(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117291239 A
(43)申请公布日 2023.12.26

(21)申请号 202311182467.8
(22)申请日 2018.01.17
(30)优先权数据
15/599,559 2017.05.19 US
(62)分案原申请数据
201880019345.8 2018.01.17
(71)申请人 谷歌有限责任公司
地址 美国加利福尼亚州
(72)发明人 禹同爀
(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
专利代理师 邓聪惠 李宝泉
(51)Int.Cl.
G06N 3/063(2023.01)
G06F 9/50(2006.01)
G06F 9/48(2006.01)

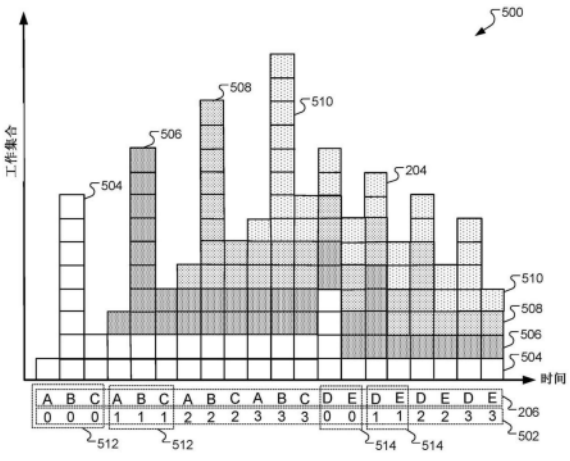
权利要求书3页 说明书13页 附图7页

(54)发明名称

调度神经网络处理

(57)摘要

本申请公开了调度神经网络处理。一种计算机实现的方法包括接收要在硬件电路上使用神经网络处理的一批神经网络输入。神经网络具有以有向图布置的多个层并且每个层具有相应的参数集合。所述方法包括确定神经网络层到超层序列的划分。每个超层是有向图的包括一个或多个层的划分。所述方法包括使用硬件电路处理该批输入,其包括:针对序列中的每个超层:i)将超层中的层的相应的参数集合加载到硬件电路的存储器中;以及ii)针对该批中的每个输入,使用硬件电路的存储器中的参数通过超层中的层中的每个层处理输入以生成针对输入的超层输出。



CN 117291239 A

1.一种使用在硬件集成电路上实现的神经网络用于处理一批神经网络输入的方法,所述神经网络包括以有向图布置的多个神经网络层,所述方法包括:

接收一批神经网络输入;

识别被划分为包括多个超层的序列的所述神经网络的层,其中,每个超层包括两个或更多个神经网络层并且是所述有向图的划分;以及

使用所述集成电路处理所述批神经网络输入,包括:针对所述多个超层中的每个超层:

获得与所述批中的所述神经网络输入中的每个神经网络输入相对应的相应超层输入;

在使用所述超层中的所述神经网络层中的任何神经网络层来处理超层输入之前,将所述超层中的层中的每个层的相应参数集合加载到所述集成电路的参数存储器中;以及

使用从所述参数存储器获得的所述神经网络层的所述相应参数集合通过所述超层中的神经网络层来处理所述超层输入。

2.根据权利要求1所述的方法,进一步包括:

基于通过所述超层中的神经网络层对所述超层输入的处理,生成超层输出。

3.根据权利要求2所述的方法,其中,所述超层输出是所述序列中第一超层的输出,并且所述方法进一步包括:

将所述超层输出作为超层输入接收到所述序列中第二超层中的神经网络层;以及

针对与所述第一超层的所述超层输出相对应的所述超层输入,通过所述序列中的所述第二超层中的神经网络层处理所述超层输入。

4.根据权利要求1所述的方法,其中,针对所述超层中的所述神经网络层中的每个神经网络层加载所述相应参数集合,包括:

基于从所述集成电路和所述参数存储器外部的主机接收到的数据值,将所述相应参数集合预加载到所述参数存储器中。

5.根据权利要求4所述的方法,其中,使用所述集成电路处理所述批神经网络输入包括:

基于调度过程处理所述批神经网络输入,所述调度过程对在所述集成电路处实现的神经网络模型的批和层维度上的神经网络计算执行全局调度。

6.根据权利要求5所述的方法,其中,所述神经网络计算的所述全局调度是使用所述主机来执行的。

7.根据权利要求1所述的方法,其中:

所述序列中超层的每个神经网络层与相应工作集合相关联;以及

所述相应工作集合部分地由存储用于处理所述工作集合中超层输入的所述神经网络层的参数所需要的存储器的量来定义。

8.根据权利要求1所述的方法,其中:

所述序列中的第一超层表示所述有向图的第一划分;以及

所述序列中的第二超层表示所述有向图中的第二不同划分。

9.一种使用在硬件集成电路上实现的神经网络用于处理一批神经网络输入的系统,所述神经网络包括以有向图布置的多个神经网络层,所述系统包括:

所述集成电路、处理器和非暂时性计算机可读存储设备,所述非暂时性计算机可读存储设备用于存储可由所述处理器执行的指令,以引起包括以下操作的执行:

接收所述一批神经网络输入;

识别被划分为包括多个超层的序列的所述神经网络的层,其中,每个超层包括两个或更多个神经网络层并且是所述有向图的划分;以及

使用所述集成电路处理所述批神经网络输入,包括:针对所述多个超层中的每个超层:

获得与所述批中的所述神经网络输入中的每个神经网络输入相对应的相应超层输入;

在使用所述超层中的所述神经网络层中的任何神经网络层来处理超层输入之前,将所述超层中的层中的每个层的相应参数集合加载到所述集成电路的参数存储器中;以及

使用从所述参数存储器获得的所述神经网络层的所述相应参数集合通过所述超层中的神经网络层来处理所述超层输入。

10.根据权利要求9所述的系统,其中,所述操作包括:

基于通过所述超层中的神经网络层对所述超层输入的处理,生成超层输出。

11.根据权利要求10所述的系统,其中,所述超层输出是所述序列中第一超层的输出,并且所述操作进一步包括:

将所述超层输出作为超层输入接收到所述序列中第二超层中的神经网络层;以及

针对与所述第一超层的所述超层输出相对应的所述超层输入,通过所述序列中的所述第二超层中的神经网络层处理所述超层输入。

12.根据权利要求9所述的系统,其中,针对所述超层中的所述神经网络层中的每个神经网络层加载所述相应参数集合,包括:

基于从所述集成电路和所述参数存储器外部的主机接收到的数据值,将所述相应参数集合预加载到所述参数存储器中。

13.根据权利要求12所述的系统,其中,使用所述集成电路处理所述批神经网络输入包括:

基于调度过程处理所述批神经网络输入,所述调度过程对在所述集成电路处实现的神经网络模型的批和层维度上的神经网络计算执行全局调度。

14.根据权利要求13所述的系统,其中,所述神经网络计算的所述全局调度是使用所述主机来执行的。

15.根据权利要求9所述的系统,其中:

所述序列中超层的每个神经网络层与相应工作集合相关联;以及

所述相应工作集合部分地由存储用于处理所述工作集合中超层输入的所述神经网络层的参数所需要的存储器的量来定义。

16.根据权利要求9所述的系统,其中:

所述序列中的第一超层表示所述有向图的第一划分;以及

所述序列中的第二超层表示所述有向图中的第二不同划分。

17.一种非暂时性计算机可读存储设备,被配置为存储使用在硬件集成电路上实现的神经网络用于处理一批神经网络输入的指令:

所述神经网络包括以有向图布置的多个神经网络层,并且所述指令可由处理器执行,以引起包括以下操作的执行:

接收所述一批神经网络输入;

识别被划分为包括多个超层的序列的所述神经网络的层,其中,每个超层包括两个或

更多个神经网络层并且是所述有向图的划分;以及

使用所述集成电路处理所述批神经网络输入,包括:针对所述多个超层中的每个超层:

获得与所述批中的所述神经网络输入中的每个神经网络输入相对应的相应超层输入;

在使用所述超层中的所述神经网络层中的任何神经网络层来处理超层输入之前,将所述超层中的层中的每个层的相应参数集合加载到所述集成电路的参数存储器中;以及

使用从所述参数存储器获得的所述神经网络层的所述相应参数集合通过所述超层中的神经网络层来处理所述超层输入。

18.根据权利要求17所述的非暂时性计算机可读存储设备,其中,所述操作进一步包括:

基于通过所述超层中的神经网络层对所述超层输入的处理,生成超层输出。

19.根据权利要求18所述的非暂时性计算机可读存储设备,其中,所述超层输出是所述序列中第一超层的输出,并且所述操作进一步包括:

将所述超层输出作为超层输入接收到所述序列中第二超层中的神经网络层;以及

针对与所述第一超层的所述超层输出相对应的所述超层输入,通过所述序列中的所述第二超层中的神经网络层处理所述超层输入。

20.根据权利要求17所述的非暂时性计算机可读存储设备,其中,针对所述超层中的所述神经网络层中的每个神经网络层加载所述相应参数集合,包括:

基于从所述集成电路和所述参数存储器外部的主机接收到的数据值,将所述相应参数集合预加载到所述参数存储器中。

调度神经网络处理

[0001] 分案说明

[0002] 本申请属于申请日为2018年01月17日的中国发明专利申请201880019345.8的分案申请。

技术领域

[0003] 本说明书涉及用于执行神经网络计算的存储器管理过程。

背景技术

[0004] 神经网络是采用一个或多个操作层以针对接收到的输入生成输出的机器学习模型,所述输出例如是分类。除了输出层之外,一些神经网络还包括一个或多个隐藏层。每个隐藏层的输出被用作对网络中的下一层——即下一隐藏层或网络的输出层——的输入。网络中的层中的一些或全部根据相应的参数集合的当前值从接收到的输入生成输出。

[0005] 一些神经网络包括一个或多个卷积神经网络层。每个卷积神经网络层具有相关联的核集合。每个核包括由用户创建的神经网络模型建立的值。在一些实施方式中,核标识特定的图像轮廓、形状或颜色。核可以被表示为权重输入的矩阵结构。每个卷积层还可以处理激活输入集合。激活输入集合也可以被表示为矩阵结构。

发明内容

[0006] 本说明书中描述的主题包括用于接收要在硬件电路上使用神经网络处理的一批神经网络输入的系统和方法。神经网络可以包括以有向图布置的多个层并且每个层可以具有相应的参数集合。根据所描述的技术的方法包括确定神经网络层到超层(superlayer)序列的划分。每个超层可以是有向图的包括一个或多个层的划分。

[0007] 所描述的方法可以包括:使用硬件电路处理该批输入。例如,处理一批输入可以包括将序列的每个超层中的层的相应参数集合加载到硬件电路的存储器中。另外,针对批中的每个输入,所描述的方法可以包括使用硬件电路的存储器中的参数通过超层中的层中的每个层处理输入以基于输入生成超层输出。

[0008] 本说明书中描述的主题的一个方面可以体现在计算机实现的方法中。方法包括:接收要在硬件电路上使用神经网络处理的一批神经网络输入,神经网络具有以有向图布置的多个层,每个层具有相应的参数集合;以及确定神经网络层到超层序列的划分,其中每个超层是有向图的包括一个或多个层的划分。

[0009] 方法进一步包括使用硬件电路处理该批神经网络输入,包括:针对序列中的每个超层:将超层中的层的相应的参数集合加载到硬件电路的存储器中;以及针对批中的每个神经网络输入:使用硬件电路的存储器中的参数通过超层中的层中的每个层处理与神经网络输入相对应的超层输入以针对神经网络输入生成超层输出。

[0010] 这些和其他实施方式均可以可选地包括以下特征中的一个或多个。例如,在一些实施方式中,针对序列中的第一超层,与神经网络输入相对应的超层输入是神经网络输入。

在一些实施方式中,对在第一超层输出之后的每个超层的超层输入是由序列中的前一超层生成的超层输出。

[0011] 在一些实施方式中,使用硬件电路处理该批神经网络输入包括,针对每个超层:通过超层中的层中的每个层按顺序处理与该批神经网络输入相对应的超层输入,使得在与批中的第二神经网络输入相对应的超层输入随后通过超层中的层中的每个层被处理之前,针对批中的第一神经网络输入的超层输入通过超层中的层中的每个层被处理。

[0012] 在一些实施方式中,相应的超层中的层与工作集合相关联,其中每个工作集合至少由以下定义:i)要在硬件电路上使用神经网络处理的该批神经网络输入中的一个或多个输入,或者超层的前一层的一个或多个输出;以及ii)大小参数,所述大小参数指示通过超层中的层中的每个层处理一个或多个输入所需的存储器的量。

[0013] 在一些实施方式中,确定神经网络层到超层序列的划分包括:i)确定至少一个工作集合的特定大小参数;ii)确定硬件电路的存储器的特定总计的参数容量;以及iii)基于以下中的至少一个:至少一个工作集合的特定大小参数或者硬件电路的存储器的特定总计的参数容量,确定神经网络层到超层序列的划分。

[0014] 在一些实施方式中,硬件电路的存储器具有阈值存储容量,并且确定神经网络层到超层序列的划分包括:基于硬件电路的存储器的阈值存储容量来将神经网络层划分为超层序列。

[0015] 在一些实施方式中,神经网络层被划分为超层序列以在硬件电路处理该批神经网络输入时不超过存储器的阈值存储容量。

[0016] 在一些实施方式中,该批神经网络输入和相应的参数集合从硬件电路外部的源被接收,并且其中,通过超层的每个层处理与神经网络输入相对应的超层输入包括在不接收来自外部源的任何附加参数的情况下处理超层输入。

[0017] 这个方面和其他方面的其他实施方式包括:被配置为执行方法的动作的对应的系统、装置和编码在计算机存储设备上的计算机程序。一个或多个计算机或硬件电路的计算系统可以借助于安装在系统上的软件、固件、硬件或者它们的组合来这样配置,所述系统在操作中使系统执行动作。一个或多个计算机程序可以借助于具有指令来配置,所述指令在被数据处理装置执行时使装置执行动作。

[0018] 本说明书中描述的主题可以在特定实施例中实现以实现以下优点中的一个或多个。通过将神经网络层划分为超层序列,在神经网络使用参数集合处理输入时,由神经网络硬件电路进行的外部通信可以被最小化。在计算过程期间最小化的由硬件电路进行的外部通信可以导致改进的由硬件电路进行的带宽消耗和能量优化。

[0019] 进一步地,超层序列可以提供全局调度过程,所述全局调度过程混合神经网络模型的“批”和“层”维度以优化一个或多个存储器工作集合以便通过神经网络层处理输入。例如,通过对批和层维度执行全局调度,神经网络应用的实时存储器工作集合可以被最小化,从而增强给定硬件电路的输入的无批执行。实时存储器工作集合可以与通过神经网络的层处理的数据相对应,其中数据当前存在于数据处理装置或处理器硬件电路的物理存储器空间中。

[0020] 另外,示例硬件电路可以包括片上存储器(例如,SRAM),使得最小化的工作集合的输入和参数可以使用SRAM容量被存储在片上。因此,如果在SRAM容量基于提供超层序列的

全局调度过程被有效地利用时不再需要附加的存储器资源来存储输入和参数,则可以实现成本节省。在一些实施方式中,片上SRAM容量可以根据需要按比例放大或缩小以满足特定的设计要求并且提供可以或可以不包括形成超层序列的调度过程。

[0021] 在下面的附图和描述中陈述了本说明书中描述的主题的一个或多个实施方式的细节。主题的其他潜在特征、方面和优点将通过描述、附图和权利要求书而变得显而易见。

附图说明

[0022] 图1图示了用于通过均具有相应的参数集合的神经网络层处理神经网络输入的示例硬件电路。

[0023] 图2A图示了与使用神经网络的相应层处理单个批元素相关的示例图。

[0024] 图2B图示了与神经网络的给定层处理多批元素相关的示例图。

[0025] 图3图示了与在形成超层的神经网络的多个层中处理单个批元素相关的示例图。

[0026] 图4是通过神经网络的超层处理神经网络输入的方法的示例流程图。

[0027] 图5图示了表示被划分为超层序列以使用超层的多个层来处理单个批元素的神经网络层的示例图表。

[0028] 图6A图示了表示神经网络层的工作集合大小的示例图表。

[0029] 图6B图示了表示神经网络层的超层的工作集合大小的示例图表。

[0030] 在各个附图中,相同的附图标记和名称指示相同的要素。

具体实施方式

[0031] 具有多个层的神经网络可以被用于计算推断。例如,给定输入,神经网络可以计算针对输入的推断。神经网络通过经由神经网络中的层中的每个层处理输入来计算此推断。具体地,神经网络中的层可以以有向图布置,其中层中的一些或全部具有相应的参数集合。每个层接收输入并且根据该层的参数集合处理输入以生成输出。该输出可以用作下一神经网络层处的输入。

[0032] 因此,为了根据接收到的输入计算推断,神经网络接收输入并且通过有向图中的神经网络层中的每个层处理输入以生成推断,其中来自一个神经网络层的输出作为输入被提供到下一神经网络层。对神经网络层的数据输入,例如对神经网络的输入或者对神经网络层的在有向图中连接到该层的一个或多个层的输出,可以称为对层的激活输入。

[0033] 有向图中的任何特定层可以接收多个输入、生成多个输出或者两者。神经网络中的层也可以被布置为使得层的输出可以作为输入被发送回前一层。根据所描述的技术的方法可以包括:确定将神经网络层到超层序列的划分,使得每个超层是有向图的包括一个或多个层的划分。

[0034] 所描述的方法可以包括:在硬件电路上通过神经网络的序列中的相应超层的层来处理一批输入。处理该批输入可以包括:将层的参数加载到硬件电路的存储器中,并且使用参数来处理神经网络输入以针对输入生成相应的超层输出。

[0035] 在一些实施方式中,本说明书中描述的一个或多个功能可以使用系统的硬件电路或电子组件来执行。硬件电路可以从被电耦合至硬件电路的控制设备接收控制信号。硬件电路可以是包括一个或多个非暂时性机器可读存储介质(例如,存储器)的封装电子设备,

所述非暂时性机器可读存储介质被用于储对神经网络层的输入以及被用于处理该输入的参数。

[0036] 硬件电路可以包括形成封装集成电路或诸如处理器微芯片(例如,CPU或GPU)的处理器设备的多个组件。因此,在此实例中,相对于形成微芯片的多个其他组件,硬件电路的存储器可以是“片上(on chip)”存储器。如本说明书中所使用的,封装硬件电路或电子设备可以包括被密封或封闭在支撑壳体内的半导体材料,诸如硅片。支撑壳体可以包括从壳体的外围延伸出来以将设备连接至印刷电路板的一个导线。

[0037] 控制设备可以是与硬件电路间隔开并且至少在由硬件电路的组件封装(例如,支撑壳体)封闭的片上存储器外部的外部控制器。外部控制器可以是向硬件电路提供控制信号以使硬件电路使用上面讨论的输入和参数来执行神经网络推断计算的系统级控制器。外部控制器可以包括“片外(off-chip)”存储器,其中因为存储器不与封装硬件电路的片上存储器位于一处,所以存储器至少在芯片外。

[0038] 在一些实施方式中,在执行推断计算而不使用片外存储器时,外部控制器可以使用硬件电路的片上存储器来存储输入和参数。响应于从系统的至少一个控制器接收控制信号,硬件电路访问片上存储器并且使用所存储的输入和参数来执行神经网络计算。

[0039] 图1示出了可以被用于执行神经网络计算的硬件电路100的示例。执行神经网络计算可以包括通过神经网络中的层处理神经网络输入的电路100,所述神经网络中的层每一个均具有相应的参数集合。在一些实施方式中,电路100与包括一个或多个处理器、处理器微芯片或者体现神经网络的其他电路组件的硬件电路相对应。在其他实施方式中,电路100可以包括一个或多个硬件电路、处理器以及形成一个或多个神经网络的其他相关电路组件。通常,根据所描述的技术的方法可以被应用于各种处理器架构或者可以使用各种处理器架构来实现,所述处理器架构诸如是CPU、GPU、数字信号处理器(DSP)或其他相关处理器架构。

[0040] 电路100通常包括控制器108,所述控制器108提供一个或多个控制信号110以使与存储器104相关联的输入被存储到存储器102的存储器地址或者从存储器102的存储器地址被检索。同样地,控制器108还提供一个或多个控制信号110以使参数存储器106的参数被存储到存储器102的存储器地址或者从存储器102的存储器地址被检索。

[0041] 电路100进一步包括一个或多个乘法累加(MAC)单体/单元107、输入激活总线112和输出激活总线114。例如,控制信号110可以使存储器102向输入激活总线112提供一个或多个输入、使存储器102提供来自参数存储器106的一个或多个参数和/或使MAC单元/单元107使用输入和参数来执行产生被提供到输出激活总线114的输出激活的计算。

[0042] 控制器108可以包括一个或多个处理单元和存储器。控制器108的处理单元可以包括一个或多个处理器(例如,微处理器或中央处理单元(CPU))、图形处理单元(GPU)、专用集成电路(ASIC)或者不同处理器的组合。控制器108还可以包括提供附加处理选项的一个或多个的其他存储或计算资源/设备(例如,缓冲器、寄存器、控制电路系统等)以执行本说明书中描述的确定和计算。

[0043] 在一些实施方式中,控制器108的(多个)处理单元执行存储在存储器中的指令以使控制器108和电路100执行本说明书中描述的一个或多个功能。控制器108的存储器可以包括一个或多个非暂时性机器可读存储介质。本文描述的非暂时性机器可读存储介质可以

包括固态存储器、磁盘、光盘、便携式计算机磁盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(例如,EPROM、EEPROM或闪速存储器)或者能够存储信息的任何其他有形介质。

[0044] 电路100可以是示例计算单元或计算瓦片(tile),并且可以包括附加的硬件结构以执行与诸如张量、矩阵和/或数据阵列的多维数据结构相关联的计算。在一些实施方式中,使用由电路100从与神经网络计算系统相关联的外部或高级别控制设备接收到的数据值,输入值可以被预加载到激活存储器104和/或参数/权重值可以被预加载到参数存储器106。

[0045] 电路100可以接收定义要通过使用系统的神经网络来执行的特定计算操作的指令。一般而言,存储在存储器102中的数据值通常均被写入到相应的存储器地址位置。在需要诸如输入的数据值来执行特定计算操作时,存储器102中的地址位置然后可以被示例控制设备(例如,控制器108)访问。

[0046] 控制器108可以向存储器102提供一个或多个控制信号110以将输入从存储器102加载到输入激活总线112上并且向包括MAC 107的计算单元阵列提供这些值。激活存储器104的索引可以包括具有输入的全部存储器地址位置。数据总线112可由计算阵列的一个或多个单元访问。计算阵列的单元可以从数据总线112接收一个或多个激活值以基于接收到的激活值来执行与矩阵乘法相关的计算。

[0047] 针对给定的计算周期,电路100可能需要访问激活存储器104和参数存储器106的元素以执行与神经网络层的推断计算相关联的乘法操作。针对执行计算的周期,控制器108可以一次提供一个输入值,并且包括MAC单元107的计算单元阵列将使激活与权重/参数相乘以产生针对给定输入的不同输出激活。

[0048] 在一些实施方式中,计算单元阵列的每个MAC单元107可以负责神经网络层的不同输出深度。计算单元阵列可以由控制器108完全控制,并且控制器108可以基于对激活值的检测来确定何时需要执行特定计算。

[0049] 此外,输入值在到达电路100时可以被分析以存储在存储器102中。响应于分析输入,控制器108可以执行编程指令以通过仅将特定输入值(例如,仅非零激活值)存储在存储器102中来有效地压缩激活数据,从而节省存储器存储空间以及对应的带宽。

[0050] 在电路100接收到输入和参数时,控制器108可以例如执行一个或多个直接存储器访问操作。这些存储器访问操作的执行包括将与激活存储器104的维度元素相对应的输入存储在存储器102的地址位置中。同样地,控制器108还可以将与参数存储器106的维度元素相对应的参数存储在存储器102的地址位置中。控制器108可以进一步包括一个或多个保持存储器地址的地址寄存器,特定输入将从所述存储器地址中被提取。而且,一个或多个寄存器还将存储要与特定输入相乘的对应参数从其中被提取的存储器地址。

[0051] 在按顺序处理第一和第二输入时,控制器108可以引用上面提到的寄存器以确定第一输入的对应参数(和存储器地址)以及确定第二输入的对应参数(和存储器地址)。在一些实施方式中,在第一神经网络层处计算的输出激活被用作对网络中的下一/后续第二层的输入,所述下一/后续第二层例如是下一隐藏层或网络的输出层。通常,神经网络中的每个层根据相应的参数集合的当前值来从接收到的输入生成输出。

[0052] 在替代实施方式中,可能存在一些计算操作,其中,单个输入被用作覆盖参数存储

器106的给定维度元素的各种权重的多个乘法操作的操作数(例如,以迭代“X”或“Y”维度)。根据所描述的技术,电路100可以被配置为从计算系统或机器学习系统的外部控制器接收控制信号。外部控制器可以提供存储在电路100的片上存储器中的多批神经网络输入和参数。如下面更详细地描述的,外部控制器可以被配置为通过电路100上的神经网络实现用于批元素处理的调度策略。

[0053] 例如,系统的外部控制器可以向电路100提供控制信号以使电路100使用存储在电路100的片上存储器中的输入和参数通过神经网络中的层处理神经网络输入。根据所描述的技术,特定的调度策略可以被使用以将神经网络中的层划分为层的分组,所述层的分组形成一个或多个(下面描述的)超层序列。然后,系统控制器可以使用电路100访问存储在片上存储器中的输入和参数并且然后通过超层序列中的每个层处理神经网络输入的批。

[0054] 图2A图示了与使用神经网络的相应层处理单个批元素相关的示例图表200A。在一些实施方式中,下面描述的图表200A/图表200B以及图表300、图表500和图表600A/图表600B与可以表示神经网络的拓扑的示例有向图不同。

[0055] 图表200A示出了在通过神经网络中的层处理批元素期间工作集合的大小如何变化。工作集合的大小依据存储单元204来表示。通常,给定神经网络层的工作集合包括对神经网络层的输入、来自神经网络层的输出以及被用于通过神经网络层处理输入的参数。工作集合通常包括给定神经网络计算所需的一个或多个数据结构的分组并且在下面更详细地描述。

[0056] 一个或多个存储单元204被用于存储工作集合的输入以及相关联的神经网络层的参数。存储单元204可以与上述存储器102的存储器资源相关联。批元素是指在硬件电路上使用示例神经网络处理的单个神经网络输入。

[0057] 如上面提到的,神经网络可以包括被用于计算推断的多个层,并且推断通过经由神经网络中的层处理神经网络输入来计算。因此,图表200A进一步示出了神经网络层206,包括层A、层B、层C、层D和层E。图表200A示出了批元素首先通过层A处理,然后通过层B处理,然后通过层C处理,然后通过层D处理,然后通过层E处理。在一些实施方式中,层206中的相应层可以是以下类型的神经网络层中的一个:卷积层、约简层(reduction layer)、全连接(FC)层、分类器层、逐元素乘法层或者池化层,例如平均池化层或最大池化层。

[0058] 神经网络层的工作集合可以包括一个或多个批元素以及被用于通过神经网络中的相应层处理批元素的参数。工作集合可以由以下定义:i)要在硬件电路上使用神经网络处理的一批输入的一个或多个输入/批元素;以及ii)指示存储输入和参数所需的存储器的量的大小参数或存储单元204的数量。除了输入之外,工作集合还可以包括输出激活。在一些实施方式中,神经网络可以被描述为具有与上述批元素相关联的“批(batch)”维度以及与层206相对应的“层(layer)”维度。

[0059] 通常,参考例如图3至图6,图2A的以下描述提供了下文中描述的改进的神经网络调度过程的背景。例如,层206可以是示例机器学习模型的神经网络层,所述机器学习模型包括至少五个层(例如,层A、B、C、D和E)。由机器学习模型执行的推断计算可能会经历特征深度或输出跨步的突然或意外增加。在发生这种情况时,神经网络计算过程中的给定点处的有效工作集合可能会随时间增加输入和输出激活量或者减少输入和输出激活量。

[0060] 例如,如图2A所示,对于在层A处发生的批处理,由机器学习模型处理的单个批元

素的工作集合可能需要单个存储单元204。在层B处的批处理期间,针对给定工作集合所处理的输入激活的增加可能会发生。因此,在层B处的批处理期间,机器学习模型可能需要使用8个存储单元204,而不是在层A处的单个存储单元204。进一步,在图2A的实施方式中,在层C、D和E处处理的工作集合可能分别需要2、4和1个存储单元。

[0061] 在一些实施方式中,输入/输出激活量以及对应的所需存储单元的增加或减少可以基于神经网络中的均具有不同数量的参数或权重的层而发生。所以相对于层B,层A的工作集合可以包括更少的激活和参数,并且所以相对于可能需要更多存储资源的层B的较大工作集合,层A的工作集合可能仅需要较少的存储资源。

[0062] 在一些实施方式中,存储单元204可以与输入存储器104和参数存储器106的存储器资源相对应。例如,存储单元204可以与静态随机存取存储器(SRAM)的存储器资源相对应,所述静态随机存取存储器与电路100的硬件电路的上述电子组件的片上存储器相关联。包括存储器104、106的片上存储器资源可以具有固定的或阈值存储容量。此阈值存储容量可以小于或远小于与电路100的片外存储器相关联的动态随机存取存储器(DRAM)资源的存储容量。如上面所指示的,片外存储器可以是高级别外部控制设备的存储器。

[0063] 图2B图示了与神经网络的给定层处理多个批元素相关的示例图表200B。图表200B包括用于存储与批212的相应批元素相关联的工作集合的输入的存储单元的第一集208。图表200B进一步包括用于存储与批214的相应批元素相关联的工作集合的输入的存储单元的第二集210。

[0064] 在图2B的实施方式中,两个或更多个批均可以包括多个批元素,即批212可以具有至少一个单独的批元素“0”并且批214可以具有至少一个单独的批元素“1”。处理至少两个批212、214可以使给定工作集合的相对大小被批大小的因子放大。例如,如图2B所示,基于处理具有对应的批大小的至少两个批——批212和批214——的输入,层206中的每个层(层A至层E)处的工作集合大小可以被放大(例如,加倍)。

[0065] 如上面所讨论的,系统控制器可以被配置为包括编译时调度或其他计算逻辑以实现神经网络调度过程或策略,所述神经网络调度过程或策略定义多批输入通过神经网络中的一个或多个层被处理的方式。例如,电路100接收一批神经网络输入并且系统控制器确定针对输入应该如何被处理以对批中的每个输入执行推断的调度过程。处理输入使神经网络生成中间输入,诸如可以被提供到神经网络的后续层的输入激活。中间输入可以与第一神经网络层的作为输入激活被提供到后续神经网络层的输出激活相对应。

[0066] 在常规的调度策略中,神经网络通过第一神经网络层处理每个输入或批中的批元素以针对每个批元素生成层输出(输出激活)。然后通过第二神经网络层处理每个层输出,以此类推,直到批中的批元素的处理完成。即,在神经网络中的下一层的任何处理发生之前,针对批中的所有批元素执行给定层的处理。这种常规的神经网络调度策略可能被诸如存储器容量的约束限制,并且因此在最大化可用存储器和计算机器学习系统的资源的使用方面可能是低效的。

[0067] 在一些实施方式中,关于片上存储器的使用,所述片上存储器例如是示例硬件电路的存储器104、106的存储单元204,可以由片上存储器资源支持的最大批大小可以基于工作集合的大小来确定。具体地,由存储单元204支持的最大批大小可以部分地基于由给定的神经网络层处理的输入和参数的最大工作集合来确定。

[0068] 例如并且参考图2B,与存储器102和104相关联的总片上存储容量可以被限制于20个存储单元204。在图2B中,因为由层B处理的两个批元素的工作集合需要16个存储单元204,第三个批元素的处理将需要24个单元的存储单元204,并且因此,超过了20个存储单元容量。所以,在此示例中,在处理每个批元素需要至少8个存储单元时,神经网络可以仅支持包括两个批元素的特定最大工作集合大小。

[0069] 具体地,在图2B的实施方式中,如参考特征208所指示的,处理工作集合中的批元素“0”需要8个单元的存储,并且如参考特征210所指示的,处理工作集合中的批元素“1”也需要8个单元的存储。因此,因为处理批元素0和1共需要16个存储单元204,所以处理需要多于4个存储单元204的至少一个附加批元素将超过神经网络的硬件电路的可用存储器资源的片上存储容量(此处限于20个单元)。

[0070] 图3图示了示例图表300,所述示例图表300与处理形成一个或多个超层308和310的神经网络的多个层206之中的批元素相关,其中,超层308例如包括层A、B和C。图表300包括用于存储与相应批元素302中的批元素0相关联的工作集合的输入和参数的存储单元的第一集304。同样地,图表300进一步包括用于存储与相应批元素302中的批元素1相关联的工作集合的输入和参数的存储单元的第二集306,所述存储单元的第二集306在图3中示出为灰色。

[0071] 如上面所指示的,电路100可以包括示例电子组件或硬件电路,所述示例电子组件或硬件电路相对于电路100的其他组件或电路可以具有较少的片上或SRAM存储资源。然而,如本文所描述的,电路100可以被配置为使用可用的片上存储器执行计算密集的机器学习算法。在这些实例中,机器学习系统的神经网络可以包括加速器架构,所述加速器架构不对可以由硬件电路的片上存储器的存储单元204支持的最小或最大批大小施加不必要的约束。

[0072] 根据所描述的技术,改进的神经网络调度过程可以被用于有效地利用通过使用电路100的硬件电路的本地片上存储资源而提供的批局部性。进一步地,使用这种片上存储以及其他本地计算资源可以优化可用带宽并且节省带宽和能量敏感计算环境中的组件能耗。更进一步地,使用这种片上存储和其他本地资源可以用于在通过神经网络的层处理输入期间最小化由硬件电路进行的外部通信。

[0073] 例如,如上面简要提到的,实现神经网络的硬件电路可以与主机设备/外部控制器外部地通信以接收被神经网络用于计算推断的神经网络输入和参数。这些外部通信可能需要使用硬件电路的片上计算资源。因此,外部通信可以减少硬件电路的可用计算带宽、增加系统延迟并且还可能导致硬件电路的电子组件的能耗增加。

[0074] 鉴于与带宽和能耗相关的这些约束,本说明书描述了混合示例神经网络模型的“批”和“层”维度以优化特定存储器工作集合的使用的全局调度策略或过程。具体地,所描述的技术的实施方式可以包括利用机器学习模型的批和层维度以最小化由神经网络处理的批元素的有效工作集合的大小的灵活的神经网络调度策略。

[0075] 例如,根据所描述的教导的改进的神经网络调度过程使得有效工作集合的大小能够被调整,使得片上存储器104、106中的包括参数的工作集合的存储不超过片上存储器资源的阈值存储容量。因此,本文描述的方法使得能够实现由神经网络处理的批元素的有效调度。例如,效率可以基于调度策略来实现,所述调度策略使得工作集合能够以不对输入和

用于处理输入的参数的批大小施加不必要的约束的方式存储在硬件电路的片上存储中。

[0076] 进一步地,根据所描述的教导的改进的调度策略可以最大化用于存储输入和参数的可用片上资源的有效使用,使得访问片外资源的外部通信最小化。片上资源的有效使用和减少的外部通信可以导致可用系统带宽的增加以及系统组件的能耗的总体降低。

[0077] 在一些实施方式中,改进的调度过程或策略的各个方面可以使用软件指令或程序代码来编码。指令可以由以下来执行:电路100的至少一个处理器、控制器108的至少一个处理器,或者电路100或控制器108的示例硬件电路的至少一个处理器或者两者。

[0078] 图4是使用电路100通过神经网络的超层处理神经网络输入的方法400的示例流程图。方法或过程400与由神经网络进行的批元素处理的改进的调度策略相对应。在框402中,电路100接收要在系统的硬件电路上使用神经网络处理的一批神经网络输入。神经网络可以具有以有向图布置的多个层,并且每个层可以具有相应的参数集合。如上面所讨论的,在一些实施方式中,电路100的硬件电路可以从示例神经网络硬件系统的主机接口设备或较高级别的控制器接收输入。

[0079] 在框404中,电路100确定神经网络层到超层序列的划分。例如,电路100可以包括或者访问被配置为确定神经网络层到超层序列的一个或多个划分的编译器逻辑。可替代地或者除了编译器逻辑之外,电路100可以包括或者访问至少一个被配置为确定神经网络层到超层序列的一个或多个划分的硬件块。在一些实施方式中,超层序列中的每个超层是有向图的包括一个或多个层的划分。

[0080] 在框406中,电路100使用系统的硬件电路处理该批神经网络输入。在一些实施方式中,使用硬件电路处理一批神经网络输入可以包括将超层中的层的相应参数集合加载到存储器106中。在一些实例中,针对超层序列中的每个超层,加载超层中的层的参数。进一步地,使用硬件电路处理一批神经网络输入还可以包括:针对批中的每个神经网络输入,使用硬件电路的存储器中的参数通过超层中的层中的每个层处理神经网络输入以针对神经网络输入生成超层输出。

[0081] 针对序列中的第一超层,对该超层的神经网络输入(例如,超层输入)的输出是第一超层输出。另外,对第一超层之后的每个超层的超层输入是由序列中的前一超层生成的超层输出。在一些实施方式中,处理一批神经网络输入包括:通过序列中的第一超层的所有层处理输入,并且然后通过序列中的每个后续超层的全部层处理输入,直到已经通过神经网络中的全部超层处理了批中的所有输入。

[0082] 再次参考图3,在使用改进的神经网络调度过程时,对于多个层308和310,一个批元素可以以无批方式执行。根据所描述的技术,多个层308可以形成第一超层,而多个层310可以形成不同于第一超层的第二超层。下面参考图4更详细地描述被划分以形成超层的多个层的分组。

[0083] 如图3所示,在一些实施方式中,相对于处理较小的工作集合的情况下层C处的所需要的存储单元的数量,示例机器学习模型的层B可能需要大量存储单元204以处理大的工作集合。在批元素的工作集合足够小时,改进的调度过程可以包括:机器学习模型切换到由诸如超层/层308的多个层的特定分组(例如,超层)处理的下一批元素。

[0084] 例如,在电路100的硬件电路上实现的神经网络可以被配置为对神经网络的“批”和“层”维度执行全局调度。具体地,对神经网络层的输入的批处理可以通过在第一过程迭

代中针对第一批的元素0执行一组层308(A、B、C)并且然后在第二过程迭代中针对第二批的元素1执行同一组层(A、B、C)308来执行。

[0085] 如图3所示,相对于上述的常规调度策略的最大工作集合的大小,根据改进的调度策略在不同批元素之间交替减少了工作集合的最大大小。例如,至少对于批元素1的B层处的批处理,在不同批元素之间交替可以将层B的最大工作集合大小减小到10个单元,而不是在使用上述的常规调度策略时需要16个单元的最大工作集合大小。例如,8个单元可以用于批元素1的层B处的批处理,并且2个单元可以用于存储批元素0的层A、B、C处的先前批处理的输出和/或与批元素0相关联的工作集合的输入和参数以便在层D和E处处理

[0086] 图5图示了表示被划分为超层序列的神经网络层的示例图500,所述超层序列用于使用被划分以形成超层的多个层来处理至少单个批元素。图表500包括用于存储相应的批元素502的批元素0的工作集合的输入的存储单元的第一集504。

[0087] 同样地,图表500进一步包括:a)用于存储相应的批元素502的批元素1的工作集合的输入的存储单元的第二集506;b)用于存储相应的批元素502的批元素2的工作集合的输入的存储单元的第三集508;以及c)用于存储相应的批元素502的批元素3的工作集合的输入的存储单元的第四集510。

[0088] 图表500进一步包括沿着图表的X轴的超层序列。例如,图500包括:i)通过层A、B、C中的每个层处理批元素0、1、2和3的第一超层512;以及ii)通过层D、E中的每个层处理批元素0、1、2和3的第二超层514。根据所描述的教导,基于改进的神经网络调度策略定义的超层序列可以支持相对高的工作集合批大小而不超过执行神经网络的硬件电路的片上存储器容量或阈值容量。

[0089] 例如,如图5所示,在输入在示例“B3”层和批阶段期间被处理时,工作集合的最大大小针对四个批元素可能仅需要14个存储单元,例如,如由相应存储单元204的有区别的阴影图案所指示的批元素0、1、2和3。与常规的(例如,需要16个存储单元的)调度过程相比,所需存储单元的这种减少允许对所接收的和经由硬件电路的片上存储器存储的输入和参数的局部性的改进的利用。片上资源的这种改进的利用可以导致部分地基于片外或DRAM、存储器资源的减少使用来实现的增加的带宽和能量节省。

[0090] 另外,如上面简要提到的,改进的调度策略可以被用于处理一批或多批输入或输入而不超过电路100的硬件电路的片上存储器容量。在一些实施方式中,通过序列中的超层的层处理一批或多批神经网络输入可以包括由序列中的第一超层(512)生成第一超层输出以作为对后续层的输入由至少神经网络的后续层接收。

[0091] 在一些实例中,对超层序列中的第二超层的神经网络输入可以与由序列中的第一超层生成的第一超层输出相对应。进一步地,通过序列中的超层的层处理一批输入可以包括使用硬件电路的存储器中的参数通过第二超层中的每个层处理神经网络输入以针对与第一超层输出相对应的神经网络输入生成第二超层输出。

[0092] 在一些实施方式中,通过超层序列中的超层的层处理一批神经网络输入可以包括通过超层的每个层逐个处理批元素的输入。例如,处理一批输入可以包括通过超层中的每个层按顺序处理两个或更多个神经网络输入。这样的按顺序处理可以包括:通过超层的每个层处理第一神经网络输入并且然后通过超层的每个层处理第二神经网络输入。

[0093] 在一些实施方式中,针对序列中的每个超层,通过超层的层处理输入可以包括:通

过超层中的每个层按顺序处理与一批神经网络输入相对应的超层输入,使得在随后通过超层中的层中的每个处理与批中的第二神经网络输入相对应的超层输入之前,通过超层中的层中的每个处理批中的第一神经网络输入的超层输入。

[0094] 在一些实施方式中,超层序列中的第一超层可以包括单个神经网络层。在此实施方式中,通过超层序列处理输入可以包括通过包括单个神经网络层的第一超层处理第一输入。在此第一输入通过第一超层的单个层处理之后,在通过跟随序列中的第一超层的后续超层的全部层处理第一输入之前,可以立即通过第一超层处理第二输入。由序列中的后续超层处理的第一输入可以是包括单个神经网络层的第一超层的超层输出。

[0095] 超层和一个或多个超层序列可以基于根据改进的神经网络调度策略的划分多组层来形成。在一些实施方式中,电路100包括用于改进的调度策略的编程指令,并且这些指令可以包括确定神经网络层到超层序列的划分。每个超层可以是有向图的包括一个或多个层的划分。

[0096] 改进型调度过程的各个方面可以使神经网络层形成为多个超层,使得可以从电路100的硬件电路的片上存储装置访问给定超层的所有输入和参数。如上面所指示的,对输入和参数的片上访问可以最小化硬件电路的外部通信。例如,可以最小化外部通信,因为硬件电路可以避免与重复提取操作相关联的计算过程以从片外接口获得附加数量的输入和参数。

[0097] 在一些实施方式中,片外接口可以将硬件电路耦合至向电路100提供输入和参数的外部控制设备。具体地,超层序列中的每个超层可以接收用于处理该超层的一个或多个神经网络输入的特定数量的参数。在一些实例中,通过超层的层处理一个或多个神经网络输入可以包括在不接收后续的数量的参数的情况下处理输入以处理超层的特定数量的输入。

[0098] 在一些实施方式中,电路100执行程序代码以确定超层序列的一个或多个超层划分或分界。例如,电路100可以确定或计算给定层的激活工作集合和总计的参数容量的总和。电路100然后可以部分地基于硬件电路的存储器资源的预定义的或阈值片上存储容量(例如,存储器104和106)使用所确定的总和来确定神经网络层到超层序列的划分。因此,在电路100的硬件电路处理一批或多批神经网络输入时,神经网络层可以被划分为超层序列以不超过片上存储器的阈值存储容量。

[0099] 在一些实施方式中,确定神经网络层到超层序列的划分包括:i)电路100确定包括用于由神经网络处理的输入的至少一个工作集合的特定大小参数;ii)电路100确定硬件电路的存储器的特定合计的输入激活和参数容量;以及iii)电路100至少基于至少一个工作集合的特定大小参数或者硬件电路的存储器的特定合计的输入激活和参数容量来确定层到超层序列的划分。

[0100] 例如,片上存储器的存储容量或阈值容量可以是500兆字节(MB)。电路100可以基于等式1[总使用量=(工作集合*N)+参数]来确定总片上存储器使用量,其中等式1的变量N是批大小。电路100然后可以确定存储神经网络的每个层的相应参数集合所需的存储器的量。在一些实施方式中,参考图5,电路100可以确定:i)层A的参数集合需要25MB的存储器;ii)层B的参数集合需要125MB的存储器;以及iii)层C的参数集合需要50MB的存储器。

[0101] 因此,在此示例中,电路100确定层A、B和C的相应参数集合的总计的存储器使用量

是200MB,留下300MB的可用片上存储器用于存储输入(例如,500MB片上存储器容量减去200MB的总计的存储器使用量)。针对相应层A、B、C,电路100可以确定要由相应层处理的工作集合的输入的特定大小参数以及工作集合的对应的批大小。使用工作集合的输入的大小参数和对应的批大小,电路100可以确定存储器的总计的激活和参数容量。电路100可以使用存储器的总计的激活和参数容量来确定层到超层序列的划分。

[0102] 在一些实施方式中,电路100使用等式1、输入的大小参数(例如以存储器单元为单位)、批大小以及用于参数的总计的存储器以确定一组或多组层的总片上存储器使用量。电路100可以将每一组层的总存储器使用量与500MB片上存储容量相比较。电路100然后可以基于比较结果来确定形成超层序列的层的划分或分组。在硬件电路处理工作集合的一批神经网络输入时,电路100确定层到超层序列的划分以不超过片上存储器的阈值存储容量(500MB)。

[0103] 图6A图示了表示神经网络层的激活工作集合大小的示例图表600A,而图6B图示了表示神经网络的超层的激活工作集合大小的示例图表600B。如上面所讨论的,并且如图表600A和图表600B所指示的,在与被布置为超层的神经网络层的工作集合的大小相比时,未被布置为超层的神经网络层的工作集合可以包括明显更大的工作集合大小。

[0104] 例如,使用上述的传统调度策略的批处理的工作集合可以导致包括数百万输入的工作集合大小。在片上存储单元204被用于存储输入和被用于处理输入的参数时,这样的大量的输入可以超过硬件电路的片上存储器资源的存储或阈值容量。相反,基于本文描述的改进的调度策略,使用超层划分的批处理的工作集合可以导致包括明显更少的输入的工作集合大小。使用片上存储单元204可以有效地存储明显更少量的输入,使得不超过片上存储器容量。

[0105] 本说明书中所描述的主题和功能操作的实施例可以在包括在本说明书中所公开的结构及其结构等同物的数字电子电路、有形地体现的计算机软件或固件、计算机硬件或者它们中的一个或多个的组合中实现。本说明书中描述的主题的实施例可以被实现为一个或多个计算机程序,即,在有形的非暂时性程序载体上编码以由数据处理装置执行或者以控制所述数据处理装置的操作的计算机程序指令的一个或多个模块。

[0106] 替代地或者另外,程序指令可以在例如是机器生成的电气、光学或者电磁信号的人工生成的传播信号上编码,所述人工生成的传播信号被生成以编码信息以传输至合适的接收器装置以便由数据处理装置执行。计算机存储介质可以是机器可读存储设备、机器可读存储基底、随机或串行存取存储器设备或者它们中的一个或多个的组合。

[0107] 本说明书中所描述的过程和逻辑流程可以由执行一个或多个计算机程序的一个或多个可编程计算机来执行以通过对输入数据操作并且生成(多个)输出来执行功能。过程和逻辑流程也可以通过专用逻辑电路系统来执行,并且装置可以被实现为专用逻辑电路,所述专用逻辑电路例如是FPGA(现场可编程门阵列)、ASIC(专用集成电路)、GPGPU(通用图形处理单元)或一些其他处理单元。

[0108] 适合于执行计算机程序的计算机包括,例如,可以基于通用或者专用微处理器或两者或者任何其他类型的中央处理单元。通常,中央处理单元将接收来自只读存储器或者随机存取存储器或者两者的指令和数据。计算机的必要元件是用于执行或运行指令的中央处理单元以及用于存储指令和数据的一个或多个存储器设备。通常,计算机还将包括用于

存储数据的一个或多个大容量存储设备或者计算机被可操作地耦合以接收来自所述大容量存储设备的数据或者将数据传送至所述大容量存储设备或者两者,所述大容量设备例如是磁盘、磁光盘或者光盘。然而,计算机无需具有这样的设备。

[0109] 适合于存储计算机程序指令和数据的计算机可读介质包括全部形式的非易失性存储器、介质和存储器设备,包括:例如,半导体存储器设备,例如EPROM、EEPROM和闪速存储器设备;磁盘,例如内部硬盘或者可移动盘。处理器和存储器可以由专用逻辑电路系统补充或者可以并入所述专用逻辑电路系统中。

[0110] 虽然本说明书包含了许多具体实施细节,但是这些具体实施细节不应该被解释为对任何发明或者可能被要求保护的内容的范围的限制,而是作为针对特定发明的特定实施例的特征的描述。在本说明书中在单独实施例的上下文中描述的某些特征还可以以组合在单个实施例中实现。相反,在单个实施例的上下文中描述的各种特征也可以单独地或者按照任何合适的子组合在多个实施例中实现。而且,虽然特征在上文中可以被描述为以某些组合的方式起作用并且甚至最初要求这样的保护,但来自所要求保护的组合的一个或多个特征在某些情况下可以从组合中去除,并且所要求的组合可以针对子组合或者子组合的变化。

[0111] 同样地,虽然在附图中按照特定顺序描绘了操作,但是这不应该将其理解为需要以所示的特定次序或者以按顺序的次序来执行这样的操作或者执行全部图示的操作以实现期望的结果。在某些情况下,多任务处理和并行处理可能是有利的。而且,不应该将在上述实施例中描述的各种系统模块和组件的分离不应被理解为在全部实施例中都需要这种分离,并且应该理解的是,所描述的程序组件和系统通常可以被一起集成在单个软件产品中或者被封装到多个软件产品中。

[0112] 已经描述了本主题的特定实施例。其他实施例在以下权利要求书的范围内。例如,在权利要求书中叙述的动作可以以不同的顺序来执行并且仍然实现期望的结果。作为一个示例,在附图中描绘的过程不必须需要所示的特定次序或者按顺序的次序以实现期望的结果。在某些实施方式中,多任务处理和并行处理可以是有利的。

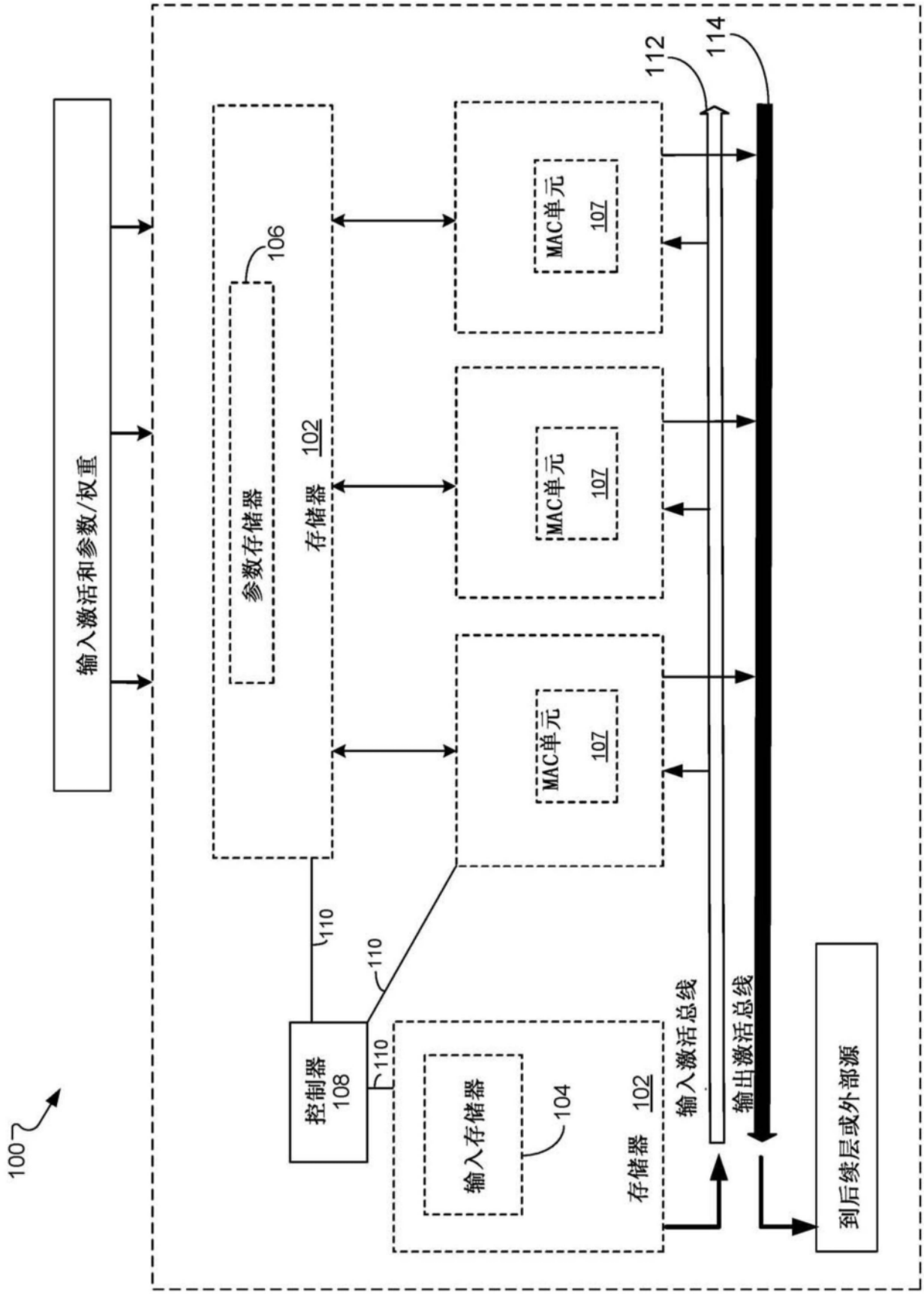

100
输入激活和参数/权重
参数存储器
106
存储器
102
MAC单元
107
MAC单元
107
MAC单元
107
112
114
110
110
110
控制器
108
输入存储器
104
存储器
102
输入激活总线
输出激活总线
到后续层或外部源

图1

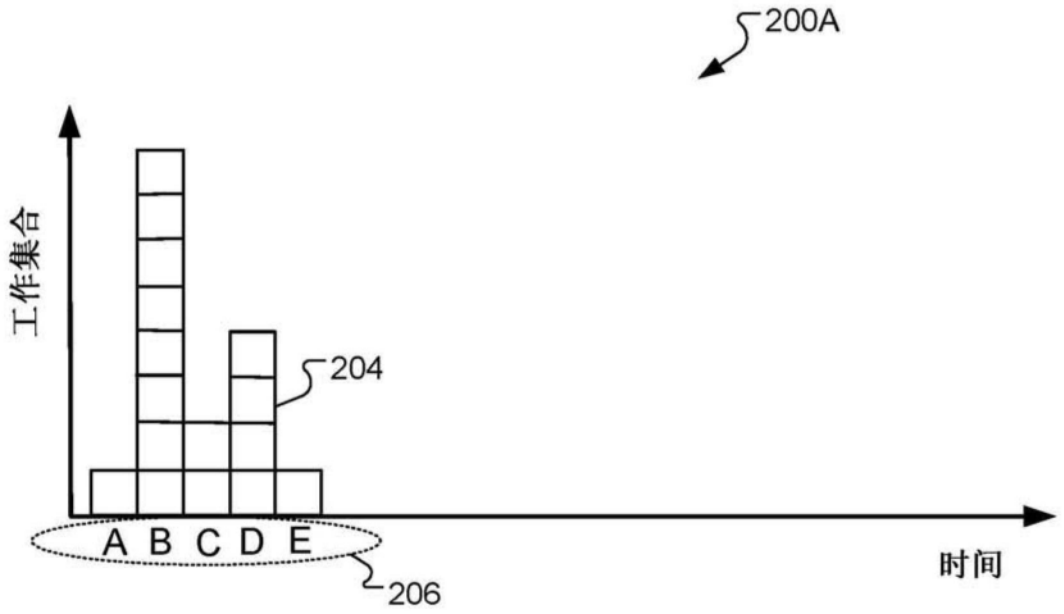

200A
工作集合
204
A B C D E
206
时间

图2A

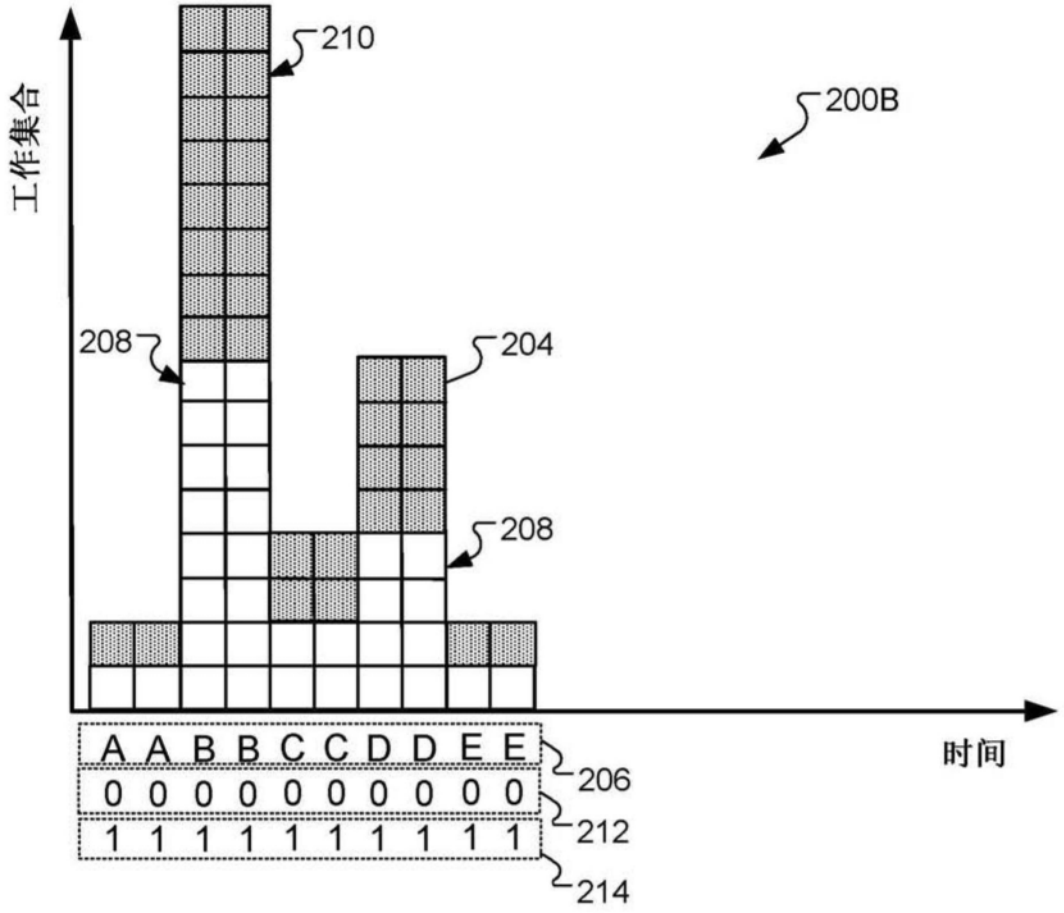

210
200B
工作集合
208
204
208
A A B B C C D D E E
206
0 0 0 0 0 0 0 0 0 0
212
1 1 1 1 1 1 1 1 1 1
214
时间

图2B

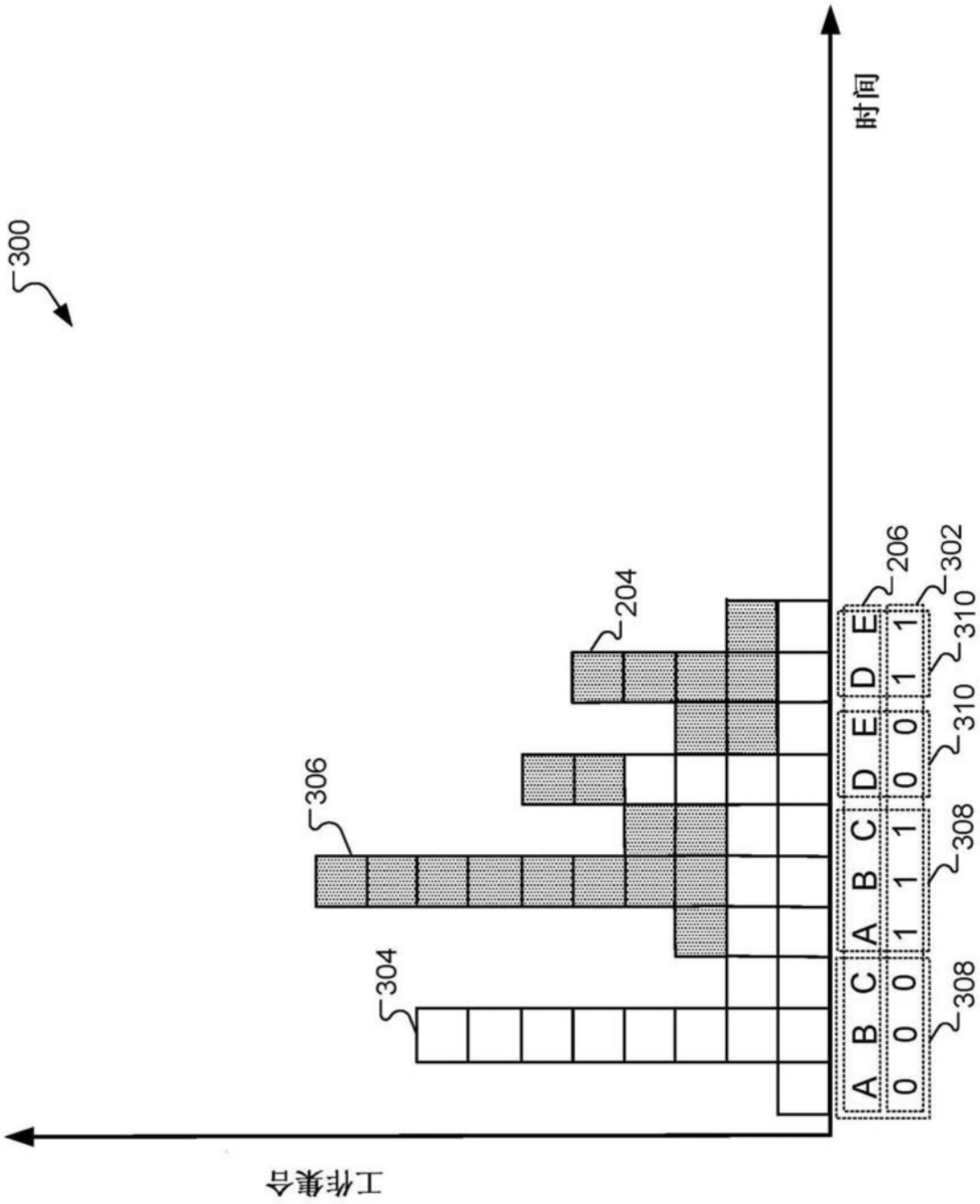

300
时间
工作集合
204
206
302
306
304
A B C D E
0 0 0 1 1
308
A B C D E
1 1 1 0 0
310
308
A B C D E
1 1 1 1 1
310
308

图3

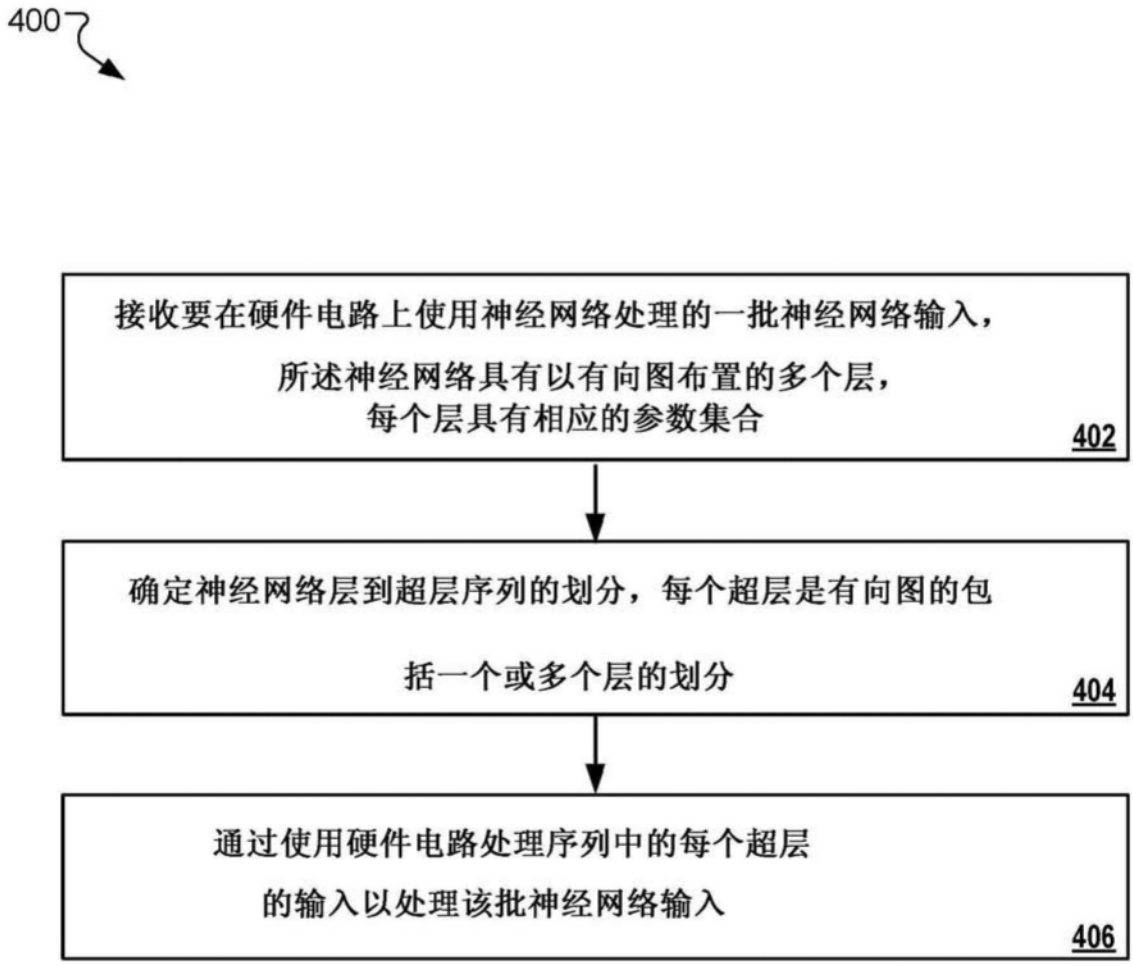

400
接收要在硬件电路上使用神经网络处理的一批神经网络输入，
所述神经网络具有以有向图布置的多个层，
每个层具有相应的参数集合
402
确定神经网络层到超层序列的划分，每个超层是有向图的包
括一个或多个层的划分
404
通过使用硬件电路处理序列中的每个超层
的输入以处理该批神经网络输入
406

图4

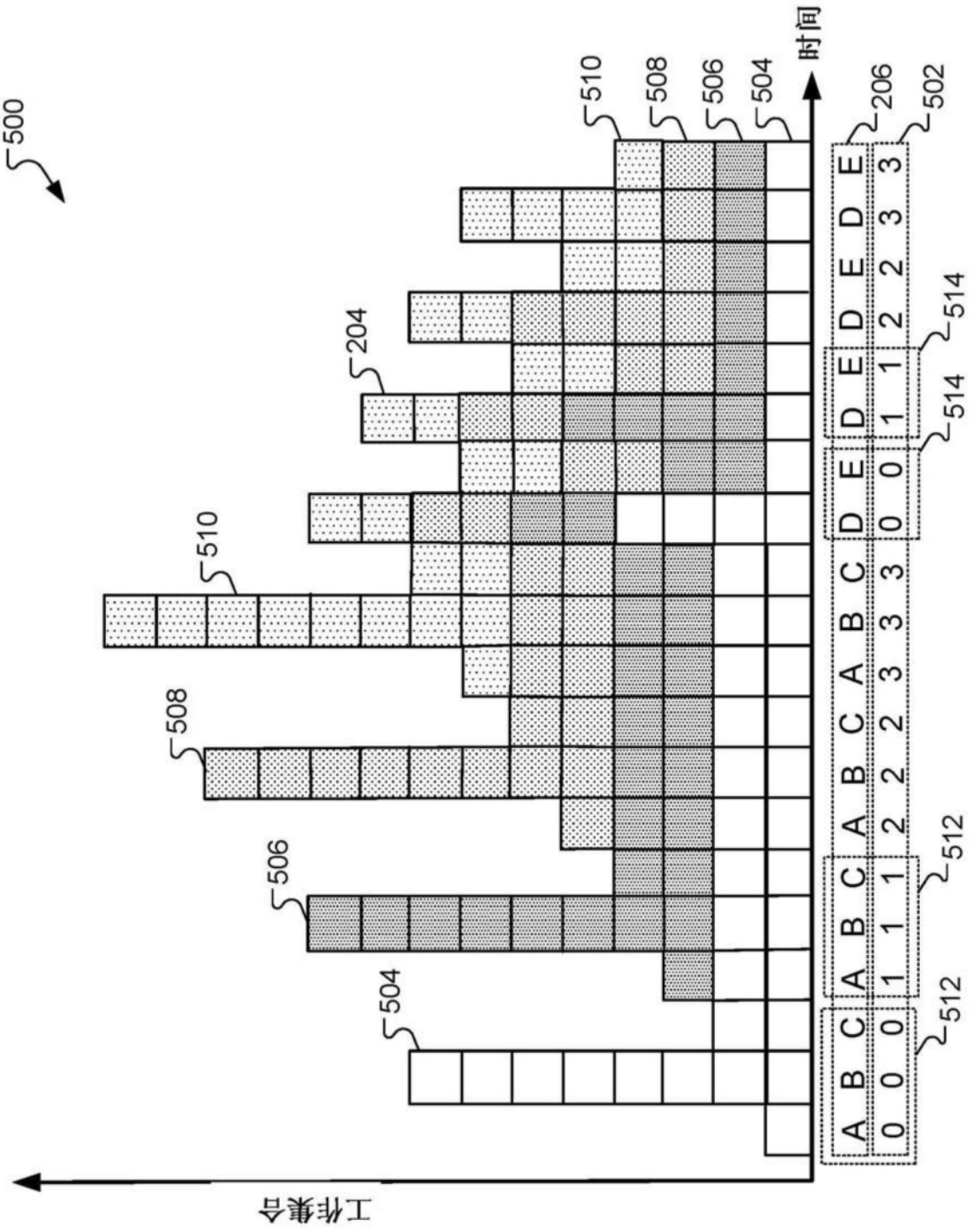

500
510
508
506
504
204
206
502
514
512
时间
工作集合
A B C A B C A B C D E D E D E D E
0 0 0 1 1 1 2 2 2 3 3 3 0 0 1 1 2 2 3 3

图5

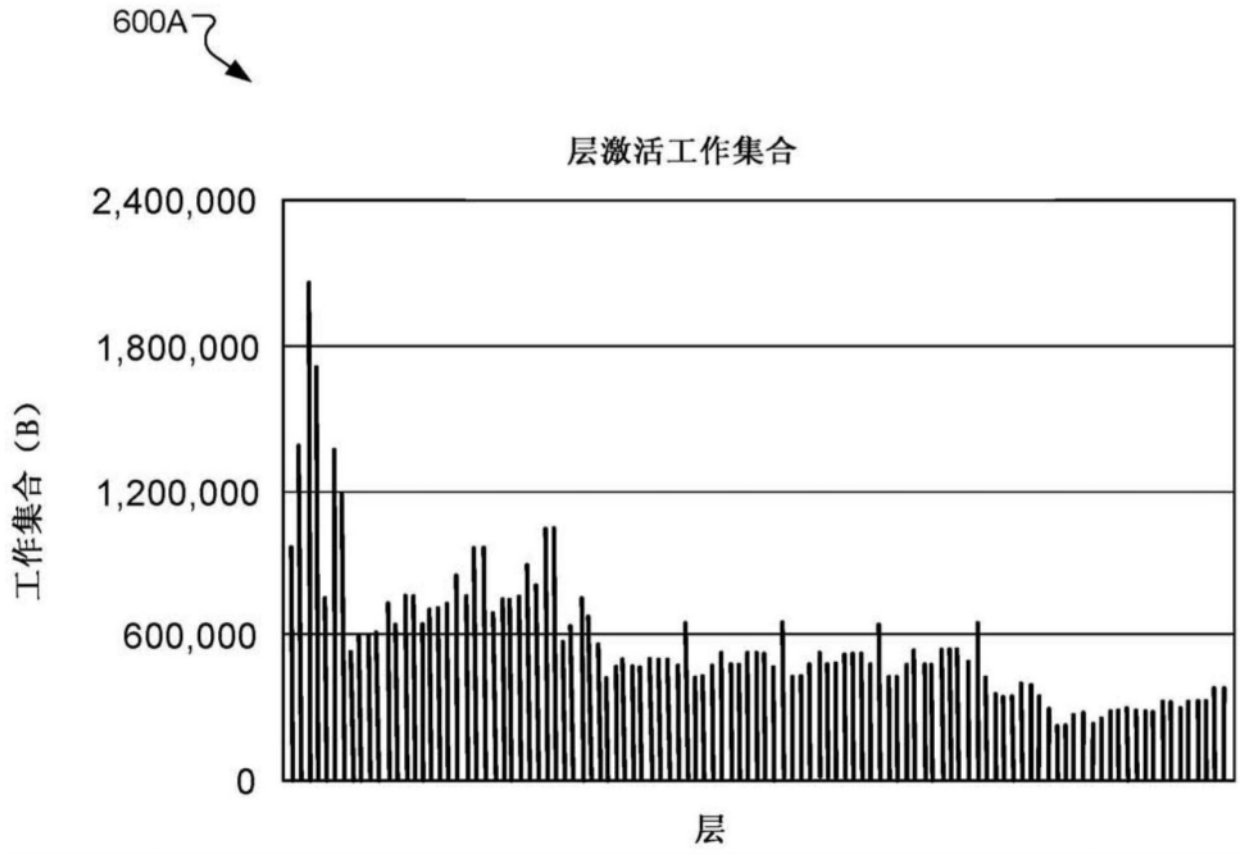


图6A

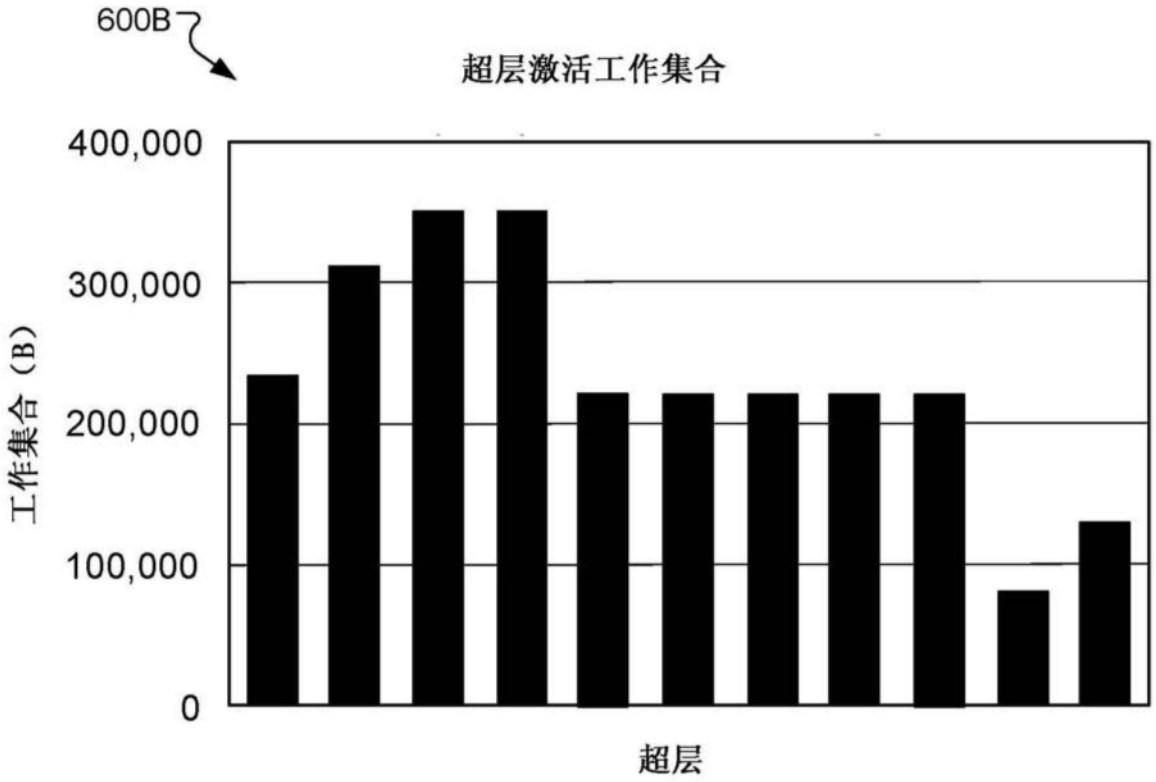


图6B