106] 示例条款U,根据示例条款O到T所述的计算机可读介质,其中协作处理单元与可操作以接收包括音频数据、视频数据、触觉感觉数据和其他数据的人类交互处理输入数据的一个或多个输出物理部件电子地协作。

[0107] 结论

[0108] 最后，虽然已经以特定于结构特征和/或方法动作的语言描述各个实施例，但是将理解到，所附表示中定义的主题不必限于上文所描述的特定特征或动作。相反，特定特征或动作被公开为实现要求保护的主题的示例形式。

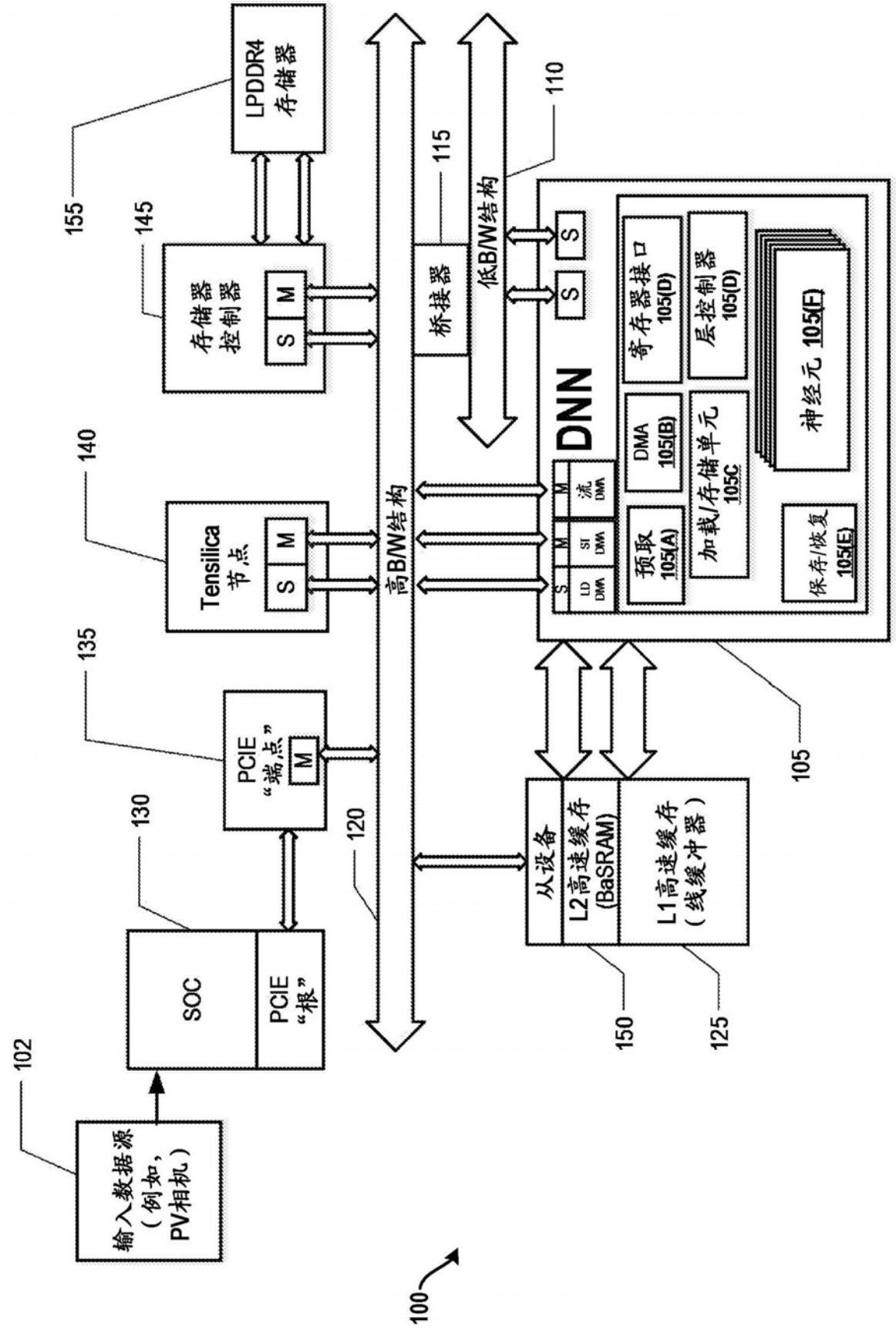

LPDDR4
存储器
155
145
存储器
控制器
S
M
S
M
桥接器
115
低B/W结构
110
S
S
140
Tensilica
节点
S
M
高B/W结构
DNN
寄存器接口
105(D)
层控制器
105(D)
神经元 105(F)
M
流
DMA
DMA
105(B)
135
PCIE
“端点”
M
M
SI
DMA
预取
105(A)
加载/存储单元
105C
S
LD
DMA
保存/恢复
105(E)
105
130
120
SOC
PCIE
“根”
从设备
L2高速缓存
(BaSRAM)
L1高速缓存
(线缓冲器)
150
125
102
输入数据源
(例如,
PV相机)
100

图1

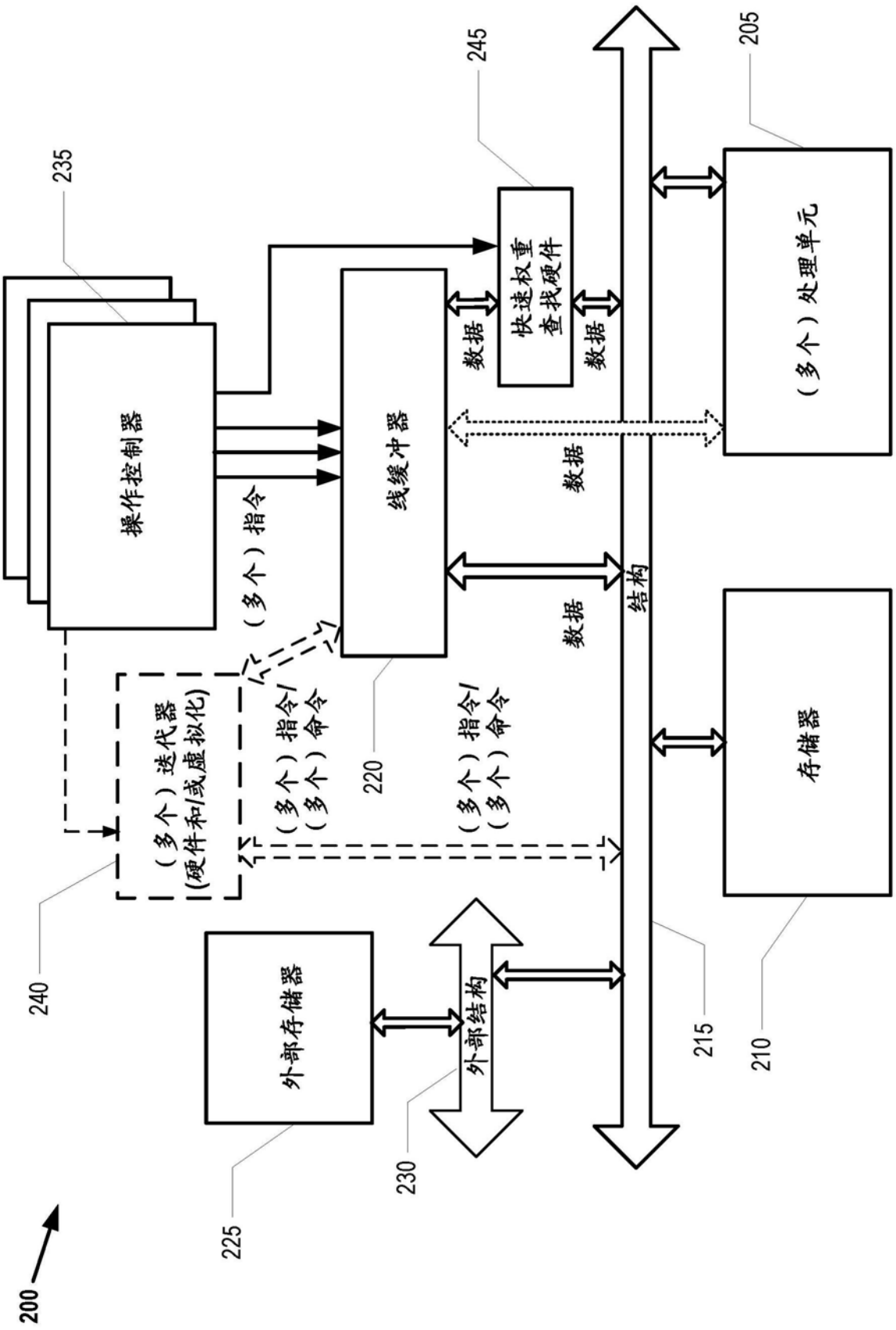

200
235
操作控制器
（多个）指令
线缓冲器
220
245
快速权重查找硬件
数据
数据
205
（多个）处理单元
数据
结构
（多个）迭代器
(硬件和/或虚拟化)
240
（多个）指令/
（多个）命令
（多个）指令/
（多个）命令
存储器
210
外部存储器
225
外部结构
230
215

图2

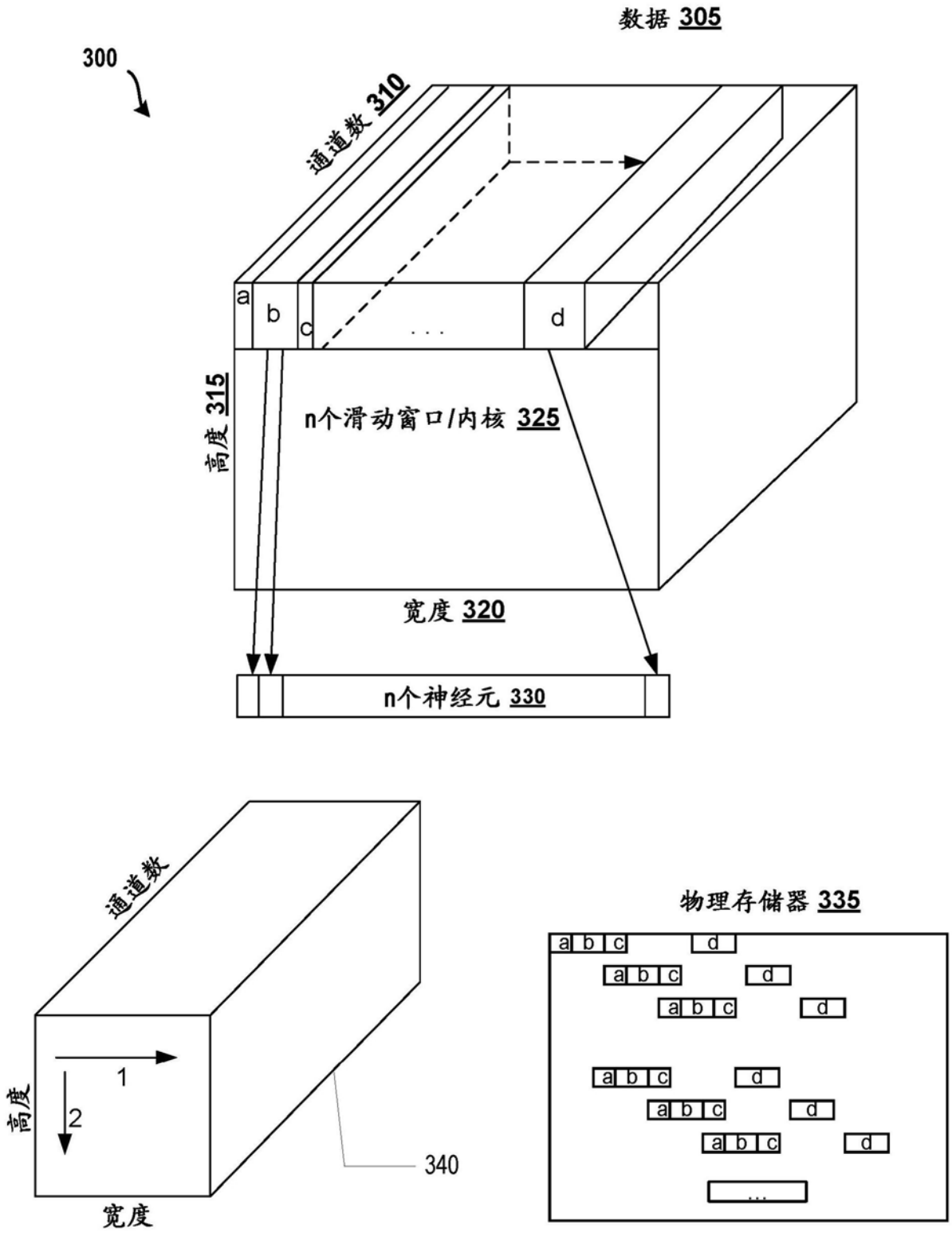

数据 305
300
通道数 310
a
b
c
. . .
d
高度 315
n个滑动窗口/内核 325
宽度 320
n个神经元 330
通道数
高度
1
2
宽度
340
物理存储器 335
a b c
d
...

图3

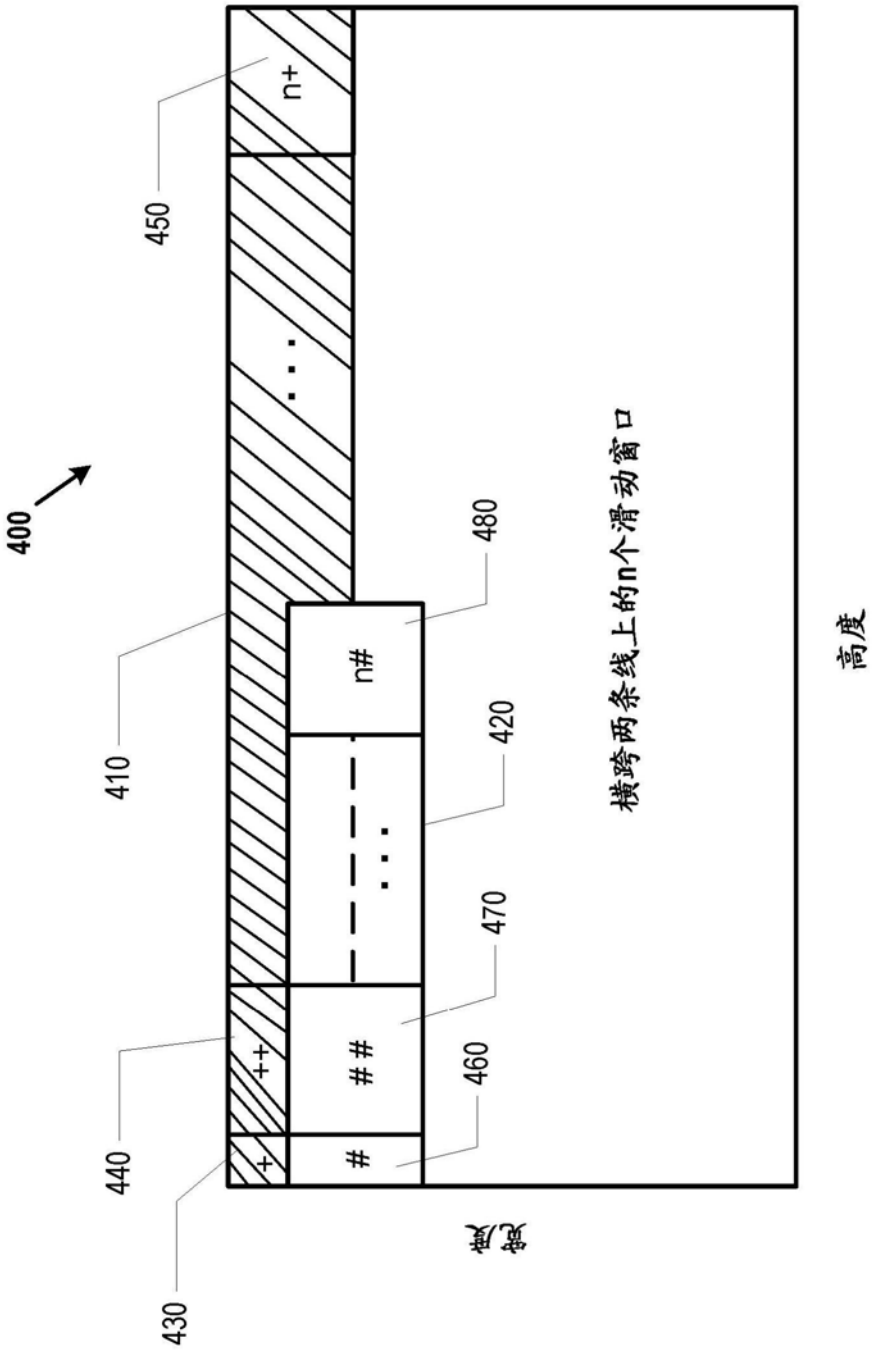


图4

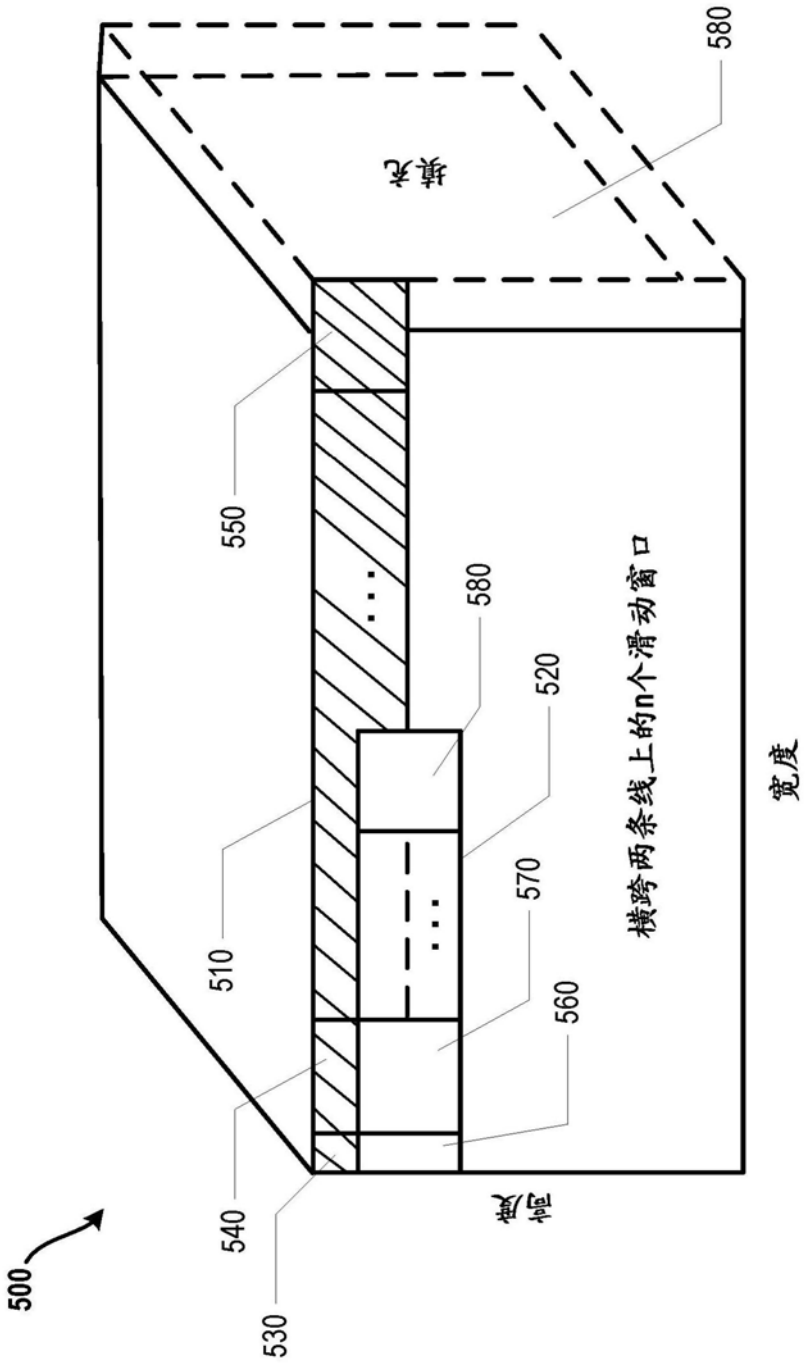


图5

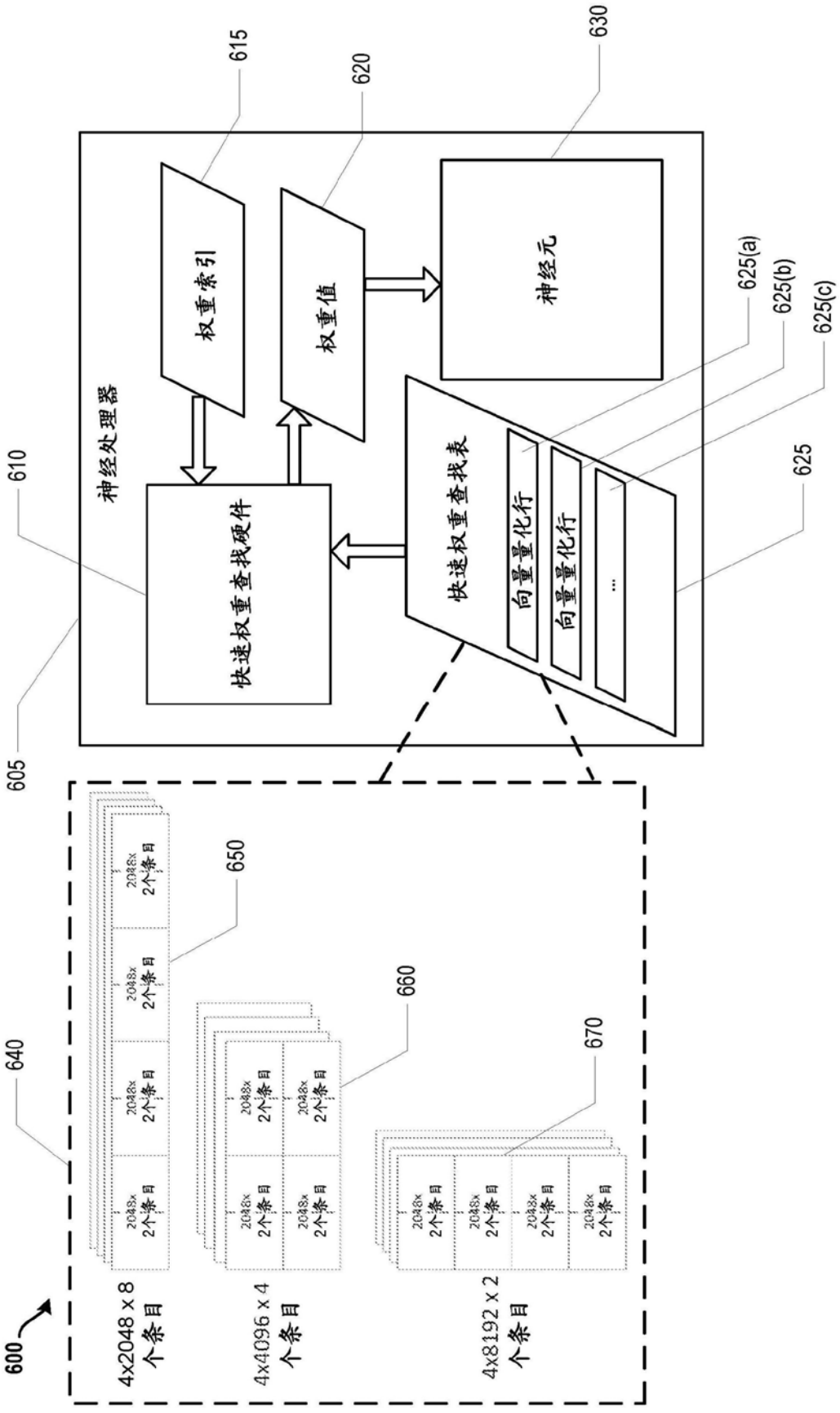

615
620
630
权重索引
权重值
神经元
神经处理器
610
快速权重查找硬件
快速权重查找表
625(a)
625(b)
625(c)
625
向量量化行
向量量化行
...
605
650
660
670
640
2048x 2个条目
4x2048 x 8 个条目
4x4096 x 4 个条目
4x8192 x 2 个条目
600

图6

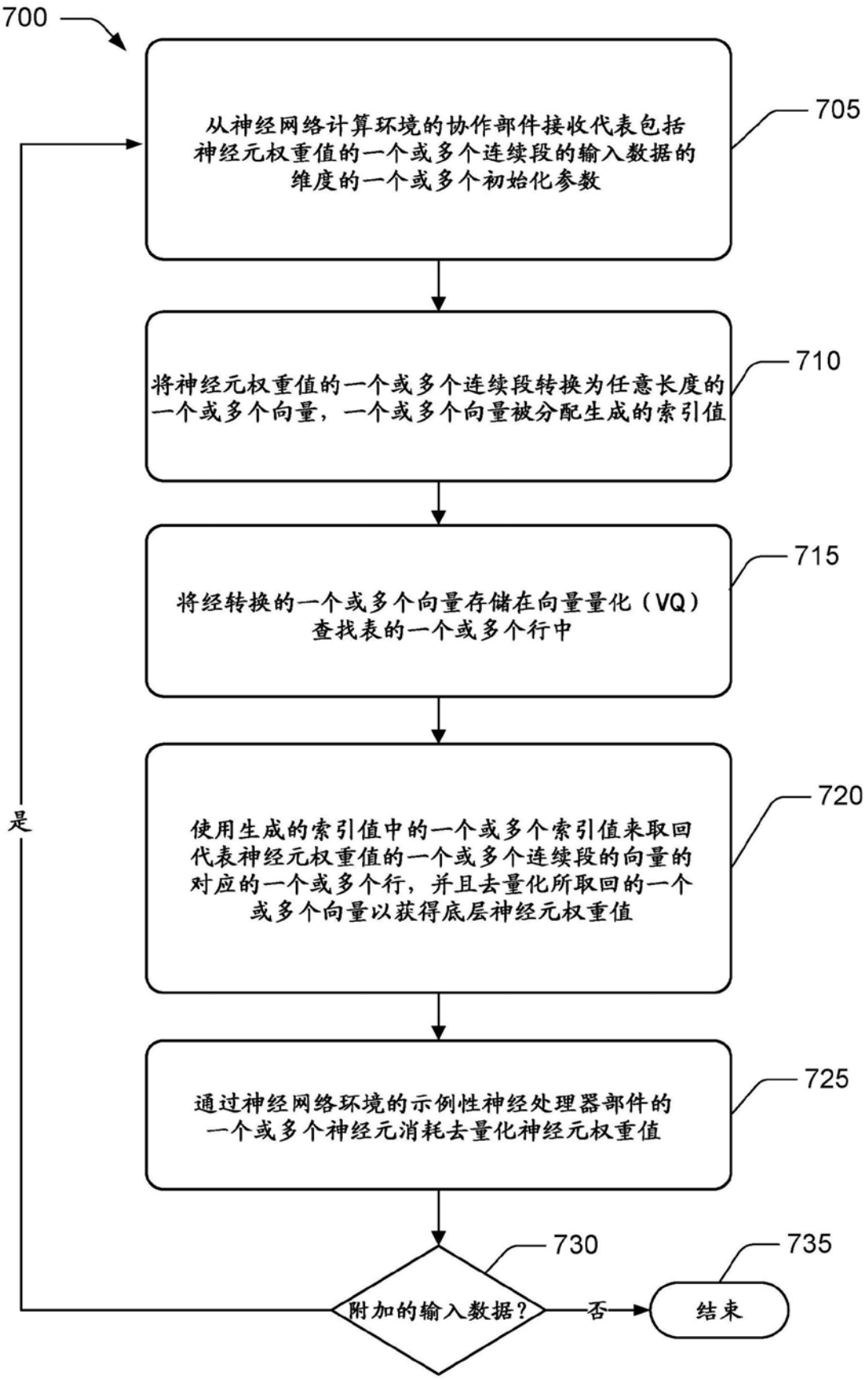

700
从神经网络计算环境的协作部件接收代表包括神经元权重值的一个或多个连续段的输入数据的维度的一个或多个初始化参数
705
将神经元权重值的一个或多个连续段转换为任意长度的一个或多个向量，一个或多个向量被分配生成的索引值
710
将经转换的一个或多个向量存储在向量量化（VQ）查找表的一个或多个行中
715
使用生成的索引值中的一个或多个索引值来取回代表神经元权重值的一个或多个连续段的向量的对应的一个或多个行，并且去量化所取回的一个或多个向量以获得底层神经元权重值
720
通过神经网络环境的示例性神经处理器部件的一个或多个神经元消耗去量化神经元权重值
725
附加的输入数据？
730
是
否
结束
735

图7

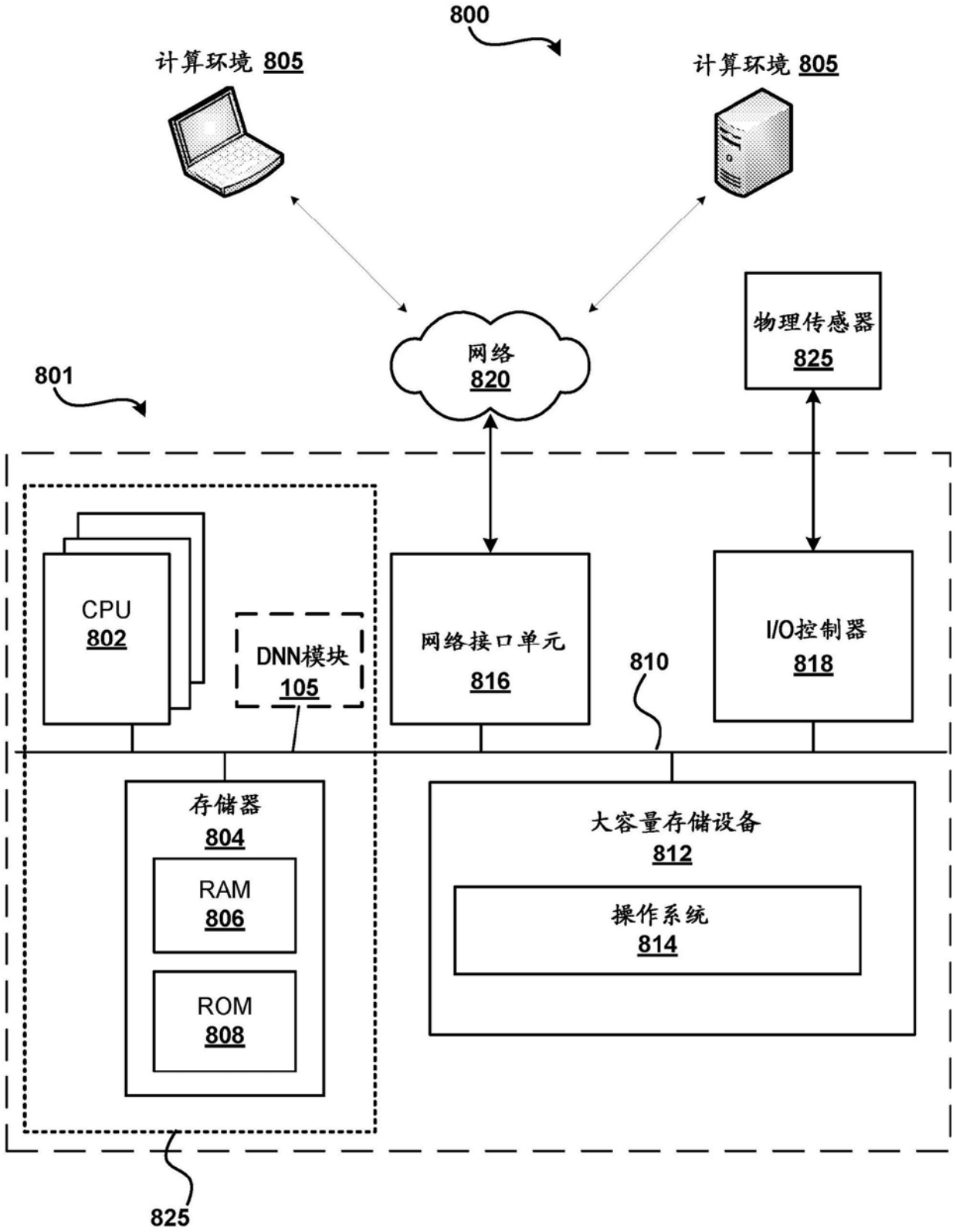

800
计算环境 805
计算环境 805
网络
820
物理传感器
825
801
CPU
802
DNN模块
105
网络接口单元
816
810
I/O控制器
818
存储器
804
RAM
806
ROM
808
大容量存储设备
812
操作系统
814
825

图8

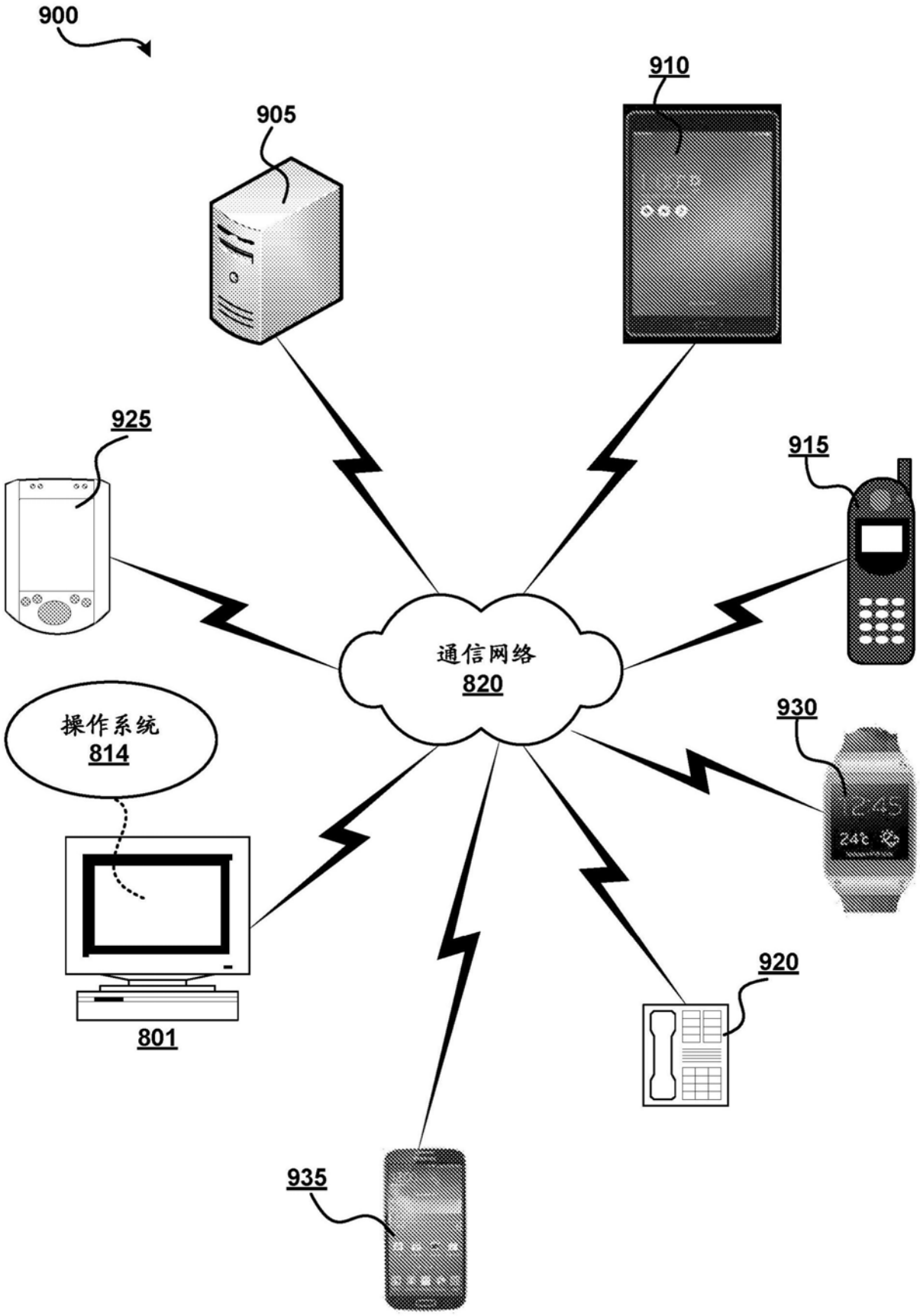


图9