(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119452368 A
(43)申请公布日 2025.02.14

(21)申请号 202380044796.8
(22)申请日 2023.05.16
(30)优先权数据
17/848,679 2022.06.24 US
(85)PCT国际申请进入国家阶段日
2024.12.03
(86)PCT国际申请的申请数据
PCT/US2023/022447 2023.05.16
(87)PCT国际申请的公布数据
WO2023/249754 EN 2023.12.28
(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 D·R·怕特尔 左薇 余原
(74)专利代理机构 北京市金杜律师事务所 11256
专利代理师 丁君军
(51)Int.Cl.
G06N 3/045(2023.01)
G06N 3/063(2023.01)
G06N 3/08(2023.01)

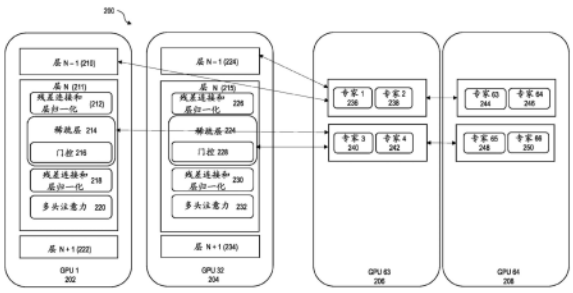
权利要求书2页 说明书12页 附图12页

(54)发明名称

用于分布专用混合专家机器学习模型的层的系统和方法

(57)摘要

一些所公开的实施例涉及具有不同加速器的计算系统,使得第一加速器集合具有比第二加速器集合更大的存储器能力,而第二加速器集合具有比第一加速器集合更大的处理能力。具有不同密集层和稀疏层的机器学习模型分布在不同的加速器上,使得密集层分布在从第一加速器集合所选择的一个或多个加速器上,并且稀疏层分布在第二加速器集合中的一个或多个加速器上。



CN 119452368 A

1.一种计算系统,包括:

一个或多个硬件处理器;

一个或多个硬件存储设备,所述一个或多个硬件存储设备被配置为存储计算机可执行指令,所述计算机可执行指令由所述一个或多个硬件处理器可执行,以用于管理混合专家(MOE)模型的稀疏层和密集层在多个加速器上的分布;

多个加速器,所述多个加速器包括第一加速器集合和第二加速器集合,所述第二加速器集合中的所述加速器具有比所述第一加速器集合中的加速器更大的存储器容量,并且所述第一加速器集合中的加速器具有比所述第二加速器集合中的加速器更大的处理能力;以及

机器学习模型,所述机器学习模型包括多个密集层和多个稀疏层,所述机器学习模型分布在所述多个加速器上,使得(i)所述多个密集层分布在从所述第一加速器集合所选择的一个或多个加速器上,(ii)所述多个稀疏层分布在从所述第二加速器集合所选择的一个或多个加速器上,以及(iii)存储所述机器学习模型的一个或多个稀疏层的至少一个加速器被配置为从存储所述机器学习模型的密集层的一个或多个加速器接收多个输入。

2.根据权利要求1所述的计算系统,其中被包括在所述多个稀疏层中的每个稀疏层还包括一个或多个专家,使得所述机器学习模型被配置为混合专家机器学习模型。

3.根据权利要求1所述的计算系统,其中每个专家在唯一任务上被训练。

4.根据权利要求1所述的计算系统,其中一个或多个加速器是硬件加速器。

5.根据权利要求4所述的计算系统,其中一个或多个加速器被配置为图形单元处理器。

6.根据权利要求1所述的计算系统,其中所述第一加速器集合被配置为密集加速器,所述密集加速器具有用于处理密集层的能力,所述能力比被配置为稀疏加速器的所述第二加速器集合大至少10%。

7.根据权利要求1所述的计算系统,其中所述第二加速器集合被配置为稀疏加速器,所述稀疏加速器在处理稀疏数据方面比被配置为密集加速器的所述第一加速器集合效率高至少10%。

8.根据权利要求1所述的计算系统,其中所述多个稀疏层与所述多个密集层交错。

9.一种用于管理机器学习模型的分布的方法,所述机器学习模型包括在计算系统上的混合专家(MOE),所述方法包括:

访问包括多个密集层和多个稀疏层的机器学习模型;

标识被包括在所述机器学习模型中的若干密集层和若干专家;

访问包括多个加速器的计算系统,其中至少一些加速器具有比其他加速器更大的存储器容量并且至少一些加速器具有比其他加速器更大的处理能力;

标识哪些加速器是可用的;

确定所述可用加速器中的哪些加速器被配置为支持密集层以及哪些加速器被配置为支持专家;

将所述多个密集层被分布到被配置为支持密集层的一个或多个可用加速器;以及

将所述若干专家被分布到被配置为支持专家的一个或多个可用加速器,使得被配置为支持专家的所述一个或多个加速器被配置为从所述被配置为支持密集层的一个或多个加速器接收多个输入。

10.根据权利要求9所述的方法,其中被包括在所述多个稀疏层中的每个稀疏层还包括一个或多个专家,使得所述机器学习模型被配置为混合专家机器学习模型。

11.根据权利要求9所述的方法,还包括:

访问被配置为支持一个或多个专家的附加加速器;以及

将所述附加加速器整合到所述计算系统中,使得所述计算系统是扩展的计算系统。

12.根据权利要求11所述的方法,还包括:

访问与所述机器学习模型相关联的一个或多个附加专家;以及

将所述一个或多个附加专家分布到所述附加加速器上。

13.根据权利要求9所述的方法,还包括:

选择跨一个或多个加速器分布的有限数目的专家,所述有限数目的专家被选择来处理输入数据;

访问输入数据集合;以及

使用至少所述有限数目的专家来处理所述输入数据。

14.根据权利要求9所述的方法,其中所述多个稀疏层与所述多个密集层交错。

15.根据权利要求9所述的方法,其中所述多个加速器是硬件加速器。

用于分布专用混合专家机器学习模型的层的系统和方法

背景技术

[0001] 使用混合专家(MOE)技术的机器学习模型通常由N个数目的层来组成,这些层被广泛地分类为MOE层和非MOE层。各种分布策略被用于将大型MOE机器学习模型分布到计算系统硬件中。

[0002] 当模型根据常规MOE分布策略分布时,单个加速器或图形处理单元(GPU)将被分派模型的层中的一些或所有层,包括MOE层以及非MOE层。然而,存在许多与这种分配相关联的问题。例如,某些组件将保持空闲,而其它组件仍在处理输入数据。此外,由于现有计算系统的当前硬件设备的限制,这种模型是不可缩放的。附加地,训练具有许多分布式层和专家的MOE模型在计算方面可能是繁重而且耗时。

[0003] 鉴于上述内容,存在针对用于可以分布在不同类型的硬件配置上的MOE机器学习模型的改进的系统和方法的持续需要。本文所要求保护的主题内容不限于解决任何缺点或仅在诸如上文所描述的环境中操作的实施例。相反,该背景仅被提供以示出可以在其中实践本文所描述的一些实施例的一个示例性技术领域。

发明内容

[0004] 所公开的实施例包括用于将MOE模型分布到各种系统配置中的系统、方法和设备。

[0005] 在一些实施例中,MOE模型分布在包括多个加速器的系统中,该多个加速器包括第一加速器集合和第二加速器集合,其中根据这些实施例,第二加速器集合中的加速器具有比第一加速器集合中的加速器更大的存储器能力,并且第一加速器集合中的加速器具有比第二加速器集合中的加速器更大的处理能力。

[0006] 在一些实施例中,MOE模型被配置有分布在多个系统加速器(例如上面所提及的第一加速器集合和第二加速器集合)上的多个密集层和多个稀疏层,使得多个密集层分布在从第一加速器集合所选择的一个或多个加速器上,并且多个稀疏层分布在第二加速器集合中的一个或多个加速器上。在一些情况下,机器学习模型还以这种方式分布:存储机器学习模型的一个或多个稀疏层的至少一个加速器被配置为从存储机器学习模型的密集层的一个或多个加速器接收多个输入。

[0007] 如本文所描述的,所公开的实施例也涉及用于将机器学习模型的不同密集层和稀疏层分布到定制计算系统上的系统和方法。例如,系统被配置为访问包括多个密集层和多个稀疏层的机器学习模型,并且标识被包括在机器学习模型中的若干密集层和若干专家。系统也被配置为访问包括多个加速器的计算系统,其中至少一些加速器具有比其它加速器更大的存储器能力并且至少一些加速器具有比其它加速器更大的处理能力,其标识哪些加速器可用,并且确定可用加速器中的哪些加速器被配置为支持密集层以及哪些加速器被配置为支持专家。

[0008] 所公开的系统也被配置为将多个密集层分布到被配置为支持密集层的一个或多个可用加速器,并且将若干专家被分布到被配置为支持专家的一个或多个可用加速器,使得该被配置为支持专家的一个或多个加速器被配置为从被配置为支持密集层的一个或多

个加速器接收多个输入。

[0009] 本发明内容被提供来以简化的形式介绍一系列概念,这将在下面的具体实施方式中被进一步描述。本发明内容不旨在于标识所要求保护的主题内容的关键特征或必要特征,也不旨在于被用作对确定所要求保护的主题内容的范围的帮助。

[0010] 附加的特征和优点将在下面的描述中被阐述,并且部分地将从描述中显而易见,或者可以通过实践本文的教导来被获知。本发明的特征和优点可以通过所附权利要求中被特定指出的手段和组合的方式来实现和获得。本发明的特征将从以下描述和所附权利要求中变得更完全地显而易见,或者可以通过实践如下文中所阐述的本发明来被获知。

附图说明

[0011] 为了描述可以在其中获得上述和其他优点和特征的方式,将通过参考在附图中被示出的具体实施例来呈现对以上被简要描述的主题内容的更具体的描述。要理解的是,这些附图仅描绘了典型的实施例并且因此不被认为是对范围的限制,实施例将通过附图的使用来利用附加的具体性和细节而被描述和解释,在其中:

[0012] 图1A至图1D示出了现有MOE系统的各种示例实施例。

[0013] 图2示出了根据所公开的实施例的分布在计算系统上的MOE机器学习模型的示例图。

[0014] 图3示出了在其中计算系统并入和/或被用于执行所公开的实施例的所公开的方面的计算环境。

[0015] 图4A至图4D示出了分布式的特殊MOE机器学习模型的各种示例实施例。

[0016] 图5示出了具有用于确定包括MOE的机器学习模型的分布的多个动作的流程图的一个实施例。

[0017] 图6示出了具有用于将机器学习模型的一个或多个专家迁移到计算系统的一个或多个不同加速器的多个动作的流程图的一个实施例。

具体实施方式

[0018] 所公开的实施例涉及用于将包括MOE的机器学习模型分布到计算系统上的系统和方法。在该方面,将理解的是,所公开的实施例中的一些实施例具体涉及用于基于在定制的硬件设备上将稀疏层与密集层分离来确定机器学习的分布的改进的系统和方法。所公开的实施例提供优于现有系统的许多技术优点。

[0019] 例如,传统的基于变换器的机器学习模型使用按顺序处理输入数据的变换器层的堆叠来被构造。例如,来自先前变压器层的输出被用作到下一变压器层的输入。来自典型变换器层的所有神经元参与处理每个输入。采用该层内的神经元中的所有或大多数神经元的变换器层被标识为密集层,而采用该层内的一个或有限数目的神经元的变换器层被标识为稀疏层。密集层需要大量浮点运算(FLOPS)和大量GPU存储器来处理输入。以这种方式被配置有密集层的机器学习模型难以缩放。

[0020] 一些数据科学家已经开始使用传统变压器层的变体,其已经被称为混合专家(MOE)层,该MOE层作为缩放机器学习模型的方式。在一些情况下,作为稀疏层类型的MOE层使用专家的集合来被构建。例如,如果模型正在被训练以执行特定任务,则该特定任务(例

如,预测性建模任务)可以被分解成两个或多个子任务。然后每个专家在子任务中的一个子任务上被训练。虽然在一些情况下,专家被配置作为模型,例如具有其自己的节点或神经元集合的神经网络,但是当特定机器学习模型层内的专家集合形成神经网络时,专家也可以被称为节点或神经元。因此,在MOE层(即,稀疏层)的情况下,每个输入可以由来自MOE层的专家(即,神经元)的有限子集来处理。

[0021] 这与其中所有或大多数神经元参与数据处理的密集层相反,而不是如针对稀疏层的情况那样选择几个。在一些现有系统中,包括密集层和稀疏层的整个机器学习模型分布在单个硬件上,其在本文中被称为加速器(例如,GPU 1),如图1A中所示出的。例如,如所示出的,GPU 1包括多个层(例如,层N-1、层N和层N+1)。层N还包括残差连接和层归一化层、前馈网络层(例如,FFN1、FFN2、FFNe等)上的一个或多个专家、门控层、附加的残差连接和层归一化层以及多头注意力层。以这种方式,层N-1和层N+1是密集层,而层N内的稀疏层包括不同的专家。

[0022] 关于前述内容,将理解的是,存在许多不同类型的加速器,包括但不限于硬件加速器、图形加速器、加密加速器、web加速器、PHP加速器或其它类型的加速器,其中所有的加速器在本文中被统称为加速器,并且其中的每个加速器包括对应的机器学习模型的一个或多个密集层或稀疏层。

[0023] 基于MOE的变压器机器学习模型的不同层以各种配置可配置。在一些现有系统中,机器学习模型的不同层分布在多个加速器(例如,GPU 1和GPU N)上,其中每个加速器在其稀疏层中具有单个专家,如图1B中所示出的。例如,虽然GPU 1和GPU N中的每个GPU具有与图1A中的GPU 1类似的层配置,但是GPU 1和GPU N两者各自仅在其稀疏层(层N)内包括单个专家(例如,GPU 1中的FFN1和GPU N中的FFNn)。

[0024] 在一些配置中,密集层和稀疏层是交错的。例如,如果机器学习模型使用两个密集层(例如,密集层1,密集层2)和两个稀疏层(例如,稀疏层1,稀疏层2)来被构造,则机器学习模型可以根据图1C来被配置。如所示出的,输入数据首先由密集层1来处理,然后由稀疏层1来处理,然后由密集层2来处理,然后由稀疏层2来处理,以便生成最终输出。

[0025] 在图1D中,每个稀疏层由至少两个专家组成。例如,稀疏层1包括第一多个专家(例如,S1E1、S1E2)以及稀疏层2包括第二多个专家(例如,S2E1、S2E2)。为了适应这种大型机器学习模型(即,整个模型将不适合单个加速器),机器学习模型的层被分布到多于一个加速器上。例如,密集层1、稀疏层1和密集层2分布在GPU1上,而密集层3、稀疏层2和密集层4分布在GPU2上。然而,这种通常被称为模型并行性的分布方案是非常低效的,因为当GPU1正在处理输入时,GPU2保持空闲,而当GPU2正在处理输入时,GPU1是空闲的。

[0026] 一些工作已经集中于通过引入处理流水线来减轻该低效,使得当GPU2正在处理第一输入(在第一输入已经由GPU1处理之后)时,GPU1开始处理第二输入。然而,该配置仍具有缺点,其在于GPU利用率保持较低,因为不参与处理给定输入的一个或多个稀疏层中的任何专家仍占用大量GPU存储器。

[0027] 附加的改进已经被探索,其被称为专家并行性,这提供了其中专家跨GPU均匀分布的模型配置。在这种配置中,系统可以基于N个GPU来同时处理多达N个输入。在一个示例中,其中存在四个GPU和四个专家,每个GPU从每个稀疏层仅分配单个专家。在该配置中,系统可以同时处理多达四个输入。稀疏层将交换输入,使得每个输入被发送给GPU,其中已经被选

择来处理输入的专家被存储。然而,即使该改进仍具有有限的能力。例如,除了稀疏层之外,每个GPU处理密集层。在一些情况下,因为大量的存储器由稀疏层来占用,这确实需要比密集层更少的处理,所以这是低效的。这降低了系统的整体计算效率。附加地或备选地,每个GPU上的模型消耗相同数量的存储器和计算资源。因此,可缩放性由具有最小计算功率和存储器的GPU来约束,其导致系统中具有较大存储器存储和/或计算功率的GPU的低效。

[0028] 注意力将聚焦于图2,其示出了特殊专家机器学习模型的示例实施例,其中稀疏层被分布到稀疏硬件上以及密集层被分布到密集硬件上,并且其中稀疏层与密集层交错,使得稀疏硬件可以处理来自多个密集硬件设备的多个输出。所公开的实施例提供了优于现有系统的许多技术优点。

[0029] 例如,计算系统200被显示具有多个加速器(例如,加速器202、加速器204、加速器206、加速器208,以及未被示出的一个或多个其他加速器)。机器学习模型被分布到各种加速器上。例如,第一多个模型层(例如,层210、层211和层222)被显示被分布到加速器202上。每层还包括一个或多个层(即,子层)。例如,层211包括层212(例如,残差连接和层归一化)、层214(例如,稀疏层)(其还包括门控层216)、层218(例如,残差连接和层归一化)和层220(例如,多头注意力)。

[0030] 类似地,显示被分布到加速器204上的第二多个模型层(例如,层224、层215和层234)。每层还包括一个或多个层(即,子层)。例如,层215包括层226(例如,残差连接和层归一化)、层224(例如,稀疏层)(其还包括门控层228)、层230(例如,残差连接和层归一化)和层232(例如,多头注意力)。

[0031] 如上面所描述的,在传统系统中,稀疏层将包括一个或多个专家。然而,在所示出的实施例中,其是本文中所公开的一些实施例的代表,来自不同稀疏层的专家已经被分布到单独的加速器上。例如,来自与层210相关联的稀疏层的专家236和专家238以及来自与层211相关联的层214的专家240和专家242被分布到加速器206上。附加地,专家244、专家246、专家248和专家250被分布到加速器208上。如所显示的,专家被配置为从如跨多个加速器分布的机器学习模型的各个层接收不同的输入。这防止某些组件空闲而其它组件正在活跃地处理数据。

[0032] 以这种方式,MOE层和非MOE层被分离成单独的组,并且每组被分派其自己的加速器集合。通过根据该配置分布机器学习模型,计算系统能够实现以下技术优点。首先,系统能够以选择性的方式应用适合于密集计算和稀疏计算的特定性能优化。附加地,MOE层合并繁重的通信开销。因此,与图2中所示出的一种配置类似的配置使得系统能够利用集群的子集(例如,在单个节点上或单个机架中的节点上)内可用的更高通信带宽。

[0033] 此外,根据所公开的实施例,单个MOE层集合可以被调度以处理多个交织的输入来有效地增加模型吞吐量(或者用于训练、推断和/或实现)。

[0034] 所公开的实施例也涉及能够使用异构集群来降低计算成本并且提高计算系统/机器学习模型的性能的系统和方法。具体地,这种配置有助于显著减少机器学习训练所必须的时间,因此允许用户更快地将机器学习模型部署到实现任务中。

[0035] 现在将注意力聚焦于图3,其示出了可以包括和/或被用于实现所公开的发明的各个方面的计算系统310的组件。如所显示的,计算系统包括与机器学习引擎和模型的输入和输出相关联的多个机器学习(ML)引擎、模型和数据类型。例如,图3示出了作为计算环境300

的部分的计算系统310,该计算环境300也包括(经由网络330)与计算系统310通信的(多个)远程/第三方系统320。计算系统与包括一个或多个处理器322和一个或多个计算机可执行指令324的(多个)远程/第三方系统320通信。预期在一些情况下,(多个)远程/第三方系统320还包括容纳可以被用作训练数据的数据的数据库,例如外部扬声器数据。附加地或备选地,(多个)远程/第三方系统320包括计算系统310外部的机器学习系统。在一些实施例中,(多个)远程/第三方系统320是软件程序或应用。

[0036] 例如,计算系统310包括一个或多个处理器312(例如一个或多个硬件处理器)和存储计算机可执行指令318的存储装置(即,(多个)硬件存储设备340)。(多个)硬件存储设备340中的一个或多个硬件存储设备能够容纳任何数目的数据类型和任何数目的计算机可执行指令318,当计算机可执行指令318由一个或多个处理器312执行时,计算系统310由该计算机可执行指令318来配置为实现所公开实施例的一个或多个方面。

[0037] 计算系统310还包括多个加速器(例如,密集加速器313和稀疏加速器314)。在一些配置中,密集加速器313被配置为使用密集层来处理输入数据,其中密集加速器313是针对处理功率被优化的定制硬件。在这种配置中,稀疏加速器314被配置为使用稀疏层来处理输入数据,其中稀疏加速器314是针对存储器存储装置被优化的定制硬件。稀疏加速器314在处理稀疏数据(例如,稀疏张量、稀疏层)方面比密集加速器更有效。加速器中的每个加速器可以包括能够存储和/或执行对应的密集层和稀疏层(分别为344和346)的专用处理器或其它硬件。

[0038] 在一些情况下,稀疏加速器具有比密集加速器大至少10%的存储器或存储容量,或者甚至更多(例如,比密集加速器大20%、30%、40%、50%、60%、70%、80%、90%或甚至超过100%的存储器存储容量)。附加地或备选地,稀疏加速器在处理稀疏数据方面比密集加速器效率高至少10%,或者甚至更多(例如,在处理稀疏数据方面比密集加速器效率高20%、30%、40%、50%、60%、70%、80%、90%或甚至超过100%)。

[0039] 在一些情况下,密集加速器具有比稀疏加速器大至少10%的处理能力,或者甚至更多(例如,比稀疏加速器的处理能力大20%、30%、40%、50%、60%、70%、80%、90%或甚至100%)。例如,密集加速器在处理密集数据(例如,密集层、密集张量)方面比稀疏加速器更有效。

[0040] 在一些情况下,稀疏加速器至少基于它们在处理稀疏数据中的更高效率而区别于密集加速器。附加地或备选地,稀疏加速器基于其与密集加速器相比增加的存储器容量和/或减少的原始FLOP的数目而区别于密集加速器。

[0041] 计算系统310也被显示包括(多个)用户接口315和(多个)输入/输出(I/O)设备316。

[0042] 如图3中所显示的,(多个)硬件存储设备340被显示为单个存储单元。然而,将理解的是,在一些实施例中,(多个)硬件存储设备340是分布式存储装置,其分布到若干单独的并且有时是(多个)远程和/或第三方系统320(例如,硬件存储设备324)。在一些实施例中,计算系统310也可以包括分布式系统,其中计算系统310的组件中的一个或多个组件由彼此远离并且各自执行不同任务的不同分立系统来维护/运行。在一些情况下,例如在分布式云环境中,多个分布式系统执行类似和/或共享的任务以用于实现所公开的功能。

[0043] 以该方式,机器学习模型的不同层可分布到计算系统310上和/或跨包括计算系统

310和一个或多个第三方系统320的分布式计算系统300上。

[0044] (多个)硬件存储设备340被配置为存储包括诸如机器学习模型342的各种模型的不同数据(例如,输入数据348),该机器学习模型342包括被用于处理输入数据348的密集层344和稀疏层346。

[0045] 存储装置(例如,(多个)硬件存储设备340)包括用于实例化或执行计算系统310中显示的模型和/或引擎中的一个或多个模型和/或引擎的计算机可执行指令318。模型被配置为机器学习模型或机器所学习的模型,例如深度学习模型和/或算法。在一些情况下,一个或多个模型被配置为引擎或处理系统(例如,被集成在计算系统310内的计算系统),其中每个引擎(即,模型)包括一个或多个处理器(例如,(多个)硬件处理器312)和对应于计算系统310的计算机可执行指令318。

[0046] 用于存储(多个)机器学习(ML)引擎350的附加存储单元目前在图3中被显示为存储多个机器学习模型和/或引擎。例如,计算系统310包括以下中的一个或多个:数据检索引擎351、分布引擎352和实现引擎353,它们被单独地和/或共同地配置为实现本文所描述的不同功能。

[0047] 例如,数据检索引擎351被配置为定位和访问包括一个或多个数据类型的数据源、数据库和/或存储设备,数据检索引擎351可以从这些数据类型提取被用作输入数据的集合或子集。数据检索引擎351从数据库和/或硬件存储设备接收数据,其中该数据检索引擎351被配置为重新格式化或以其他方式扩充所接收的数据,以被用作训练数据或输入数据。附加地或备选地,数据检索引擎351与包括远程/第三方数据集和/或数据源的一个或多个远程/第三方系统(例如,(多个)远程/第三方系统320)通信。在一些情况下,这些数据源包括记录语音、文本、图像和/或视频的视听服务。

[0048] 数据检索引擎351访问包括声音数据、文本数据和/或其它类型的视听数据(包括视频数据、图像数据、全息数据、3-D图像数据等)的电子内容。数据检索引擎351是智能引擎,其能够学习最优数据集提取过程以及时地提供足够数量的数据以及检索最适于期望应用的数据,机器学习模型/引擎将针对该期望被使用。例如,数据检索引擎351可以学习哪些数据库和/或数据集将生成训练数据,该训练数据将训练模型(例如,针对特定查询或特定任务)以提高该模型在期望层分布配置中的准确度、效率和功效。

[0049] 在一些实施例中,计算系统310包括被配置为确定跨不同加速器的机器学习模型342的不同层的分布的分布引擎352。分布引擎352也被配置为促进在模型的实例化之前的分布或在模型的实例化之后的机器学习模型的再分布。在一些情况下,再分布基于标识计算系统的一个或多个属性(例如,模型吞吐量、计算效率)的潜在改进和/或基于机器学习模型的按比例放大或缩小。在一些实施例中,再分布通过将一个或多个专家从一个加速器迁移到不同的加速器来实现。

[0050] 分布引擎352被配置为动态地标识组成计算系统310的加速器的总数目,以及标识哪些加速器是针对密集层与稀疏层相比的专业化或优化的硬件设备。附加地,分布引擎352被配置为标识哪些加速器是满的和/或哪些加速器具有可用的或预期的处理能力和/或用于存储和执行一个或多个附加层的存储器空间。

[0051] 分布引擎352也被配置为标识有多少层组成机器学习模型,以及标识哪些层是密集层并且哪些层是稀疏层。分布引擎352还被配置为标识在不同稀疏层中的每一层中有多

少专家,以及标识专家的属性(例如,什么专业化任务与专家相关联)。

[0052] 因此,基于密集层的数目和稀疏层的数目(或跨一个或多个稀疏层的专家的数目),分布引擎352被配置为将不同层动态地以及自动地分布到计算系统310的一个或多个加速器上。在一些情况下,分布引擎352被配置为将密集层344分布到密集加速器313(例如,被确定为具有用于存储和/或执行密集层的能力的加速器)上,并且将稀疏层346分布到稀疏加速器314(例如,具有比密集加速器更有限的容量和能力的加速器)上。

[0053] 在一些实施例中,计算系统310包括与被包括在计算系统310中的模型和/或(多个)ML引擎350中的任何一个模型和/或ML引擎(或所有模型/引擎)通信的实现引擎353,使得实现引擎353被配置为实现、启动或运行多个ML引擎350的一个或多个功能。在一个示例中,实现引擎353被配置为操作数据检索引擎351,使得数据检索引擎351在适当的时间处检索数据以能够处理输入数据348。实现引擎353促进(多个)ML引擎350中的一个或多个ML引擎之间的过程通信和通信定时。

[0054] 在其他示例中,实现引擎353被配置为将机器学习模型342的一个或多个功能(即,处理输入数据348)实现为被分布到计算系统310上或跨计算系统310和(多个)第三方系统320分布。实现引擎353也被配置为实现分布引擎352以便标识机器学习模型的不同层的分布或再分布。

[0055] 此外,实现引擎353被配置为选择在处理输入数据348方面将使用分布在各种加速器上的哪些专家。通过根据这些所公开的实施例实现系统和方法,计算系统310能够实现诸如可定制和可缩放的技术益处。具体地,不同的专家可以在处理输入数据348方面的不同时间处被使用。因此,系统可以被配置为基于输入数据的类型、输入数据的格式化、输入数据的上下文和/或经处理的输入数据的下游应用来选择有限数目的专家,以在处理输入数据方面使用。因此,当一个或多个专家被标识为将被用于处理输入数据的专家时,系统可以将包括那些所标识的专家的一个或多个稀疏层分布或再分布到不同的加速器上,以便增加模型吞吐量并且增加系统的计算效率。

[0056] 现在将注意力聚焦于图4A至图4D,其示出了被分布到计算系统(例如,计算系统310)上的机器学习模型(例如,机器学习模型342)的各种实施例。首先注意力聚焦于图4A,其示出了新颖配置的示例实施例。如所示出的,计算系统400包括多个加速器(例如,加速器402、加速器404、加速器406、加速器408和加速器410)。在一些情况下,如图4A中所显示的,加速器被配置为GPU(例如,GPU1、GPU2、GPU3、GPU4和GPU5)。加速器402和加速器410被配置为使用密集层来处理输入。例如,多个密集层(例如D1、D2、D3和D4)被分布到加速器402和加速器410上。加速器404、加速器406和加速器408被配置为使用稀疏层来处理来自先前层的输入。

[0057] 例如,包括第一多个专家(例如,S1E1、S1E2、S1E3和S1E4)的第一稀疏层被分布到加速器404上,包括第二多个专家(例如,S2E1、S2E2、S2E3和S2E4)的第二稀疏层被分布到加速器406上,以及包括第三多个专家(例如,S3E1、S3E2、S3E3和S3E4)的第三稀疏层被分布到加速器408上。在上述专家中,符号如下:“S”指代稀疏层,其具有以下指示稀疏层的数目的数字,以及“E”指代专家,其具有以下指示专家的数目的数字。虽然每个加速器被显示为具有至少四层,但是应当理解的是,任何数目的层,包括没有层,可分布到加速器中的任何一个加速器上或跨多个加速器可分布。

[0058] 如图4A中所显示的,加速器404、加速器406和加速器408中分别被配置为通过交错同时接收来自加速器402和加速器410的多个输入。以这种方式,多个加速器能够同时处理输入,因此增加了计算系统的计算效率和处理功率。

[0059] 通过以这种方式分布不同的层,进一步的技术优点被实现。例如,系统能够配备已经被优化以用于处理密集层的第一定制硬件集合和已经被优化以用于处理稀疏层的第二定制硬件集合。这在加速器被配置为处理具有密集层和稀疏层两者的输入的先前配置中是不可能的。

[0060] 在一些情况下,针对密集层被优化的硬件设备由高处理功率属性来特性化。附加地,这种硬件设备也由较小的存储器存储装置来特性化。在一些情况下,针对稀疏层被优化的硬件设备由高存储器存储属性来特性化。附加地,这种硬件设备也由比针对密集层被优化的硬件设备更低的处理功率属性来特性化。

[0061] 现在将注意力聚焦于图4B,其示出了被分布到包括定制硬件的计算系统401上的机器学习模型。例如,图4B包括与图4A类似的组件,其中加速器402和加速器410被配置用于密集层(即,密集硬件),并且其中加速器404、加速器406和加速器408被配置用于稀疏层。通过定制计算系统400的不同硬件组件,计算系统的计算效率和处理功率增加。

[0062] 在一些情况下,效率由在处理中被使用的层与其在特定加速器上占据的存储器相比多长时间来特性化。因此,有效的系统是这样一种系统,其中进行很少处理的层比占用更大数量存储的参与大量处理的层占用更小数量的存储器。备选地,进行很少处理但是需要大量存储器的层(例如,稀疏层)可以被分布到专用硬件(例如,密集硬件)上,该专用硬件针对超过处理能力的存储器存储装置被优化,而层(例如,密集层)分布在专用硬件(例如,稀疏硬件)上,该专用硬件针对超过存储容量的处理能力被优化。

[0063] 所公开的实施例有利地实现了附加的技术益处,包括高度可缩放。现在将注意力聚焦于图4C,其示出了计算系统403,该计算系统403是还包括被配置为稀疏硬件设备的附加加速器(例如,加速器412)的计算系统401是代表。包括第四多个专家的第四稀疏层被分布到加速器412上。如所示出的,第四多个专家包括两个专家(例如,S4E1和S3E4)。应当理解的是,不同加速器中的任何加速器上的任何多个专家可以包括一个或多个专家。

[0064] 因为机器学习模型的层以选择性方式分布(即,密集硬件上的密集层以及稀疏硬件上的稀疏层),所以机器学习模型是高度可缩放的,如图4C中所示出的。例如,为了增加系统的吞吐量,附加加速器(例如,GPU6)可以以较小增量被添加。以这种方式,更小或更大的加速器可以被添加到系统中以适应包括一个或多个附加专家的一个或多个附加稀疏层。这也在机器学习模型的不同层如何被分布到计算系统上的方面提供附加的灵活性和可定制性。

[0065] 附加地或备选地,来自另一加速器的一个或多个专家可以被迁移到附加加速器上。例如,如图4D中所示出的,代表具有类似硬件设备而非机器学习模型层的不同分布的计算系统403的计算系统405被显示为具有最初被分布到加速器408(见图4C)上而现在被分布到加速器412上的一些专家(例如,S3E3和S3E4)。在具有可用处理和存储器存储能力的附加加速器被标识的情况下,系统能够将一个或多个专家再分布(即迁移)到附加加速器上。例如,如果系统选择专家S3E1和专家S3E3作为将被用于处理新输入数据的专家,则系统确定将增加模型吞吐量并且促进计算处理方面中的改进的层的再分布。

[0066] 相反,如果S3E1和S3E3已经保持在相同的稀疏硬件(例如,加速器408)上,则输入数据将必须被串行处理,首先利用S3E1并且然后利用S3E3。然而,现在通过再分布,可以使用加速器408和加速器412两者并行地处理输入数据。以这种方式,减少了等待时间(一层不等待另一层以完成处理)并且增加了模型吞吐量。

[0067] 将理解的是,基于与在不同数据集上执行模型相关联的能力和约束的动态评估,专家的再分布和迁移可以发生在停机时间期间和/或在由模型正在执行的处理期间。

[0068] 因为每个专家在特定任务或数据应用上被训练,所以机器学习模型能够使用各种不同的训练任务或应用来处理输入数据。这种特殊的MOE配置允许机器学习模型容易且有效地合并新的专家或已更新的专家,而不必重新训练整个模型和/或不必重新配置被分布到计算系统上的其它层。

[0069] 附加地,在一些实施例中,新专家被添加到可用的加速器上或被添加到容易被集成到系统中的新加速器上。这是高效的配置,因为硬件不处于空闲(即,多个硬件设备同时处理多个输入)。这是对传统混合专家系统的显著改进,该传统混合专家系统由于一些硬件设备空闲而其它硬件设备正在处理输入所以不容易适于新的任务、可扩展,并且具有有限的吞吐量。

[0070] 现在将注意力聚焦于图5,其示出了包括与示例性方法相关联的各种动作(动作510、动作520、动作530、动作540、动作550、动作560和动作570)的流程图500,该示例性方法可以由计算系统310来实现以用于将混合专家机器学习模型分布到定制的计算系统上。

[0071] 所示出的第一动作包括用于访问包括多个密集层和多个稀疏层的机器学习模型的动作(动作510)。因为模型具有密集层和稀疏层,所以机器学习模型能够在稀疏层内实现多个不同的专家,其针对不同的训练任务、域和应用提供机器学习的附加灵活性和可定制性。附加地,多个稀疏层与多个密集层交错。

[0072] 然后,系统标识被包括在机器学习模型中的若干密集层和若干专家(动作520)。附加地,系统访问包括多个加速器的计算系统,其中至少一些加速器具有比其他加速器更大的存储器能力,并且至少一些加速器具有比其他加速器更大的处理能力(动作530)。通过标识计算系统和机器学习模型中的每个中的特定数目的元素,系统能够生成准确且高效的分布策略。此外,加速器是硬件加速器。例如,一个或多个硬件加速器可配置为图形单元处理器。

[0073] 随后,系统标识哪些加速器是可用的(动作540)并且确定可用加速器中的哪些加速器被配置为支持密集层以及哪些加速器被配置为支持专家(动作550)。

[0074] 在访问机器学习模型和计算系统之后,该系统被配置为将多个密集层分布到被配置为支持密集层的一个或多个可用加速器(动作560),以及将若干专家分布到被配置为支持专家的一个或多个可用加速器,使得被配置为支持专家的一个或多个加速器被配置为从被配置为支持密集层的一个或多个加速器接收多个输入(动作570)。

[0075] 通过以这种方式生成分布策略,效率针对系统和模型组合被提高。这是因为进行很少处理的层比参与占用大量存储器的大量处理的层占用更少量的存储器。备选地,进行很少处理(例如,稀疏层)但是需要大量存储器的层可以被分布到专用硬件(例如,密集硬件)上,其针对超过处理能力的存储器存储装置被优化,而层(例如,密集层)分布在专用硬件(例如,稀疏硬件)上,其针对超过存储器容量的处理能力被优化。

[0076] 应当理解的是,被包括在多个稀疏层中的每个稀疏层还包括一个或多个专家,使得机器学习模型被配置为MOE机器学习模型。

[0077] 在一些实施例中,方法还包括访问被配置为支持一个或多个专家的附加加速器的动作以及将附加加速器集成到计算系统中使得计算系统是经扩展的计算系统的动作。

[0078] 在这种实施例中,系统能够访问与机器学习模型相关联的一个或多个附加专家,并且将一个或多个附加专家分布到附加加速器上。

[0079] 附加地或备选地,系统被配置为选择跨一个或多个加速器分布的有限数目的专家,使得正在被选择来处理输入数据的有限数目的专家访问输入数据集合;以及使用至少有限数目的专家来处理输入数据。

[0080] 现在将注意力聚焦于图6,其示出了包括与示例性方法相关联的各种动作(动作610、动作620、动作630、动作640、动作650、动作660和动作670)的流程图600,该示例性方法可以由计算系统310来实现以用于确定MOE机器学习模型的一个或多个层的再分布。

[0081] 所示出的第一动作包括访问包括多个加速器的计算系统的动作,其中至少一些加速器具有比其他加速器更大的存储器能力并且至少一些加速器具有比其他加速器更大的处理能力(动作610)。

[0082] 该方法也包括访问包括多个密集层和多个稀疏层的机器学习模型的动作,每个稀疏层还包括一个或多个专家,该机器学习模型分布在被包括在多个加速器中的一个或多个加速器上,使得机器学习模型的多个密集层分布在具有更大处理能力的一个或多个加速器上,并且机器学习模型的多个稀疏层分布在具有更大存储器能力的一个或多个加速器上(动作620)。具有更大处理能力的一个或多个加速器被配置为密集加速器,其被优化以处理密集层。具有更大存储器能力的一个或多个加速器被配置为稀疏加速器,其被优化以处理稀疏层。因此,密集加速器具有比稀疏加速器更大的处理能力,并且稀疏加速器具有比密集加速器更大的存储器容量。这是因为稀疏层需要比密集层更多的存储器,但是在计算上需要更小的密集性,而密集层需要比稀疏层更少的存储器,但是在计算上需要更大的密集性。

[0083] 通过以这种方式分布不同的层,进一步的技术优点被实现。例如,系统能够配备有已经被优化用于处理密集层的第一定制硬件集合和已经被优化用于处理稀疏层的第二定制硬件集合。这在加速器被配置为处理具有密集层和稀疏层两者的输入的先前配置中是不可能的。

[0084] 在访问计算系统和机器学习模型之后,系统被配置为执行被配置为使计算系统将机器学习模型应用于输入数据的一个或多个计算机可执行指令(动作630)。

[0085] 在执行一个或多个计算机可执行指令之后,然后系统标识被包括在多个加速器中的若干可用加速器(动作640),并且根据机器学习模型在一个或多个加速器上的第一分布来确定计算系统的处理效率(动作650)。

[0086] 至少基于处理效率,系统标识机器学习模型在多个加速器上的新分布,该新分布被配置为增加计算系统的处理效率(动作660)。最后,系统将一个或多个专家中的至少一个专家迁移到具有更大存储器能力的若干可用加速器中的至少一个加速器上,以便促进计算系统的处理效率方面的提升(动作670)。通过以这种方式实现系统和方法,机器学习模型能够基于增加模型吞吐量和促进硬件分布效率方面的改进来被优化。

[0087] 此外,这种所公开的实施例提供了将机器学习模型缩放为更大或更小数目的层的

能力,包括将一个或多个不同的专家集成为MOE的部分或将该专家作为MOE的部分移除。这促进了减少用于训练和用于处理新输入数据的计算量和时间。例如,在一些情况下,方法还包括选择跨一个或多个加速器分布的有限数目的专家的动作,该有限数目的专家被选择来处理输入数据、访问输入数据集合;以及使用至少有限数目的专家来处理输入数据。

[0088] 附加地,一些方法也包括基于所选择的有限数目的专家来确定有限数目的专家跨多个加速器的附加分布的动作。以这种方式,系统能够根据附加分布将有限数目的专家中的一个或多个专家迁移到不同的加速器。

[0089] 示例计算系统

[0090] 本发明的实施例可以包括或利用包括计算机硬件的专用或通用计算机(例如,计算系统310),如下面更详细地讨论的。本发明的范围内的实施例也包括用于承载或存储计算机可执行指令和/或数据结构的物理和其它计算机可读介质。这种计算机可读介质可以是可以由通用或专用计算机系统来访问的任何可用介质。

[0091] 存储计算机可执行指令(例如,图1的计算机可执行指令318)的计算机可读介质(例如,图3的(多个)硬件存储设备340)是排除传输介质的物理硬件存储介质/设备。在一个或多个载波或信号中携带计算机可执行指令或计算机可读指令(例如,计算机可执行指令318)的计算机可读介质是传输介质。因此,通过示例而非限制的方式,本发明的实施例可以包括至少两种明显不同的类型的计算机可读介质:物理计算机可读存储介质/设备和传输计算机可读介质。

[0092] 物理计算机可读存储介质/设备是硬件,并且包括RAM、ROM、EEPROM、CD-ROM或其它光盘存储装置(例如CD、DVD等)、磁盘存储装置或其它磁存储设备,或可以被用于以计算机可执行指令或数据结构的形式存储所期望的程序代码装置,并且可以由通用或专用计算机来访问的任何其它硬件。

[0093] “网络”(例如,图3的网络330)被定义为实现电子数据在计算机系统和/或模块和/或其它电子设备之间的传输的一个或多个数据链路。当信息通过网络或另一通信连接(或者硬连线、无线或硬连线或无线的组合)被传送或提供给计算机时,计算机适当地将连接视为传输介质。传输介质可以包括网络和/或数据链路(其可以被用于承载),或者以计算机可执行指令或数据结构形式的所期望的程序代码装置,并且其可以由通用或专用计算机来访问。上文的组合也被包括在计算机可读介质的范围内。

[0094] 此外,在到达各种计算机系统组件时,以计算机可执行指令或数据结构的形式的程序代码装置可以从传输计算机可读介质自动地传送给物理计算机可读存储介质(反之亦然)。例如,通过网络或数据链路所接收的计算机可执行指令或数据结构可以被缓存在网络接口模块(例如,“NIC”)内的RAM中,并且然后最终被传送给计算机系统RAM和/或计算机系统处的较不易失性计算机可读物理存储介质。因此,计算机可读物理存储介质可以被包括在也(或甚至主要)利用传输介质的计算机系统组件中。

[0095] 例如,计算机可执行指令包括使通用计算机、专用计算机或专用处理设备来执行特定功能或功能组的指令和数据。例如,计算机可执行指令可以是二进制码、诸如汇编语言的中间格式指令,或者甚至是源代码。尽管已经以专用于结构特征和/或方法动作的语言来描述主题内容,但是应当理解的是,所附权利要求中所定义的主题内容不必限于上文所描述的特征或动作。相反,所描述的特征和动作作为实现权利要求的示例形式来被公开。

[0096] 本领域中的技术人员将理解的是,本发明可以在具有许多类型的计算机系统配置的网络计算环境中被实践,包括个人计算机、台式计算机、膝上型计算机、消息处理器、手持设备、多处理器系统、基于微处理器或可编程消费电子产品、网络PC、小型计算机、大型计算机、移动电话、PDA、寻呼机、路由器、交换机等。本发明也可以在分布式系统环境中被实践,其中通过网络被链接(或者由硬连线数据链路、无线数据链路或由硬连线和无线数据链路的组合)的本地和远程计算机系统都执行任务。在分布式系统环境中,程序模块可以位于本地和远程存储器存储设备中。

[0097] 备选地或附加地,本文所描述的功能可以至少部分地由一个或多个硬件逻辑组件来执行。可以被使用的说明性类型的硬件逻辑组件包括,例如但不限于,现场可编程门阵列(FPGA)、程序专用集成电路(ASIC)、程序专用标准产品(ASSP)、系统级片上系统(SOC)、复杂可编程逻辑设备(CPLD)等。

[0098] 在不脱离本发明的基本特性的情况下,本发明可以以其它特定形式被实施。所描述的实施例将在所有方面仅被认为是说明性的而非限制性的。因此,本发明的范围由所附权利要求而不是由前面的描述来指示。在权利要求的等同物的含义和范围内的所有变化都被包含在其范围内。

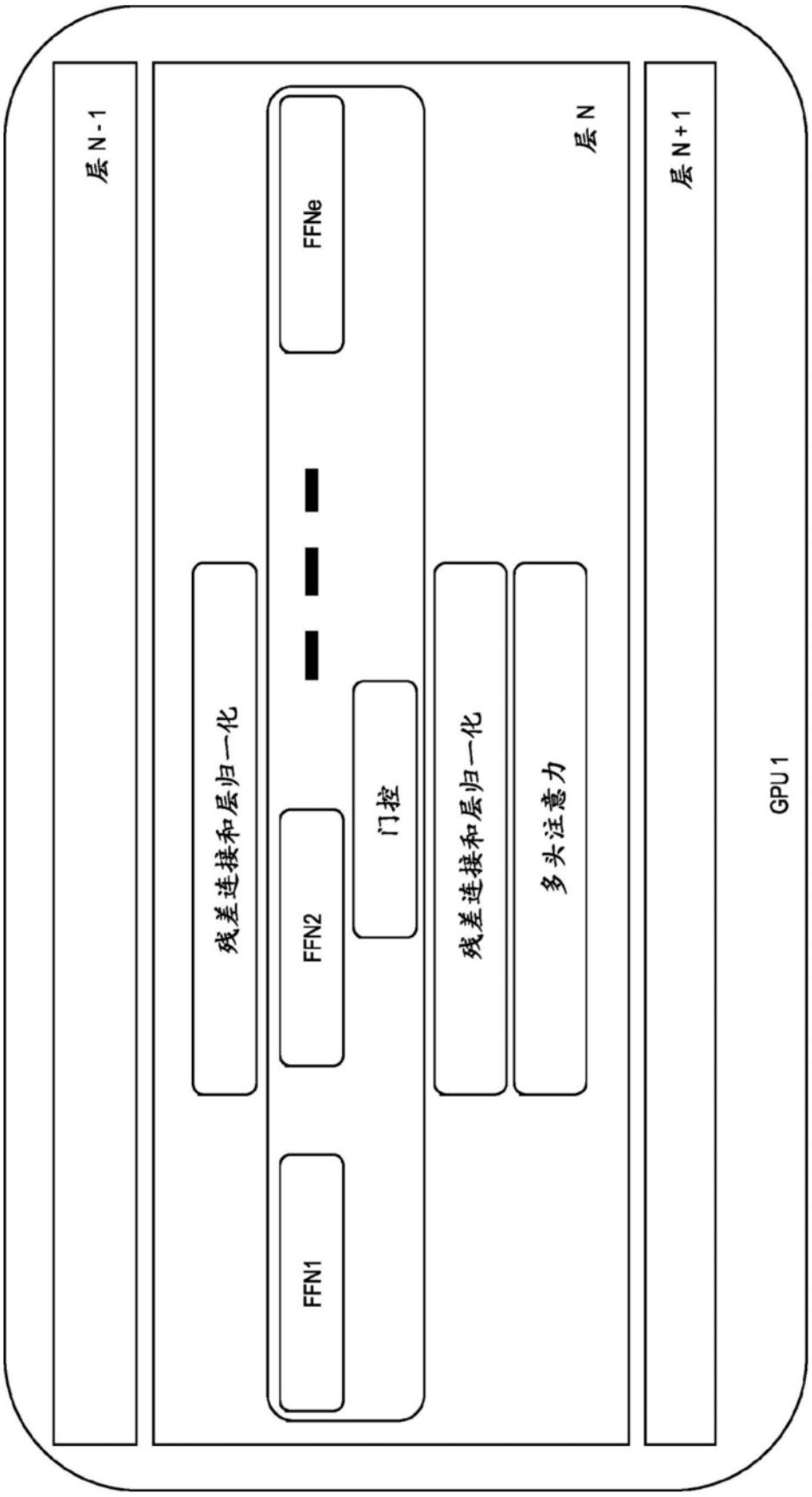
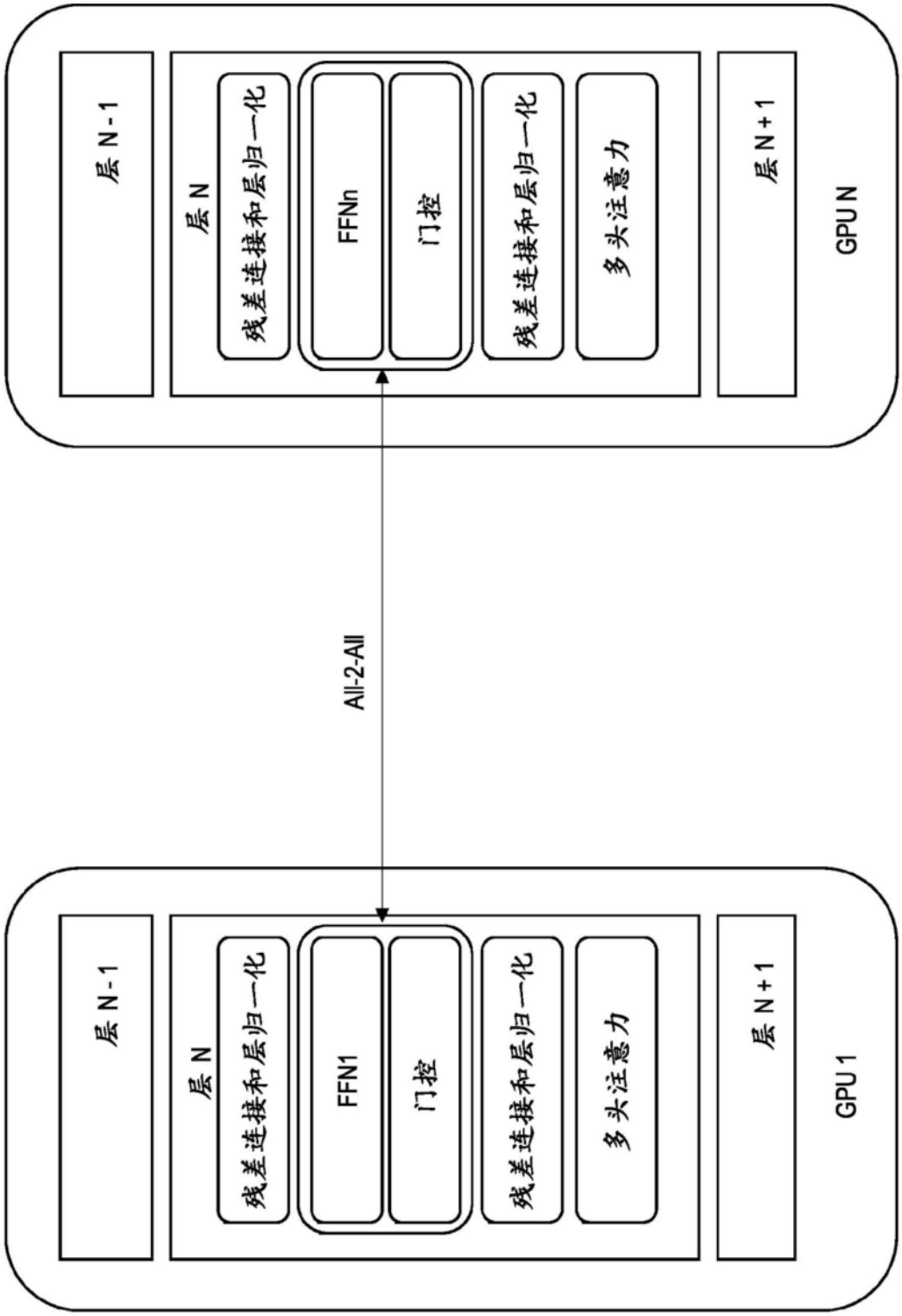

层 N-1
层 N
层 N+1
残差连接和层归一化
FFN1
FFN2
FFNe
门控
残差连接和层归一化
多头注意力
GPU 1
(现有技术)

图1A


层 N-1
层 N
残差连接和层归一化
FFNn
门控
残差连接和层归一化
多头注意力
层 N+1
GPU N
All-2-All
层 N-1
层 N
残差连接和层归一化
FFN1
门控
残差连接和层归一化
多头注意力
层 N+1
GPU 1

(现有技术)

图1B

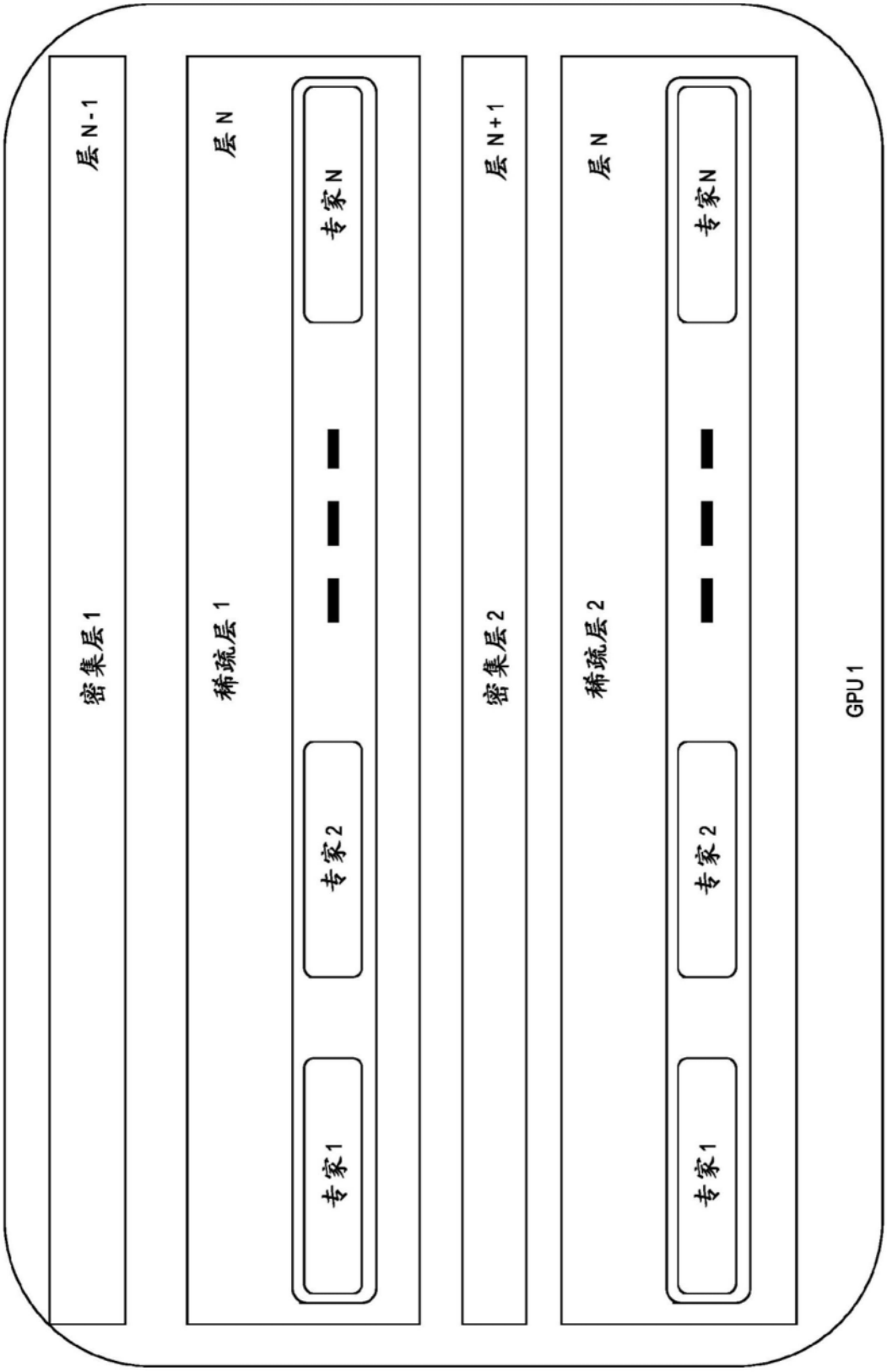
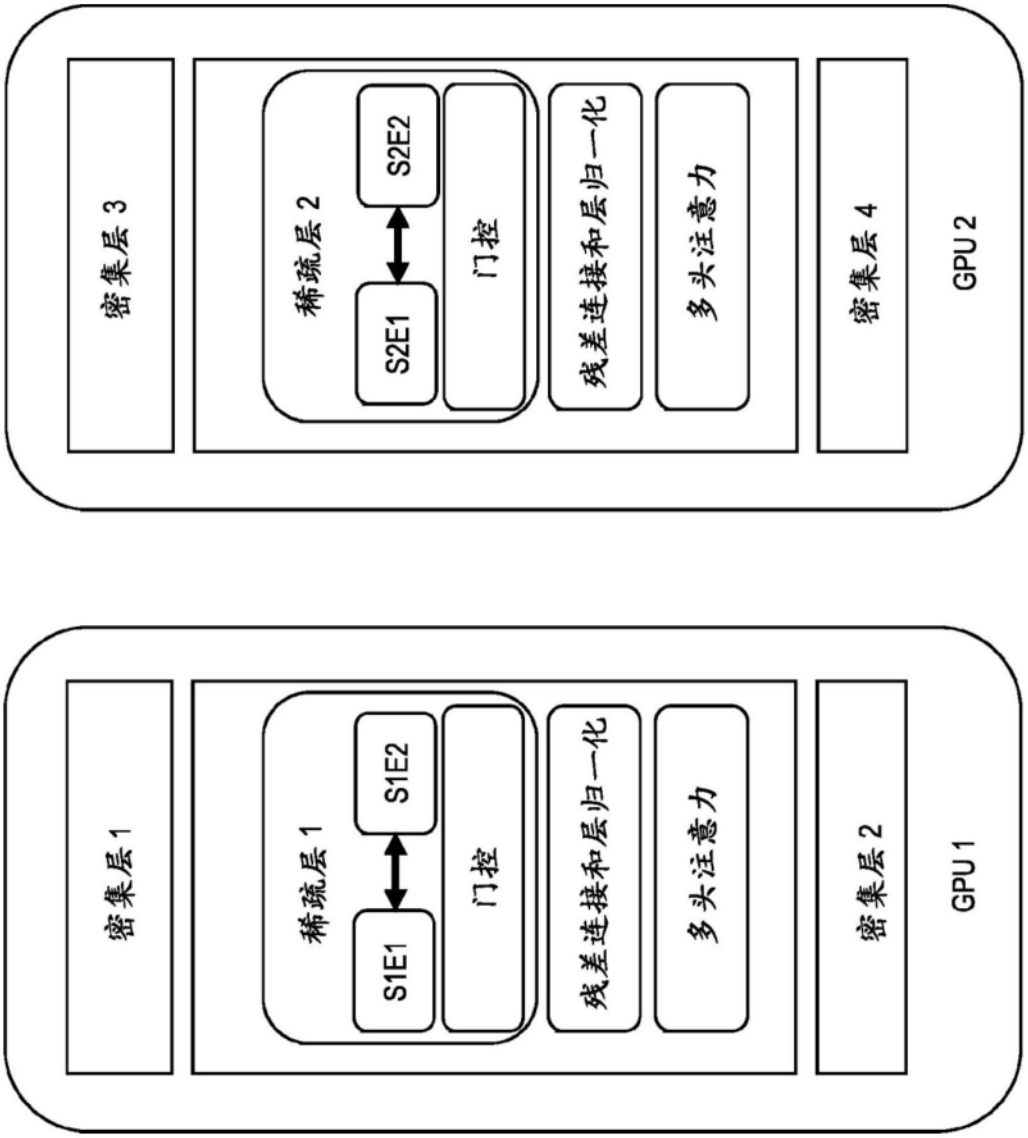

层 N-1
密集层 1
层 N
稀疏层 1
专家 1
专家 2
专家 N
层 N+1
密集层 2
层 N
稀疏层 2
专家 1
专家 2
专家 N
GPU 1
（现有技术）

图1C


密集层 1
稀疏层 1
S1E1
S1E2
门控
残差连接和层归一化
多头注意力
密集层 2
GPU 1
密集层 3
稀疏层 2
S2E1
S2E2
门控
残差连接和层归一化
多头注意力
密集层 4
GPU 2

（现有技术）

图1D

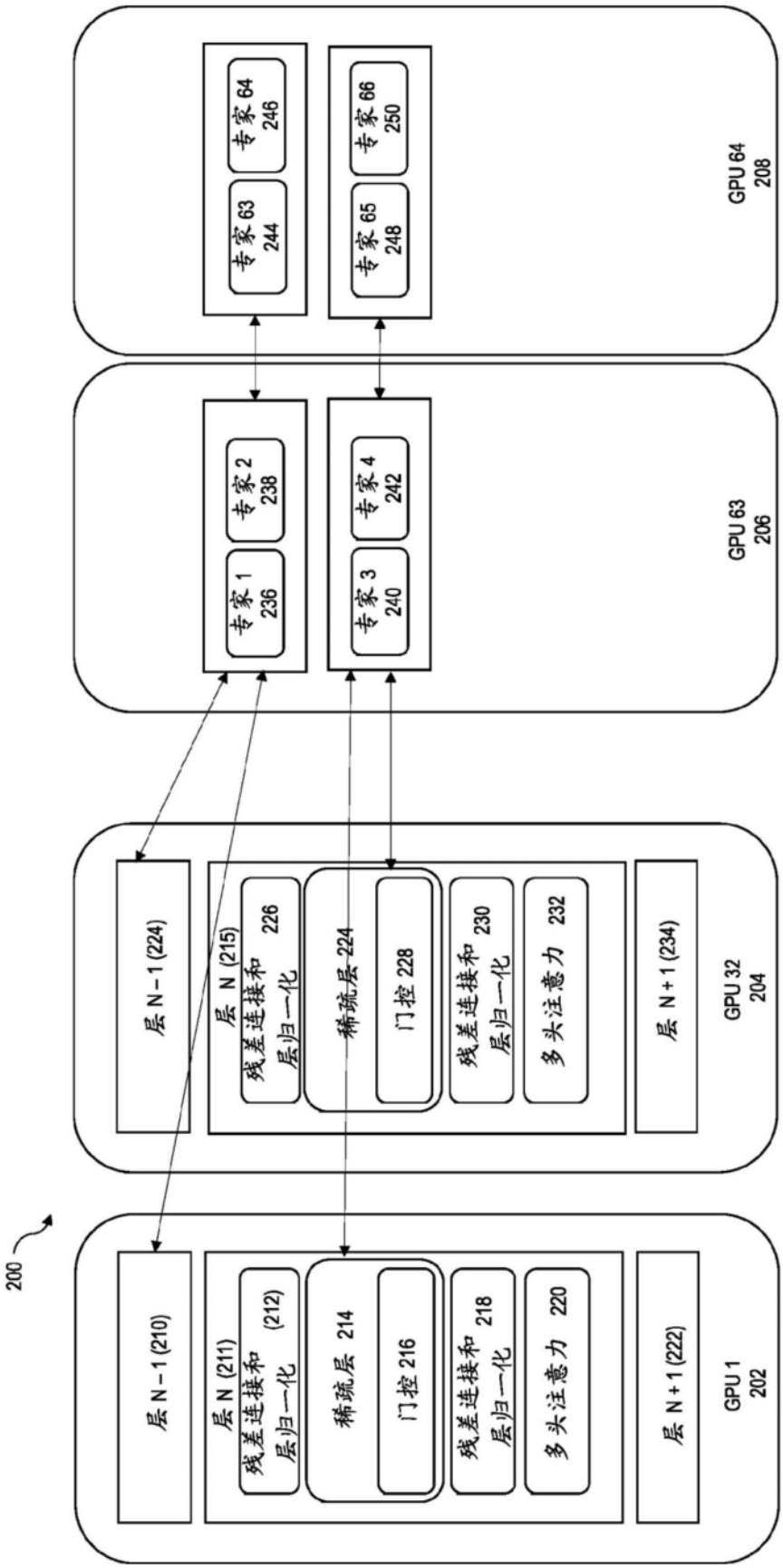

200
GPU 1
202
层 N-1 (210)
层 N (211)
残差连接和层归一化 (212)
稀疏层 214
门控 216
残差连接和层归一化 218
多头注意力 220
层 N+1 (222)
GPU 32
204
层 N-1 (224)
层 N (215)
残差连接和层归一化 226
稀疏层 224
门控 228
残差连接和层归一化 230
多头注意力 232
层 N+1 (234)
GPU 63
206
专家 1 236
专家 2 238
专家 3 240
专家 4 242
GPU 64
208
专家 63 244
专家 64 246
专家 65 248
专家 66 250

图2

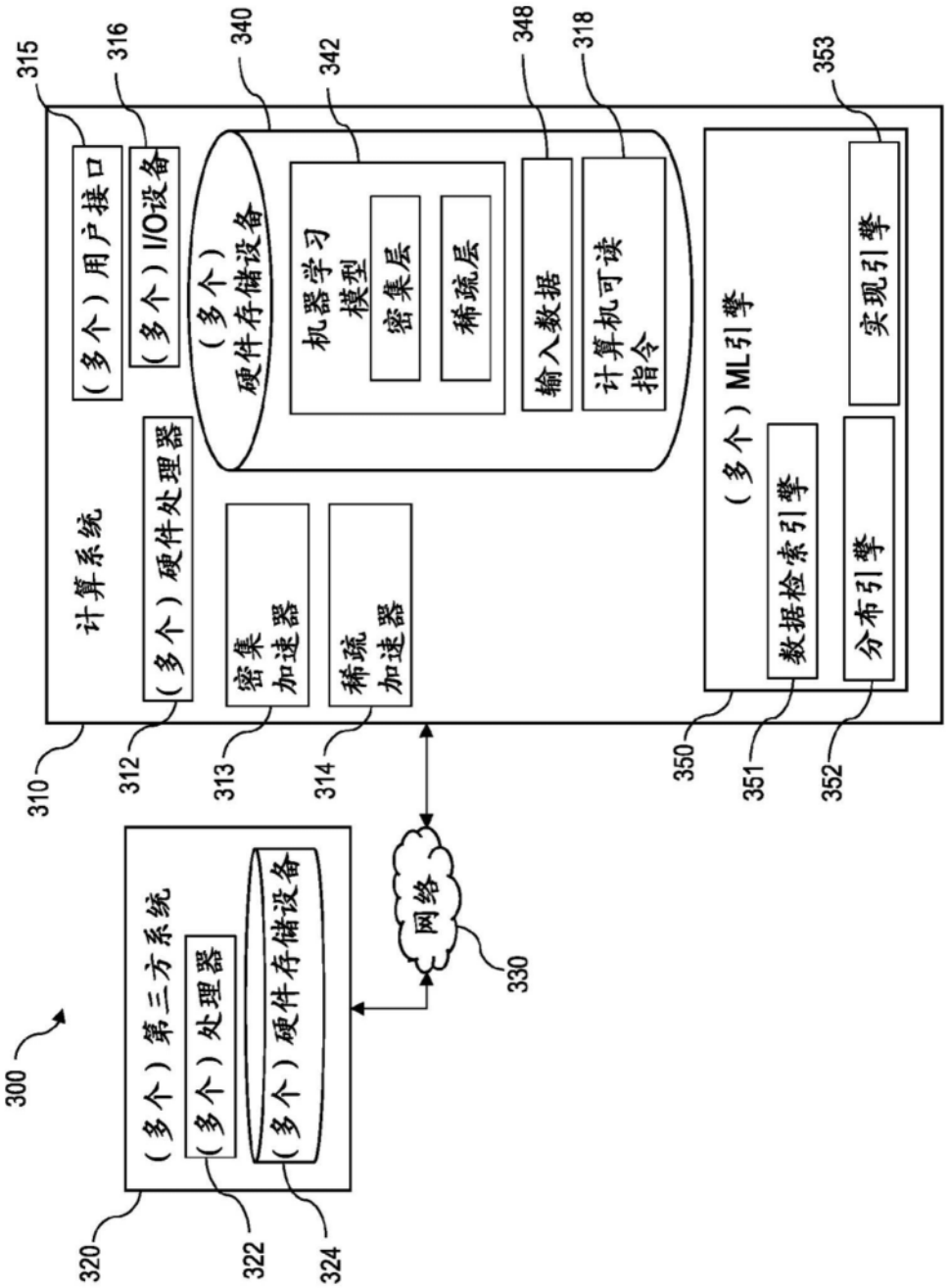

计算系统
(多个)用户接口
(多个)I/O设备
(多个)硬件处理器
(多个)硬件存储设备
机器学习模型
密集层
稀疏层
输入数据
计算机可读指令
密集加速器
稀疏加速器
(多个)ML引擎
数据检索引擎
分布引擎
实现引擎
网络
(多个)第三方系统
(多个)处理器
(多个)硬件存储设备
300
310
312
313
314
315
316
318
320
322
324
330
340
342
348
350
351
352
353

图3

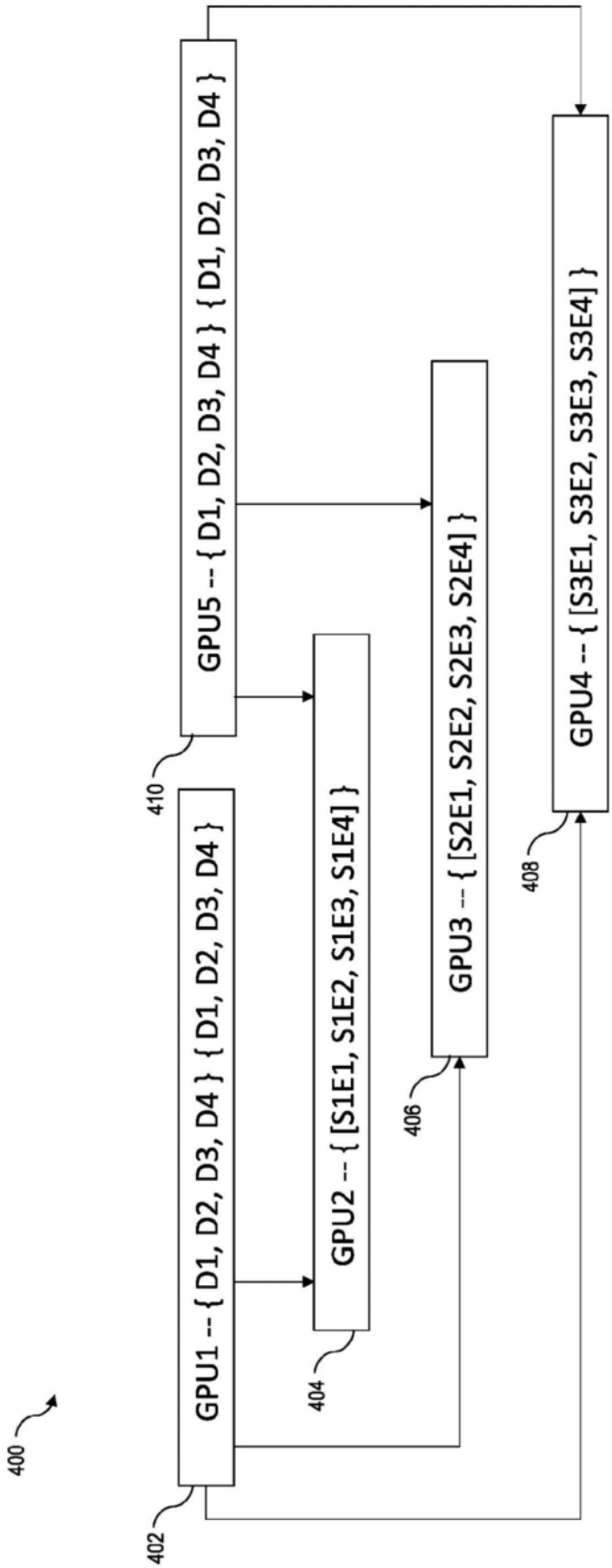

400
GPU1 -- { D1, D2, D3, D4 } { D1, D2, D3, D4 }
GPU2 -- { [S1E1, S1E2, S1E3, S1E4] }
GPU3 -- { [S2E1, S2E2, S2E3, S2E4] }
GPU4 -- { [S3E1, S3E2, S3E3, S3E4] } }
GPU5 -- { D1, D2, D3, D4 } { D1, D2, D3, D4 }
402
404
406
408
410

图4A

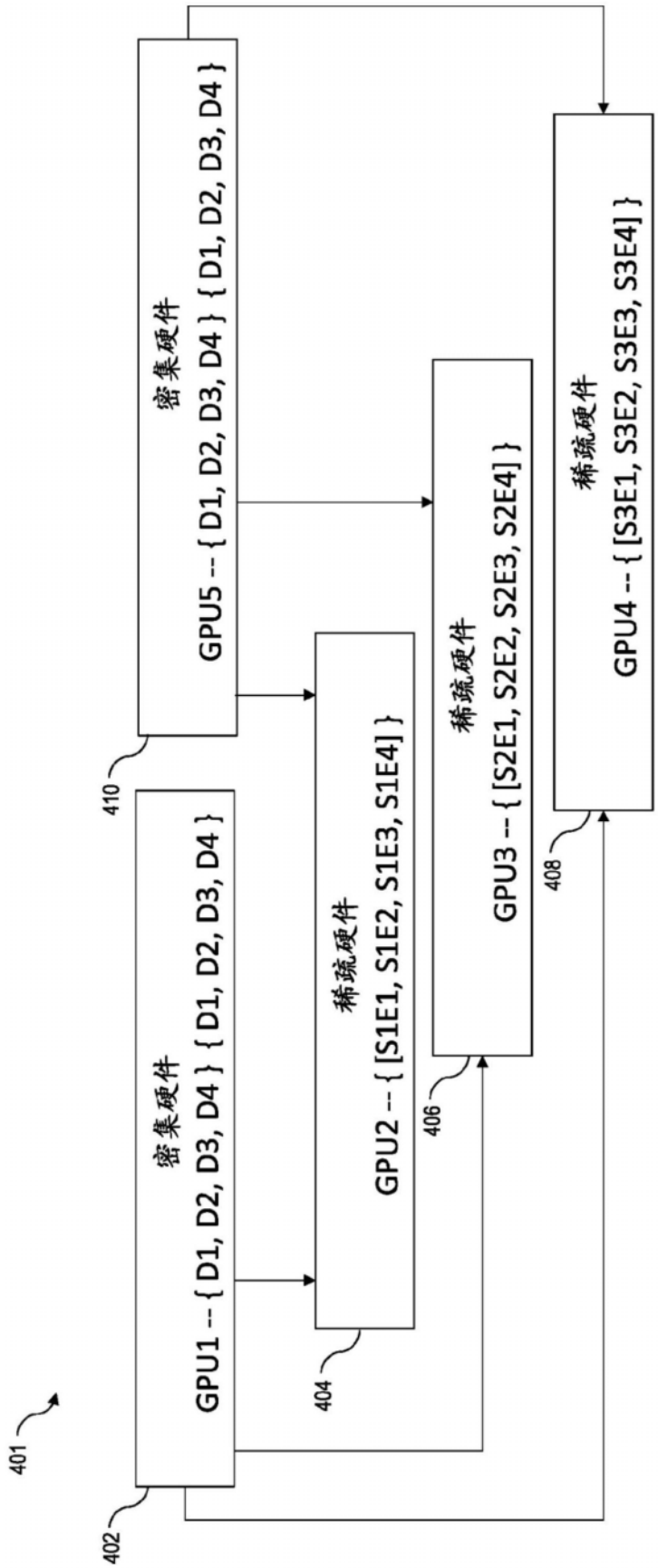

密集硬件
GPU1 -- { D1, D2, D3, D4 } { D1, D2, D3, D4 }
密集硬件
GPU5 -- { D1, D2, D3, D4 } { D1, D2, D3, D4 }
稀疏硬件
GPU2 -- { [S1E1, S1E2, S1E3, S1E4] }
稀疏硬件
GPU3 -- { [S2E1, S2E2, S2E3, S2E4] }
稀疏硬件
GPU4 -- { [S3E1, S3E2, S3E3, S3E4] }
401
402
404
406
408
410

图4B

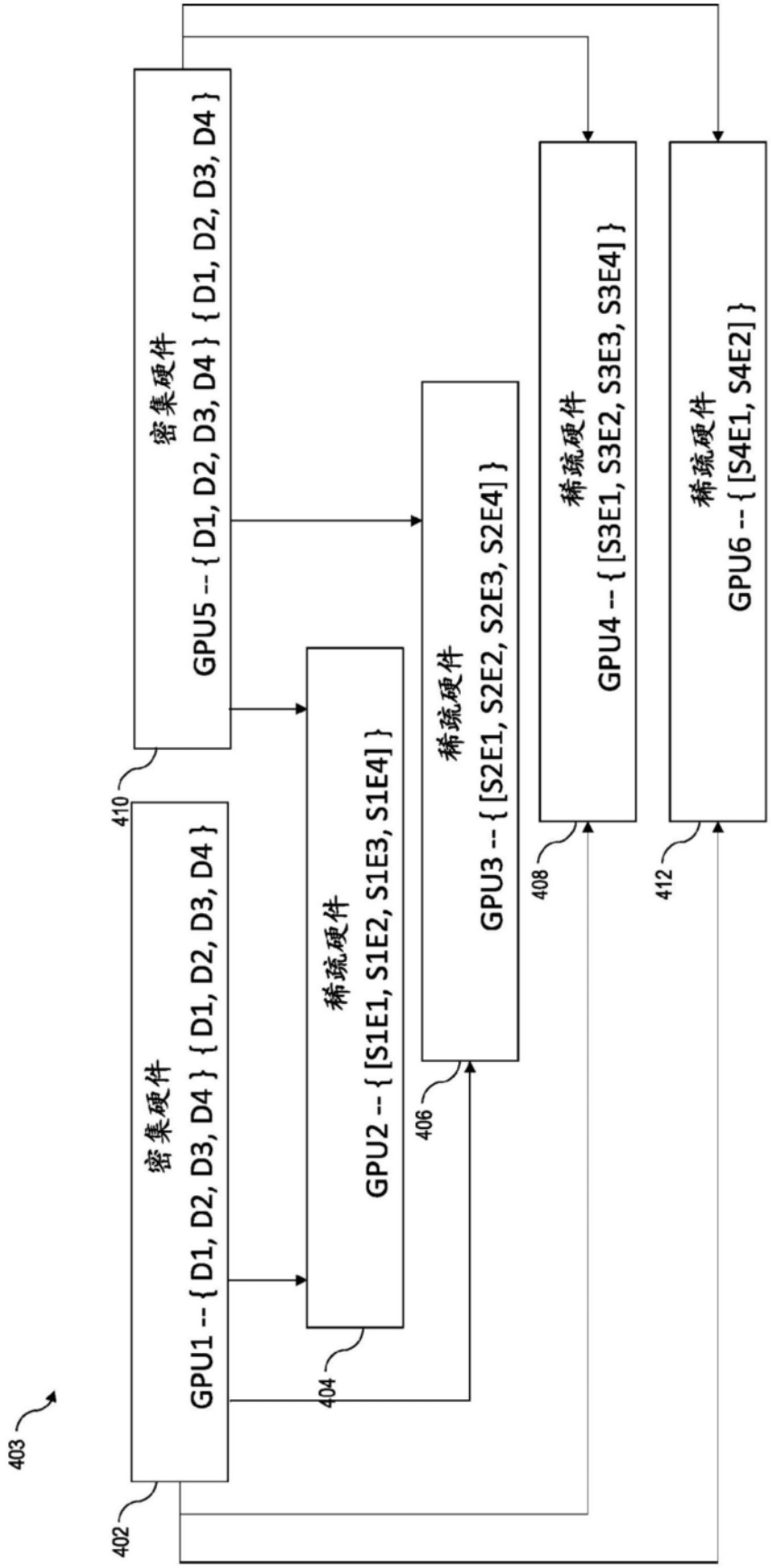

密集硬件
GPU1 -- { D1, D2, D3, D4 } { D1, D2, D3, D4 }
402
稀疏硬件
GPU2 -- { [S1E1, S1E2, S1E3, S1E4] }
404
稀疏硬件
GPU3 -- { [S2E1, S2E2, S2E3, S2E4] }
406
稀疏硬件
GPU4 -- { [S3E1, S3E2, S3E3, S3E4] }
408
稀疏硬件
GPU6 -- { [S4E1, S4E2] }
412
密集硬件
GPU5 -- { D1, D2, D3, D4 } { D1, D2, D3, D4 }
410
403

图4C

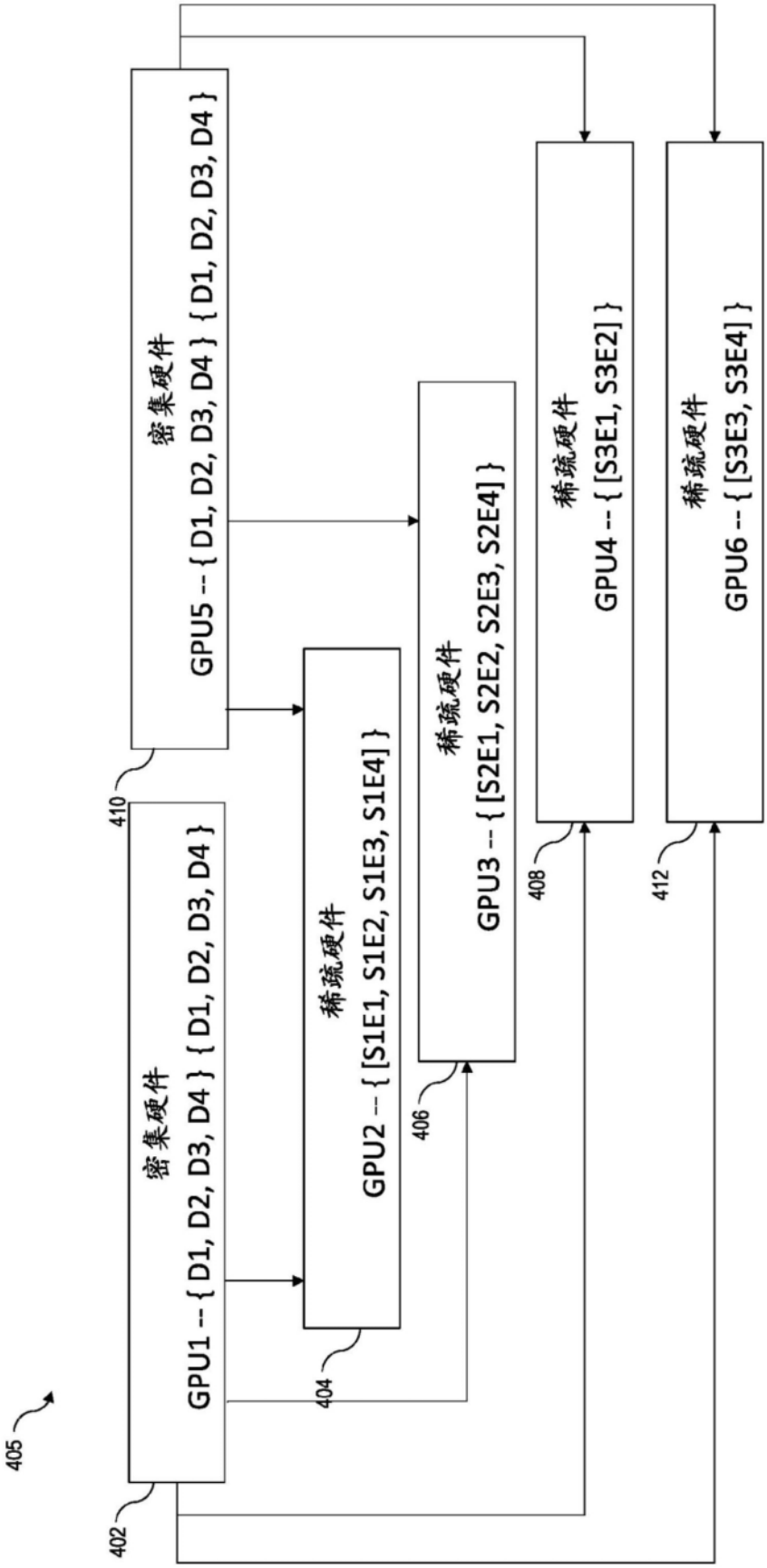

密集硬件
GPU1 -- { D1, D2, D3, D4 } { D1, D2, D3, D4 }
稀疏硬件
GPU2 -- { [S1E1, S1E2, S1E3, S1E4] }
稀疏硬件
GPU3 -- { [S2E1, S2E2, S2E3, S2E4] }
稀疏硬件
GPU4 -- { [S3E1, S3E2] }
稀疏硬件
GPU6 -- { [S3E3, S3E4] }
密集硬件
GPU5 -- { D1, D2, D3, D4 } { D1, D2, D3, D4 }
405
402
404
406
408
410
412

图4D

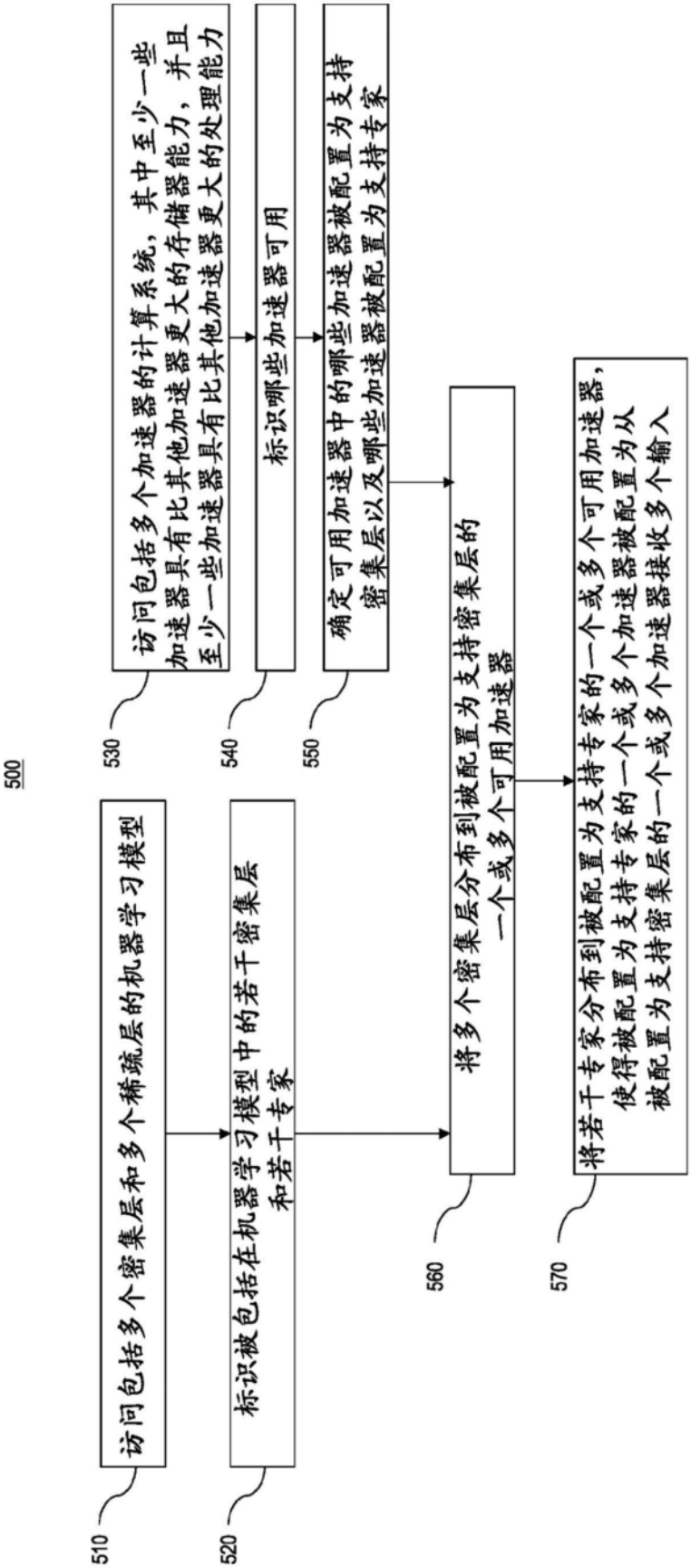

500
访问包括多个密集层和多个稀疏层的机器学习模型
510
标识被包括在机器学习模型中的若干密集层和若干专家
520
访问包括多个加速器的计算系统，其中至少一些加速器具有比其他加速器更大的存储器能力，并且至少一些加速器具有比其他加速器更大的处理能力
530
标识哪些加速器可用
540
确定可用加速器中的哪些加速器被配置为支持密集层以及哪些加速器被配置为支持专家
550
将多个密集层分布到被配置为支持密集层的一个或多个可用加速器
560
将若干专家分布到被配置为支持专家的一个或多个可用加速器，使得被配置为支持专家的一个或多个加速器被配置为从被配置为支持密集层的一个或多个加速器接收多个输入
570

图5

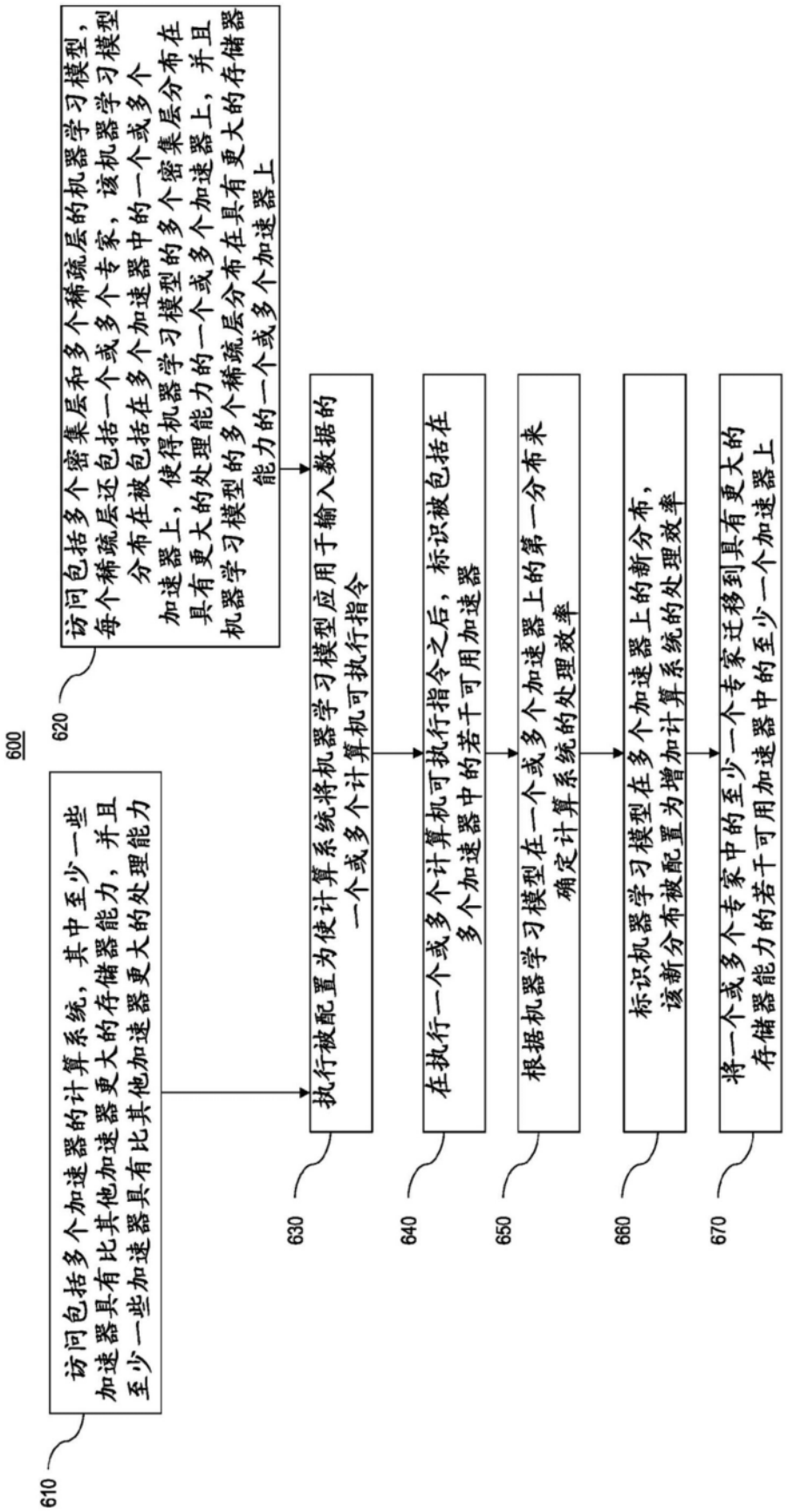

600
610
访问包括多个加速器的计算系统，其中至少一些加速器具有比其他加速器更大的存储能力，并且至少一些加速器具有比其他加速器更大的处理能力
620
访问包括多个密集层和多个稀疏层的机器学习模型，每个稀疏层还包括一个或多个专家，该机器学习模型分布在被包括在多个加速器中的一个或多个加速器上，使得机器学习模型的多个密集层分布在具有更大的处理能力的一个或多个加速器上，并且机器学习模型的多个稀疏层分布在具有更大的存储器能力的一个或多个加速器上
630
执行被配置为使计算系统将机器学习模型应用于输入数据的一个或多个计算机可执行指令
640
在执行一个或多个计算机可执行指令之后，标识被包括在多个加速器中的若干可用加速器
650
根据机器学习模型在一个或多个加速器上的第一分布来确定计算系统的处理效率
660
标识机器学习模型在多个加速器上的新分布，该新分布被配置为增加计算系统的处理效率
670
将一个或多个专家中的至少一个专家迁移到具有更大的存储器能力的若干可用加速器中的至少一个加速器上

图6