(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111919226 B
(45)授权公告日 2025.01.07

(21)申请号 201980022623.X
(22)申请日 2019.04.26
(65)同一申请的已公布的文献号
申请公布号 CN 111919226 A
(43)申请公布日 2020.11.10
(30)优先权数据
62/663,955 2018.04.27 US
(85)PCT国际申请进入国家阶段日
2020.09.27
(86)PCT国际申请的申请数据
PCT/US2019/029450 2019.04.26
(87)PCT国际申请的公布数据
WO2019/210237 EN 2019.10.31

(73)专利权人 平头哥(上海)半导体技术有限公司
地址 200120 上海市浦东新区自由贸易试验区上科路366号、川和路55弄2号5层
(72)发明人 韩亮
(74)专利代理机构 北京清源汇知识产权代理事务所(特殊普通合伙) 11644
专利代理师 冯德魁 张艳梅

(51)Int.Cl.
G06N 3/09(2023.01)

(56)对比文件
US 2018101748 A1,2018.04.12
CN 103914735 A,2014.07.09
US 2018114110 A1,2018.04.26

审查员 吴朝烨

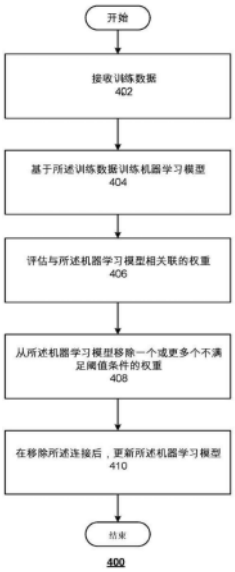
权利要求书3页 说明书11页 附图9页

(54)发明名称

执行机器学习的装置和方法

(57)摘要

本公开的实施例提供了用于执行机器学习的方法和系统。该方法可以包括:接收训练数据;以及基于训练数据训练机器学习模型,其中,机器学习模型包括多层,每一层具有一个或更多个节点,所述一个或更多个节点与来自机器学习模型的另一层的节点具有一个或更多个连接;评估与机器学习模型的连接相关联的权重,其中每个连接具有相应的权重,从机器学习模型中去除一个或更多个连接的权重不满足阈值条件的连接;并在去除连接后,更新机器学习模型。



CN 111919226 B

1.一种计算机实现的方法,包括:

接收训练数据,其中,所述训练数据包括视频数据;

基于所述训练数据训练机器学习模型,其中,所述机器学习模型包括多层,每一层具有一个或更多个节点,所述一个或更多个节点与来自所述机器学习模型的另一层的节点具有一个或更多个连接;

评估与所述机器学习模型的连接相关联的权重,其中每个所述连接具有对应的权重,所述连接包括默认权重和旁路权重,所述默认权重是根据训练确定的,连接权重是所述默认权重还是所述旁路权重基于指导信息来选择,其中所述指导信息是由过滤器基于输入数据和神经网络中的至少一个来生成;

从所述机器学习模型中去除一个或更多个权重不满足阈值条件的连接;以及

去除连接后,更新所述机器学习模型。

2.根据权利要求1所述的方法,其中,接收训练数据还包括:

通过去除一部分训练数据来减少所述训练数据。

3.根据权利要求2所述的方法,其中,所述训练数据包括多个维度,并且所述去除的部分与至少一个维度相关联。

4.根据权利要求1所述的方法,还包括:

评估与所述机器学习模型各层相关的权重,其中每一层都有对应的层权重;

从所述机器学习模型中去除具有权重不满足层阈值条件的一层或多层;和

在去除一层或多层之后,更新所述机器学习模型。

5.根据权利要求2所述的方法,其中,所述训练数据包括与第一时刻相关联的第一数据和与第二时刻相关联的第二数据,

并且通过去除所述训练数据的一部分来减少训练数据还包括:

从所述训练数据中去除与所述第一时刻相关联的所述第一数据。

6.根据权利要求1所述的方法,还包括:

生成用于评估要提供给所述机器学习模型的输入数据的过滤器。

7.一种用于执行机器学习的计算机实现的方法,包括:接收要提供给机器学习模型的输入数据,该机器学习模型包括多层,每一层具有一个或更多个节点,所述一个或更多个节点与来自所述机器学习模型的另一层的节点具有一个或更多个连接;

通过过滤器处理所述输入数据,所述过滤器引导所述输入数据在运行时绕过神经网络的至少一个连接,所述连接包括默认权重和旁路权重,所述默认权重是根据训练确定的,连接权重是所述默认权重还是所述旁路权重基于指导信息来选择,其中所述指导信息是由所述过滤器基于所述输入数据和所述神经网络中的至少一个来生成;和

将处理后的输入数据提供给所述机器学习模型;

其中,所述输入数据包括视频数据。

8.根据权利要求7所述的方法,其中,处理所述输入数据还包括:

去除一部分输入数据;和

更新所述输入数据。

9.根据权利要求8所述的方法,其中,所述输入数据包括多个维度,并且所述去除的部分与至少一个维度相关联。

10.根据权利要求9所述的方法,其中,所述输入数据包括与第一时刻相关联的第一数据和与第二时刻相关联的第二数据,

并且去除所述输入数据的所述部分还包括:

从输入数据中去除与所述第一时刻相关联的第一数据。

11.根据权利要求7所述的方法,其中,所述处理后的输入数据还包括指导信息,并且所述处理后的输入数据还被配置为基于所述指导信息来绕过至少一个连接。

12.根据权利要求11所述的方法,其中,所述至少一个连接包括与层相关联的所有连接。

13.根据权利要求7所述的方法,其中,所述过滤器是基于所述机器学习模型的训练而生成的。

14.一种非暂时性计算机可读介质,其存储有指令集,该指令集可由计算机系统的至少一个处理器执行,以使该计算机系统执行用于简化机器学习模型的方法,该方法包括:

接收训练数据,其中,所述训练数据包括视频数据;

基于所述训练数据训练机器学习模型,其中,所述机器学习模型包括多层,每一层具有一个或更多个节点,所述一个或更多个节点与来自所述机器学习模型的另一层的节点具有一个或更多个连接;

评估与所述机器学习模型的连接相关联的权重,其中每个连接具有对应的权重,所述连接包括默认权重和旁路权重,所述默认权重是根据训练确定的,连接权重是所述默认权重还是所述旁路权重基于指导信息来选择,其中所述指导信息是由过滤器基于输入数据和神经网络中的至少一个来生成;

从所述机器学习模型中去除一个或更多个权重不满足阈值条件的连接;以及

去除连接后,更新所述机器学习模型。

15.根据权利要求14所述的非暂时性计算机可读介质,其中,接收训练数据还包括:

通过去除一部分训练数据来减少所述训练数据。

16.根据权利要求14所述的非暂时性计算机可读介质,其中,所述训练数据包括多个维度,并且所述去除的部分与至少一个维度相关联。

17.根据权利要求14所述的非暂时性计算机可读介质,其中,所述指令集还进一步由所述计算机系统的至少一个处理器执行,以使所述计算机系统执行:

评估与所述机器学习模型各层相关的权重,其中每一层都有对应的层权重;

从所述机器学习模型中去除具有权重不满足层阈值条件的一层或多层;和

在去除一层或多层之后,更新所述机器学习模型。

18.根据权利要求15所述的非暂时性计算机可读介质,其中,所述训练数据包括与第一时刻相关联的第一数据和与第二时刻相关联的第二数据,并且通过去除所述训练数据的一部分来减少所述训练数据还包括:

从所述训练数据中去除与所述第一时刻相关的所述第一数据。

19.根据权利要求14所述的非暂时性计算机可读介质,其中,所述指令集还进一步由所述计算机系统的至少一个处理器执行,以使所述计算机系统执行:

生成用于评估要提供给所述机器学习模型的输入数据的过滤器。

20.一种非暂时性计算机可读介质,其存储指令集,该指令集可由计算机系统的至少一

个处理器执行,以使所述计算机系统执行用于执行机器学习的方法,所述方法包括:

接收要提供给机器学习模型的输入数据,该机器学习模型包括多层,每一层具有一个或更多个节点,该一个或更多个节点与来自所述机器学习模型的另一层的节点具有一个或更多个连接;

通过过滤器处理所述输入数据,所述过滤器引导所述输入数据在运行时绕过神经网络的至少一个连接,所述连接包括默认权重和旁路权重,所述默认权重是根据训练确定的,连接权重是所述默认权重还是所述旁路权重基于指导信息来选择,其中所述指导信息是由所述过滤器基于所述输入数据和所述神经网络中的至少一个来生成;和

将处理后的输入数据提供给所述机器学习模型;

其中,所述输入数据包括视频数据。

21.一种计算机系统,包括:

存储器,所述存储器存储有指令集;和

至少一个处理器,其被配置为执行所述指令集以使所述系统执行:

接收训练数据其中,所述训练数据包括视频数据:

基于所述训练数据训练机器学习模型,其中,所述机器学习模型包括多层,每一层具有一个或更多个节点,所述一个或更多个节点与来自所述机器学习模型的另一层的节点具有一个或更多个连接;

评估与所述机器学习模型的连接关联的权重,其中每个连接具有相应的权重,所述连接包括默认权重和旁路权重,所述默认权重是根据训练确定的,连接权重是所述默认权重还是所述旁路权重基于指导信息来选择,其中所述指导信息是由过滤器基于输入数据和神经网络中的至少一个来生成;

从所述机器学习模型中去除一个或更多个权重不满足阈值条件的连接;以及

去除连接后,更新所述机器学习模型。

22.一种用于执行机器学习的系统,包括:

存储器,所述存储器存储有指令集;和

至少一个处理器,其被配置为执行所述指令集以使所述系统执行:

接收要提供给机器学习模型的输入数据,该机器学习模型包括多层,每一层具有一个或更多个节点,该一个或更多个节点与来自所述机器学习模型的另一层的节点具有一个或更多个连接;

通过过滤器处理所述输入数据,所述过滤器引导所述输入数据在运行时绕过神经网络的至少一个连接,所述连接包括默认权重和旁路权重,所述默认权重是根据训练确定的,连接权重是所述默认权重还是所述旁路权重基于指导信息来选择,其中所述指导信息是由所述过滤器基于所述输入数据和所述神经网络中的至少一个来生成;和

将处理后的输入数据提供给所述机器学习模型;

其中,所述输入数据包括视频数据。

执行机器学习的装置和方法

[0001] 相关申请的交叉引用

[0002] 本公开要求于2018年4月27日提交的美国临时申请号62/663,955的优先权权益，该临时申请的全部内容通过引用合并于此。

背景技术

[0003] 随着机器学习程序的发展，机器学习模型的维度已经显着增加以提高模型准确性。但是，深度机器学习模型会在模型训练或推断过程中消耗大量存储空间、内存带宽、能耗和计算资源。这些问题使得难以在移动设备和嵌入式设备上部署深度机器学习模型。

[0004] 本公开的实施例通过提供用于执行机器学习的方法和系统来解决上述问题。

发明内容

[0005] 本公开的实施例提供一种计算机实现方法。该方法可以包括：接收训练数据；基于所述训练数据来训练机器学习模型，其中，所述机器学习模型包括多层，每一层具有一个或更多个节点，该一个或更多个节点与来自所述机器学习模型的另一层的节点具有一个或更多个连接；评估与所述机器学习模型的连接相关联的权重，其中每个所述连接具有对应的权重；从机器学习模型中去除一个或更多个权重不满足阈值条件的连接，并在去除连接后，更新机器学习模型。

[0006] 本公开的实施例还提供了一种用于执行机器学习的计算机实现的方法。该方法可以包括：接收要提供给机器学习模型的输入数据，该机器学习模型包括多层，每一层具有一个或更多个节点，该一个或更多个节点与来自所述机器学习模型的另一层的节点具有一个或更多个连接；通过过滤器处理所述输入数据：并将处理后的输入数据提供给机器学习模型。

[0007] 本公开的实施例还提供一种非暂时性计算机可读介质，其存储有指令集，所述指令集可由计算机系统的至少一个处理器执行，以使计算机系统执行用于简化机器学习模型的方法。该方法可以包括：接收训练数据；以及基于训练数据训练机器学习模型，其中，机器学习模型包括多层，每个层具有一个或更多个节点，所述一个或更多个节点与来自机器学习模型的另一层的节点具有一个或更多个连接；评估与所述机器学习模型的连接相关联的权重，其中每个连接具有对应的权重；从机器学习模型中去除一个或更多个权重不满足阈值条件的连接；并在去除连接后，更新所述机器学习模型。

[0008] 本公开的实施例还提供一种非暂时性计算机可读介质，其存储有指令集，所述指令集可由计算机系统的至少一个处理器执行，以使计算机系统执行用于执行机器学习的方法，该方法可以包括：接收要提供给机器学习模型的输入数据，该机器学习模型包括多层，每一层具有一个或更多个节点，该一个或更多个节点与来自所述机器学习模型的另一层的节点具有一个或更多个连接；通过过滤器处理所述输入数据；并将处理后的输入数据提供给所述机器学习模型。

[0009] 本公开的实施例还提供一种计算机系统。该计算机系统可以包括：存储器，其存储

有指令集;以及至少一个处理器,其被配置为执行所述指令集以使所述系统执行:接收训练数据;基于所述训练数据训练机器学习模型,其中,所述机器学习模型包括多层,每一层具有一个或更多个节点,所述一个或更多个节点与来自所述机器学习模型的另一层的节点具有一个或更多个连接;评估与所述机器学习模型的连接相关联的权重,其中每个连接具有对应的权重;从所述机器学习模型中去除一个或更多个权重不满足阈值条件的连接;并在去除连接后,更新机器学习模型。

[0010] 本公开的实施例还提供一种用于执行机器学习的计算机系统。所述计算机系统可以包括:存储有指令集的存储器;至少一个处理器,其被配置为执行所述指令集以使所述系统执行:接收要提供给机器学习模型的输入数据,该机器学习模型包括多层,每一层具有一个或更多个节点,该一个或更多个节点与来自所述机器学习模型的另一层的节点具有一个或更多个连接;通过过滤器处理所述输入数据;和将处理后的输入数据提供给所述机器学习模型。

[0011] 所公开的实施例的附加特征和优点将在下面的描述中部分地阐述,并且部分地将从描述中变得显而易见,或者可以通过实施例的实施而获知。公开的实施例的特征和优点可以通过权利要求中阐述的要素和组合来实现和获得。

[0012] 应当理解,前面的一般描述和下面的详细描述都仅仅是示例性和说明性的,并且不限制所要求保护的实施例。

附图说明

[0013] 在以下详细描述和附图中示出了本公开的实施例和各个方面。图中所示的各种特征未按比例绘制。

[0014] 图1示出了根据本公开实施例的神经网络处理架构。

[0015] 图2A-F示出了根据本公开的实施例的机器学习模型的示意图。

[0016] 图3示出了根据本公开的实施例的简化输入数据的示意图。

[0017] 图4示出了根据本公开实施例的计算机实现的方法。

[0018] 图5示出根据本公开的实施例的用于执行机器学习的计算机实现的方法。

具体实施方式

[0019] 下面更详细地描述本公开的特定方面。如果与通过引用并入的术语或定义冲突,则本文提供的术语和定义为准。

[0020] 本公开中描述的技术具有以下一种或多种技术效果。在一些实施方式中,本公开中描述的技术提供了一种计算机实现的方法,包括:接收训练数据;基于所述训练数据训练机器学习模型,其中,机器学习模型包括多层,每个层具有一个或更多个节点,所述一个或更多个节点与来自机器学习模型的另一层的节点具有一个或更多个连接;评估与机器学习模型的连接相关联的权重,其中每个连接具有对应的权重;从机器学习模型中删除一个或更多个连接权重不满足阈值条件的连接;并在删除连接后,更新机器学习模型。在一些实施方式中,在本公开中描述的技术通过去除机器学习模型的一个或更多个连接来简化机器学习模型。这还可以减少机器学习模型所需的存储空间、计算资源和功耗。在一些实施方式中,本公开中描述的技术还生成用于评估要提供给机器学习模型的输入数据的过滤器。该

过滤器可以在运行时处理机器学习模型的输入数据。过滤器可以通过去除一部分输入数据来减少输入数据。这允许减少机器学习模型所需的存储空间、带宽、计算资源和功耗。在一些实施方式中,过滤器还可以将指导信息合并到输入数据中,使得输入数据可以基于指导信息绕过机器学习模型的至少一个连接,而不会恶化推理结果。这允许减少机器学习模型所需的计算资源和功耗。

[0021] 如本文所用,术语“包括”、“包含”或其任何其他变体旨在覆盖非排他性包括,使得包括以下各项的处理、方法、组合物、物品或装置、元件不仅仅包括那些元件,而是可以包括未明确列出或此类过程、方法、组合物、物品或设备所固有的其他元件。术语“示例性”以“示例”而不是“理想”的意义使用。

[0022] 图1示出了根据本公开的实施例的示例性神经网络处理架构100。如图1所示。架构100可以包括芯片通信系统102、主机存储器104、存储器控制器106、直接存储器访问(DMA)单元108、联合测试操作组(JTAG)/测试访问结束(TAP)控制器110、外围接口112、总线114、全局存储器116等。应当理解,芯片通信系统102可以基于所传递的数据来执行算法操作(例如,机器学习操作)。

[0023] 芯片通信系统102可以包括全局管理器1022和多个核心1024。全局管理器1022可以包括至少一个任务管理器以与一个或更多个核心1024进行协调。每个任务管理器可以与一核心1024阵列相关联,该核心阵列为神经网络提供突触/神经元电路。例如,图1的处理元件的顶层可以提供表示神经网络的输入层的电路,而核心的第二层可以提供表示神经网络的隐藏层的电路。因此,芯片通信系统102也可以被称为神经网络处理单元(NPU)。如图1所示,全局管理器1022可以包括两个任务管理器以与两个核心阵列协调。

[0024] 核心1024可以包括一个或更多个处理元件,每个处理元件包括单个指令、多数据(S1MD)架构,所述架构包括一个或更多个被配置为基于从全局管理器1022接收到的指令来执行一个或更多个操作(例如,乘法、加法、乘法累加等)的多个处理单元。为了对所传送的数据分组执行操作,核心1024可以包括一个或更多个处理元件,用于处理数据分组中的信息。每个处理元件可以包括任何数量的处理单元。在一些实施例中,核心1024可以被认为是瓦片(tile)等。

[0025] 主机存储器104可以是片外存储器,例如主机CPU的存储器。例如,主机存储器104可以是DDR存储器(例如,DDR SDRAM)等。与充当高级缓存的、集成在一个或更多个处理器中的片上存储器相比,主机存储器104可以配置为以较低的访问速度存储大量数。

[0026] 存储器控制器106可以管理对作为主存储器的具有片上存储块(例如,8GB第二代高带宽存储器(HBM2)的4个块)的全局存储器116内的特定存储器块的数据的读取和写入。例如,存储器控制器106可以管理来自芯片通信系统102外部(例如,来自DMA单元108或与另一NPU相对应的DMA单元)或芯片通信系统102内部(例如,来自通过由全局管理器1022的任务管理器控制的2D网格控制核心1024中的本地存储器)的数据的读取和写入。此外,尽管在图1中示出了一个存储器控制器,但是应当理解,架构100中可以提供一个以上的存储控制器。例如,全局存储器内116的每个存储器块(例如,HBM2)可以有一个存储器控制器。

[0027] 存储器控制器106可以生成存储器的地址并发起存储器读取或写入周期。存储器控制器106可包含可由一个或更多个处理器写入和读取的几个硬件寄存器。这些寄存器可以包括存储器地址寄存器、字节计数寄存器、一个或更多个控制寄存器以及其他类型的寄

存器。这些寄存器可以指定源、目的地、传输方向(从输入/输出(I/O)设备读取或写入I/O设备)、传输单元的大小、一个突发中要传输的字节数、或存储控制器的其他典型功能中的一些组合。

[0028] DMA单元108可以辅助在主机存储器104和全局存储器116之间传送数据。另外,DMA单元108可以辅助在多个NPU(例如,NPU100)之间传送数据。DMA单元108可以允许片外设备访问片内和片外存储器,而不会引起CPU中断。因此,DMA单元108还可以生成存储器地址并发起存储器读取或写入周期。DMA单元108还可包含可由一个或更多个处理器写入和读取的几个硬件寄存器,包括存储器地址寄存器、字节计数寄存器、一个或更多个控制寄存器以及其他类型的寄存器。这些寄存器可以指定源、目的地、传输方向(从输入/输出(I/O)设备读取或写入I/O设备)/传输单元的大小或一个突发中传输数量的某种组合。应当理解,架构100可以包括第二DMA单元,其可以用于在其他神经网络处理架构之间传输数据,以允许多个神经网络处理架构直接通信而不涉及主机CPU。

[0029] JTAG/TAP控制器110可以指定专用的调试端口,该端口实现串行通信接口(例如JTAG接口),以实现对NPU的低开销访问,而无需直接从外部访问系统地址和数据总线。JTAG/TA控制器110还可以具有片上测试访问接口(例如,TAP接口),其实现访问一组测试寄存器的协议,该测试寄存器呈现芯片逻辑电平和各个部分的设备能力。

[0030] 外围接口112(例如PCIe接口)(如果存在)充当(通常是)在架构100和其他设备之间提供通信的芯片间总线。

[0031] 总线114包括芯片内总线和芯片间总线。芯片内总线将所有内部组件相互连接,这是系统架构所要求的。尽管并非所有组件都与每个其他组件都连接,但所有组件确实与它们需要与之通信的其他组件建立了某种连接。芯片间总线将NPU与其他设备相连,例如片外存储器或外围设备。通常,如果存在外围接口112(例如,芯片间总线),则总线114仅与芯片内总线有关,尽管在某些实现中,总线114仍可能与专门的总线间通信有关。

[0032] 芯片通信系统102可以被配置为基于神经网络执行操作。

[0033] 架构100还可以包括主机单元120。主机单元120可以是一个或更多个处理单元(例如X86中央处理单元)。在一些实施例中,具有主机单元120和主机存储器104的主机系统可以包括编译器(未示出)。编译器是一种程序或计算机软件,可将以一种编程语言编写的计算机代码转换为NPU指令以创建可执行程序。在机器应用程序中,编译器可以执行各种操作,例如,预处理、词法分析、解析、语义分析、输入程序到中间表示的转换、代码优化、和代码生成或前述操作的组合。除了编译程序之外,主机系统还可以分析要由芯片通信系统102处理的输入数据,从而可以提取输入数据的特征。

[0034] 在一些实施例中,生成NPU指令的编译器可以在主机系统上,该主机系统将命令推送到芯片通信系统102。基于这些命令,每个任务管理器可以将任意数量的任务分配给一个或更多个核心(例如核心1024)。某些命令可以指示DMA单元108将指令(由编译器生成)和数据(例如输入数据)从主机存储器104加载到全局存储器116中。然后,所加载的指令可以分配给分配了该指令的相应于的任务的每个核心,并且一个或更多个核心可以处理这些指令。

[0035] 图2A-F示出了根据本公开的实施例的机器学习模型的示意图。作为机器学习模型的示例,神经网络可以包括多个层,并且每个层可以包括多个节点(也称为人工神经元)。可

以通过用训练数据训练神经网络来建立多个节点之间的连接。在图2A-2F中,粗体箭头可以表示层之间的许多连接。神经网络的训练也可以称为机器学习,而经过训练的神经网络也可以称为机器学习模型。可以为每个连接分配包括多个位(例如32位)的权重。连接的权重可以增加或减小连接处的信号强度。例如,如图2A所示,两个节点221和231之间的连接2011的权重可以为零。因此,通过连接2011的信号减小到零。换句话说,信号无法通过连接2011,并且两个节点221和231是不连接的。

[0036] 图2A示出了示例性神经网络201的示意图。神经网络201可以包括四个层(例如,层210、220、230和240),每个层包括多个节点(例如,节点211、212、213,等等)。在本公开的一些实施例中,神经网络201可以是用于训练的初始神经网络。初始神经网络是具有默认参数(例如,连接权重)的神经网络。神经网络的参数可以与节点之间的连接以及连接的权重相关。在一些实施例中,神经网络201可以是通过基于训练数据来训练初始神经网络而得到的训练后的神经网络,从而可以在层的多个节点之间建立多个连接。

[0037] 根据本公开的实施例,可以简化神经网络201中的多个连接。在一些实施例中,可以在训练期间在神经网络201上执行简化。训练期间的简化也可以称为静态时间的简化。

[0038] 在一些实施例中,简化可以去除训练期间两个节点之间的连接。图2B示出了根据本公开的实施例的在减枝之后的神经网络202的示意图,其涉及一些连接(也称为突触)的去除。例如,在图2B中,从神经网络201中去除了图2A的连接2011。如上所述,训练后的神经网络的每个连接都被分配了连接权重。然后,可以确定连接权重是否满足阈值。例如,当连接权重大于或等于阈值时,连接权重满足阈值并维持对应的连接。否则,当连接权重小于阈值时,连接权重不满足阈值。如果连接的连接权重不满足阈值,则可以删除相应的连接。在一些实施例中,可以通过将连接的连接权重设置为零来删除连接。在一些实施例中,还可以通过从神经网络删除连接来去除连接。

[0039] 在一些实施例中,在训练后的神经网络中,可以将要移除的至少一个连接的连接权重设置为零,并且可以更新训练后的神经网络以进行准确性评估。如果更新后的神经网络的精度令人满意,则至少一个连接可以最终从神经网络中删除。另一方面,如果更新后的神经网络的精度不令人满意,则可以调整删除连接的阈值。可以理解,可以根据不同的神经网络和应用神经网络的不同因素(例如,准确性,能量消耗等)来调整阈值。

[0040] 基于其余的连接,可以将神经网络201更新为最终的神经网络(例如,如图2B所示的神经网络202)。应当理解,其余连接的权重也可以被更新。由于原始权重可以是多个位数(例如32位),因此每个更新的权重仍可以包括该位数。

[0041] 图2C示出了根据本公开的实施例的在去除层之后的神经网络203的示意图。在一些实施例中,可以在训练过程中删除神经网络的层,以进一步简化神经网络。例如,在图2C中,2A或2B中的层230已从神经网络中删除。

[0042] 在一些实施例中,可以基于神经网络的连接来确定神经网络的每一层的层权重。层的层权重可以与该层的节点的连接权重有关。例如,层权重可以是该层的节点的连接权重的总和。然后,可以确定该层的层权重是否满足阈值。例如,当层权重大于或等于阈值时,层权重满足阈值。否则,当层权小于阈值时,层权重不满足阈值。如果连接的层权重不满足阈值,则可以去除该层(例如,层230)。在一些实施例中,可以通过将与该层有关的所有连接的连接权重设置为零来“去除”该层。在一些实施例中,还可以通过在神经网络中删除该层

来去除该层。

[0043] 基于其余层,可以将神经网络201更新为最终的神经网络(例如,如图2C所示的神经网络203)。应当理解,可以更新其余层中的连接和连接权重。

[0044] 除了基于训练数据修改神经网络之外,可以在将训练数据用于训练神经网络之前对其进行修改。在一些在实施例中,可以去除一部分训练数据。例如,训练数据可以包括多个维度(例如10个维度)。在输入用于训练神经网络的训练数据之前,可以将训练数据的至少一维去除。

[0045] 在一些实施例中,时间信息可以与诸如递归神经网络(RNN)和长短期记忆网络(LSTM)的神经网络有关。训练数据的去除部分可以与时域相关。可以理解,这些神经网络可以处理数据序列。因此,神经网络的刺激不仅可以来自时间T的新输入数据,而且可以来自时间T-1的历史信息。因此,在将训练数据输入到神经网络之前,可以去除与时刻T1相关联的第一训练数据,而可以将与时刻T2相关联的第二训练数据提供给神经网络以进行训练。

[0046] 应当理解,节点和层之间的连接与训练数据有关。如上所述,由于连接和层的权重小于给定的阈值,因此可以去除一些连接或层。因此,当在运行时使用神经网络时,输入数据可以绕过(或跳过)连接或层。而且,一部分训练数据可能对神经网络的结果影响很小,因此可以被神经网络忽略。因此,可以基于神经网络的训练来生成过滤器。在一些实施例中,可以通过学习算法来生成过滤器。例如,可以基于移除的连接,移除的层,剩余的连接和层以及神经网络的输出中的至少一项来训练过滤器。在一些实施例中,过滤器可以是层之间的门和/或门控神经网络。例如,过滤器可以包括有限数量的卷积层、平均池化层和完全连接层以输出维度矢量。过滤器可以只依赖于上一层的输出,并应用少量的卷积和池化操作。

[0047] 在一些实施例中,过滤器可以被手动编程以合并过滤规则。在一些实施例中,过滤器还可以包含例如由软件工程师确定的规则。换句话说,过滤器可以由软件工程师生成或设计。

[0048] 在机器应用中,过滤器可以部署在图1的主机系统上。因此,在输入数据被传输到芯片通信系统102之前,过滤器可以引导输入数据在运行时绕过神经网络的至少一个连接。例如,过滤器可以将指导信息合并到输入数据中,并且指导信息可以包括用于通过机器学习模型的至少一个连接的输入数据的路线。应当理解,当绕过层的所有连接时,该层被绕过。例如,过滤器可以将指导信息合并到输入数据中,以便可以根据指导信息绕过某些连接或层。过滤器还可以通过删除一部分输入数据来减少输入数据。例如,当输入数据通过过滤器时,可以从输入数据中去除该部分(例如,输入数据的至少一维)。因此,除了减轻神经网络的处理负担之外,还可以降低主机单元120与芯片通信系统102之间的数据通信量。因此,可以减少架构100的功耗和带宽使用。

[0049] 图2D-2F示出了根据本公开的实施例的在运行时的神经网络204-206的示例。在一些实施例中,简化还可以在推断(reference)期间在神经网络201上执行。推断期间的简化也可以称为运行时的简化。

[0050] 如上所述,输入数据可以包含由过滤器生成的指导信息。在本公开的一些实施例中,指导信息可以输入数据绕过至少一层。可以由过滤器基于输入数据和神经网络中的至少一个来生成所述指导信息。因此,取决于输入数据和神经网络,过滤器生成的指导信息可能会有所不同。

[0051] 图2D示出了运行时神经网络204的示意图。如图2D所示,一些输入数据可以绕过层230。在一些实施例中,神经网络204可以在运行时没有层230的情况下为一些输入数据生成准确的结果。例如,更复杂的输入数据可以被路由通过更多的层,而不太复杂输入数据可以被路由通过较少的层。因此,基于输入数据,层230可以在运行时被绕过。因此,过滤器可以生成指导输入数据绕过层230的指导信息。

[0052] 应当理解,在图2D中,与层230相关联的连接仍然提供在神经网络204中,并且被示出为虚线箭头。因此,一些其他输入数据可以经由这些连接通过层230。

[0053] 与层230相关联的连接权重可以被设置为零或任何其他值。因此,在本公开的实施例中,通过绕过至少一层(例如,将层的连接权重设置为零),可以减少处理输入数据所需的计算资源。

[0054] 除了绕过至少一层之外,本公开的实施例还可以在运行时绕过节点之间的至少一个连接。图2E示出了运行时神经网络205的示意图。如图2E所示,在层210和220之间,神经网络205包括节点213和222之间的第一连接,节点212和222之间的第二连接,以及节点211和223之间的第三连接。

[0055] 在一些实施例中,神经网络205可以在没有至少一个连接(例如,如图2E所示的第一和第二连接)的情况下,在运行时为第一输入数据生成准确的结果。因此,例如,可以在运行时绕过第一和第二连接,而不会降低神经网络205的结果。因此,过滤器可以生成指示输入数据绕过第一和第二连接的指导信息。

[0056] 这样,可以在不修改神经网络的情况下(例如,从神经网络中永久移除至少一个连接)来减轻运行神经网络的计算负担。除了减少计算负担之外,这还允许运行神经网络的更大灵活性。如上所述,过滤器可以由软件工程师或基于训练来生成。因此,可以在不同的应用场景中分别将不同的过滤器部署到机器学习模型。

[0057] 可以理解的是,在运行时绕过的至少一个连接可以根据不同的输入数据而不同。例如,如在图2F中所示,基于指导信息,神经网络205的第二和第三连接可以被不同于第一输入数据的第二输入数据绕过。如上所述,指导信息可以包括用于经过机器学习模型的至少一个连接的输入数据的路线。在一些实施例中,用于不同输入数据的路由可以不同,如图2E和图2F中所示。

[0058] 因此,在本公开的实施例中,通过绕过至少一个连接,还可以减少处理输入数据所需的计算资源。

[0059] 除了在静态时间简化训练数据外,还可以在运行时对输入数据进行简化。图3示出了根据本公开的实施例的简化输入数据示意图。

[0060] 可以通过在输入数据被发送到神经网络之前去除一部分输入数据来执行输入数据的简化。

[0061] 在一些实施例中,运行时的输入数据可以涉及多个维度。至少一个维度可以从输入数据中去除。如图3所示,输入数据301可以包括多个维度(例如,至少一个维度3011)。在将输入数据301发送到神经网络300进行处理之前,可以从输入数据301中去除至少一个维度3011(例如,通过过滤器)。在一些实施例中,过滤器可以基于输入数据301和神经网络300来确定要去除的至少一个维度3011。

[0062] 在一些实施例中,输入数据的被去除的部分可以与时域相关。例如,当输入数据与

时间(例如,视频)有关时,输入数据可以包括与时刻有关的数据序列。例如,视频的数据可以包括在一段时间内分布的一系列帧,并且每帧的数据对应于一个时刻。在一些实施例中,输入数据的一部分(例如,视频的一帧或给定时间段内的多个帧)可以被神经网络绕过,对最终结果的影响非常有限。例如,在视频的数据中,过滤器可以确定从输入数据中删除前五秒内的帧,而不会显着影响最终结果。

[0063] 应当理解,取决于输入数据的性质,要从输入数据中去除的部分可以是不同的。

[0064] 通过去除输入数据的一部分,可以减少机器学习模型所占用的带宽,并且还可以减少机器学习模型所需的计算资源和功耗。

[0065] 图4示出了根据本公开的实施例的计算机实现的方法400。方法400可以由计算机系统来实现,例如图1的神经网络处理架构100。计算机系统可以包括:存储指令集的存储器和至少一个被配置为执行所述指令集以使计算机系统执行方法400的处理器。所述至少一个处理器可以包括例如图1的主机单元120和芯片单元的芯片通信系统102。该计算机系统还可以包括通信接口(例如,图1的外围接口112)。请参照图4,方法400可以包括以下步骤。

[0066] 在步骤402中,计算机系统可以(例如经由通信接口)接收训练数据。所述训练数据可以存储在计算机系统的数据库中或存储在另一计算机系统中。可以通过删除一部分训练数据来减少训练数据。

[0067] 在一些实施例中,当训练是有监督训练时,训练数据可包括大量标记数据。在一些实施例中,训练数据可包括输入向量和对应的输出向量对。因此,训练数据可以包括多个维度。每个维度都可以与训练数据的特征相关。在一些实施例中,移除的部分可以与至少一个维度相关联。

[0068] 在一些实施例中,训练数据与时域有关。例如,训练数据可以包括与第一时刻相关的第一数据和与第二时刻相关的第二数据。并且在去除训练数据的一部分时,可以去除与第一时刻相关联的第一数据。

[0069] 在步骤404中,计算机系统可以基于所述训练数据来训练机器学习模型。在训练机器学习模型之前,尚未确定机器学习模型的参数。所述参数可以例如包括机器学习模型的连通性。作为机器学习模型的示例,神经网络可以包括多个层,并且每个层可以包括多个节点。并且可以连接多个节点以在节点之间生成连接。每个连接可以具有相应的连接权重,因此对连接进行加权。通过训练,可以基于训练数据确定机器学习模型的连通性。例如,可以在节点之间建立连接,可以确定权重并将其分配给连接。

[0070] 类似地,计算机系统还可以确定与机器学习模型的各层相关联的各层权重。在机器学习模型中,每层可以具有相应的层权重。如上所述,可以基于连接权重来确定机器学习模型的每一层的层权重。

[0071] 在步骤406中,计算机系统可以评估与机器学习模型的连接关联的权重。例如,计算机系统可以确定连接的连接权重是否满足阈值条件。在一些实施例中,可以从机器学习模型移除连接,而不会显着影响机器学习模型的最终结果。可以将这种连接确定为可移除连接。相应地,如果连接的移除会导致机器学习模型的不准确结果,则可以将连接确定为不可移除的连接。阈值条件可以与用于确定可移除连接和不可移除连接的连接权重阈值相关联。例如,当连接的连接权重大于或等于连接权重阈值时,该连接不满足阈值条件,并且被评估为不可移除连接。同样例如,当连接的连接权重小于连接权重阈值时,该连接满足阈值

条件并且被评估为可移除连接。

[0072] 类似地,在一些实施例中,可以从机器学习模型中去除整个层,而不会显着影响机器学习模型的最终结果。可以将这种层确定为可移除层。相应地,如果该层的去除可以导致机器学习模型的不准确结果,则该层可以被确定为不可移除的层。因此,层阈值条件可以与用于标识可移除层和不可移除层的层权重阈值相关联,并且计算机系统可以进一步确定层的层权重是否满足该层阈值条件。例如,当层的层权重大于或等于层权重阈值时,该层不满足层阈值条件,并且被确定为不可移除层。又例如,当层的层重量小于层权重阈值时,该连接满足层阈值条件并且被确定为可移除层。

[0073] 在步骤408中,计算机系统可以从机器学习模型中删除权重不满足阈值条件的一个或更多个连接。换句话说,可以从机器学习模型中删除确定的可移除连接。

[0074] 类似地,可移除层也可以从机器学习模型中移除。

[0075] 在步骤410中,在连接被移除之后,计算机系统可以更新机器学习模型。可以理解的是,在最终将可移动连接从机器学习模型中移除之后,计算机系统可以在节点之间建立新的连接,其中原始连接已被移除。

[0076] 通过从机器学习模型移除至少一个连接或层,可以降低机器学习模型的复杂性。因此,还可以减少用于机器学习模型的存储空间和用于运行机器学习模型的功耗。

[0077] 另外,计算机系统还可以生成用于评估要提供给机器学习模型的输入数据的过滤器。该过滤器可以用于执行机器学习,下面将参考图5进一步描述。

[0078] 图5示出了根据本公开的实施例的用于执行机器学习的计算机实现的方法500。方法500可以由计算机系统来实现,例如图1的神经网络处理架构100。该计算机系统可以包括:存储有一组指令的存储;以及被配置为执行该组指令以使计算机系统执行方法500的至少一个处理器。所述至少一个处理器可以包括例如主机单元120以及图1的芯片通信系统102。该计算机系统还可以包括通信接口(例如,图1的外围接口112)。参照图5,方法500可以包括以下步骤。

[0079] 在步骤502中,计算机系统可以接收要提供给机器学习模型的输入数据。机器学习模型可以包括多层,每个层具有一个或更多个节点,该一个或更多个节点与来自机器学习模型的另一层的节点具有一个或更多个连接。机器学习模型可以在机器学习应用程序中使用。机器学习应用程序可以在神经网络处理架构100中执行。机器学习应用程序可以划分为多个任务,并且其中一个任务可以在主机单元120上执行,而另一个可以确定为机器学习任务并由在芯片通信系统的机器学期模型执行。输入数据可以与机器学习任务相关,并且可以提供给机器学习模型进行处理。

[0080] 在步骤504中,计算机系统可以通过过滤器处理输入数据。在接收到输入数据之后,可以调用过滤器来处理输入数据。过滤器可以是机器学习应用程序的一部分,可以由计算机系统执行,也可以是计算机系统提供的功能。如参考图1-3所讨论的,可以基于机器学习模型的训练来生成过滤器,或者可以由软件工程师设计过滤器。

[0081] 基于所述输入数据,可以提取输入数据的特征并将其与机器学习模型进行比较。在一些实施例中,用于输入数据的指导信息可以由过滤器产生。可以将指导信息合并到输入数据中,以便可以将处理后的输入数据配置为基于所述指导信息绕过至少一个连接。例如,要绕过连接,可以将给定输入数据的连接权重设置为零。可以理解,对于另一个输入数

据,连接的权重可以不改变。在一些实施例中,连接可以包括默认权重和旁路权重。默认权重是根据训练确定的,旁路权重为零。基于所述指导信息,计算机系统可以确定连接权重是默认权重还是旁路权重。可以理解,旁路重量可以是另一个值。

[0082] 在一些实施例中,指导信息可以指示输入数据绕过与层相关联的连接,使得该层可以被输入数据绕过。应当理解,当某一层被绕过时,至少一个连接可以包括与该层相关的所有连接。

[0083] 在一些实施例中,通过使用过滤器,计算机系统可以去除一部分输入数据,并相应地更新输入数据。例如,基于输入数据的特征,计算机系统可以确定可以删除输入数据的一部分而不会恶化运行机器学习模型的结果,并且可以删除输入数据的一部分。可以更新输入数据并将其提供给机器学习模型。在一些实施例中,输入数据可以包括多个维度,并且移除的部分与至少一个维度相关联。在一些实施例中,输入数据可以与时域相关。例如,输入数据可以包括与第一时刻相关的第一数据和与第二时刻相关的第二数据。并且在去除输入数据的一部分时,可以去除与第一时刻相关联的第一数据。

[0084] 在步骤506中,计算机系统可以将处理后的输入数据提供给机器学习模型。机器学习模型可以生成机器学习的结果。例如,作为推断引擎,芯片通信系统102(涉及一个或更多个加速器)可以使用机器学习模型来生成结果,并将结果发送回主机单元120。

[0085] 应当理解,该过滤器可以应用于未简化的机器学习模型。尽管未简化的机器学习模型包括完整连接,但是在运行时可以通过使用过滤器使得输入数据绕过一个或更多个连接。由于在运行时输入数据会绕过一个或更多个连接,因此尽管未简化机器学习模型,但可以减少用于运行机器学习模型的计算资源和功耗。移除一部分输入数据,除了可以减少计算资源和功耗之外,还可以进一步减小主机单元120和芯片通信系统102之间的通信负载。

[0086] 本公开的实施例还提供一种计算机程序产品。该计算机程序产品可以包括非暂时性计算机可读存储介质,在其上存储有计算机可读程序指令,该指令可以用于使处理器执行上述的方法。

[0087] 所述的计算机可读存储介质可以是可以存储指令以供指令执行设备使用的有形设备。计算机可读存储介质可以是例如但不限于电子存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或前述的任何合适的组合。计算机可读存储介质的更具体示例的非穷举列表包括以下内容:便携式计算机软盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦可编程只读存储器(EPROM)、静态随机存取存储器(SRAM)、便携式光盘只读存储器(CD-ROM)、数字多功能磁盘(DVD)、存储棒、软盘、机械编码的设备(例如打孔卡或上面记录了指令的凹槽中凸起的结构)以及任何合适的组合以上所述。

[0088] 用于执行上述方法的计算机可读程序指令可以是汇编程序指令,指令集架构(ISA)指令,机器指令,机器相关指令,微代码,固件指令,状态设置数据或用一种或多种编程语言(包括面向对象的编程语言和常规过程编程语言)的任何组合编写的源代码或目标代码。所述计算机可读程序指令可以完全在计算机系统上作为独立软件包执行,或者部分地在第一计算机上并且部分地在远离所述第一计算机的第二计算机上执行。在后一种情况下,第二远程计算机可以通过任何类型的网络(包括局域网(LAN)或广域网(WAN))连接到第一台计算机。

[0089] 可以将计算机可读程序指令提供给通用或专用计算机的处理器,或者其他可编程

数据处理装置，以形成机器，使得该指令经由计算机的处理器或其他可编程数据处理设备的处理器执行时，创建用于实现上述方法的。

[0090] 附图中的流程图和示意图示出了根据说明书的各种实施例的设备、方法和计算机程序产品的可能实现的架构、功能和操作。在这方面，流程图或示意图中的框可以表示软件程序、代码段或代码部分，其包括用于实现特定功能的一个或更多个可执行指令。还应注意，在一些替代实施方式中，方框中指出的功能可以不按图中指出的顺序发生。例如，取决于所涉及的功能，连续示出的两个框实际上可以基本上同时执行，或者有时可以以相反的顺序执行这些框。还应当注意，示意图或流程图的每个方框，以及示意图和流程图中的方框的组合，可以由执行指定功能或动作的基于硬件的专用系统或专用硬件的组合和计算机指令来实现。

[0091] 应当理解，为清楚起见在单独的实施方式的上下文中描述的说明书的某些特征也可以在单个实施方式中组合提供。相反，为简洁起见，在单个实施例的上下文中描述的说明书的各种特征，也可以单独地或以任何合适的子组合或如在本发明的任何其他描述的实施例中那样适当地提供。在各种实施例的上下文中描述的某些特征不应被认为是那些实施例的必要特征，除非该实施例在没有那些要素的情况下是不可操作的。

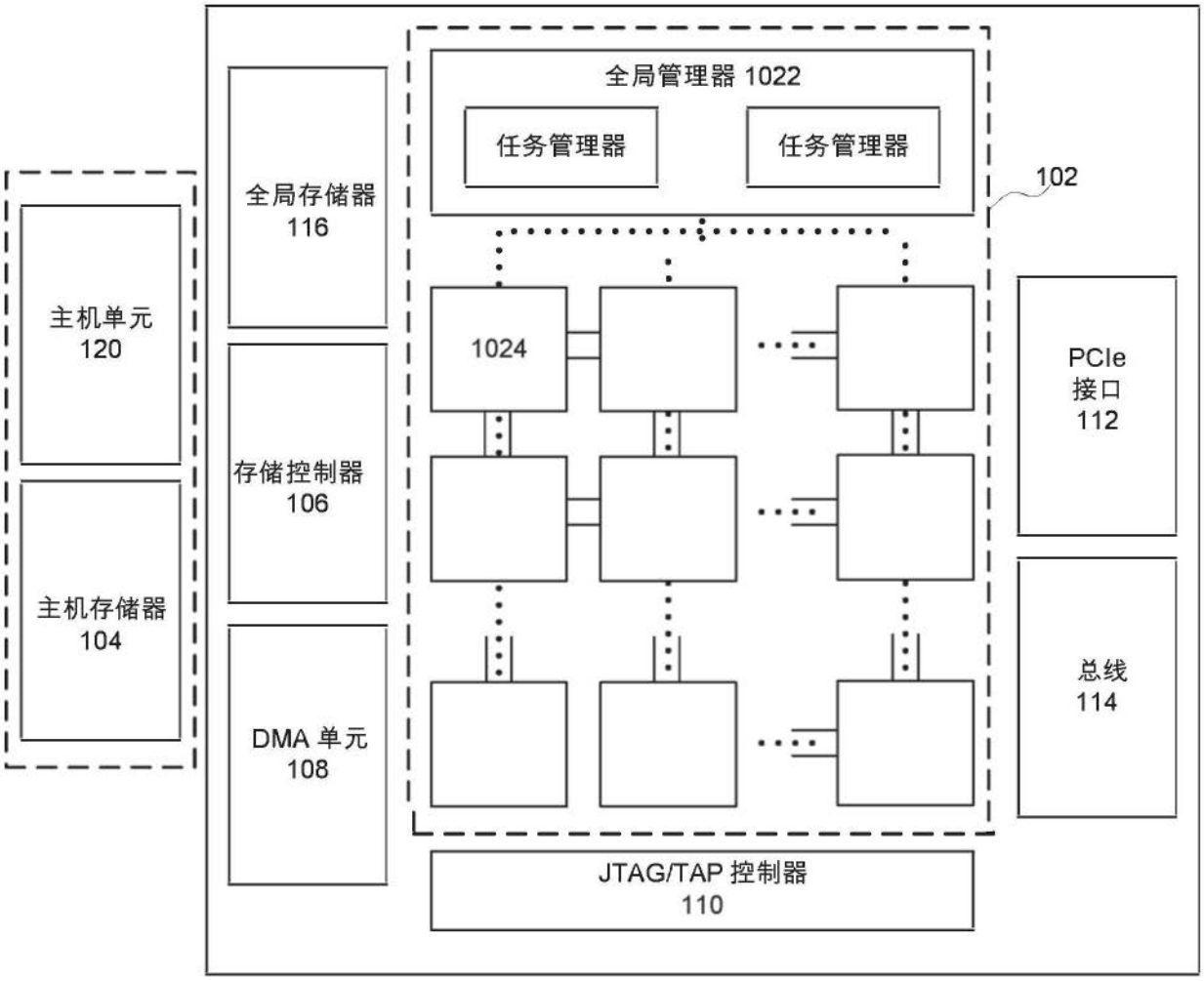

全局管理器 1022
任务管理器
任务管理器
102
全局存储器
116
主机单元
120
1024
PCIe
接口
112
存储控制器
106
主机存储器
104
总线
114
DMA 单元
108
JTAG/TAP 控制器
110
100

图1

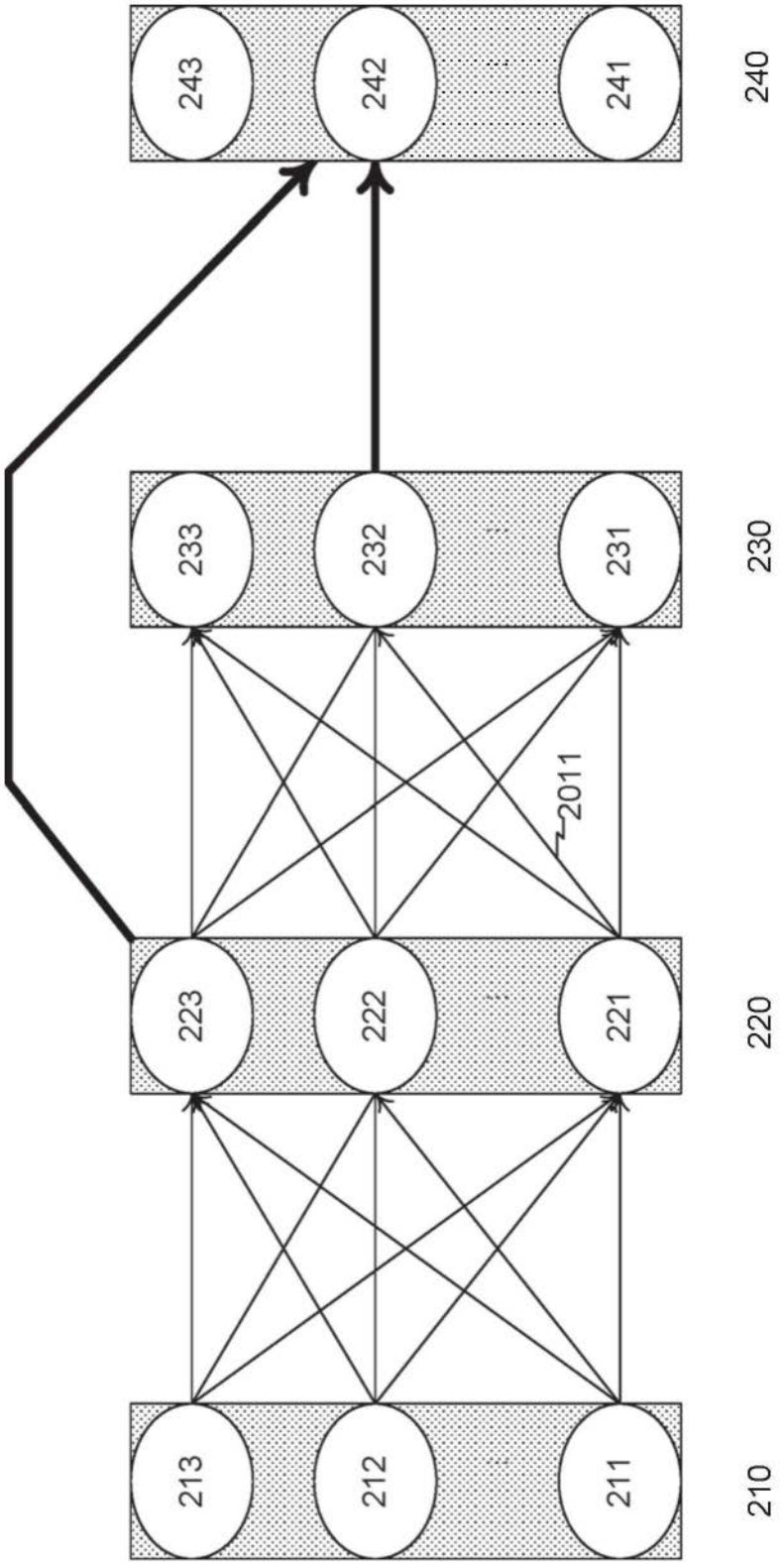

243
242
241
240
233
232
231
230
2011
223
222
221
220
213
212
211
210

图2A

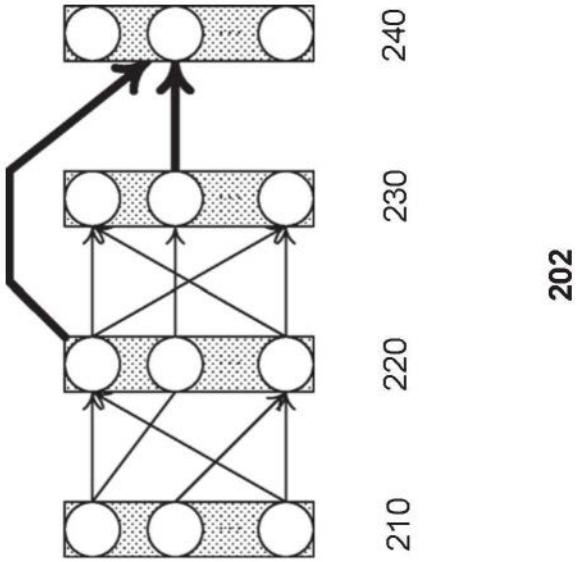


图2B

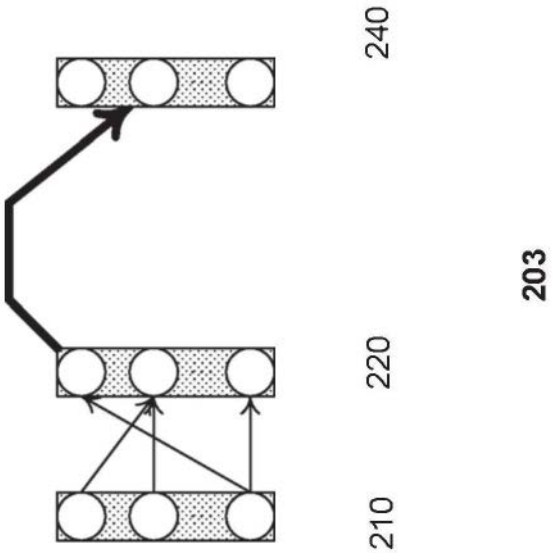


图2C

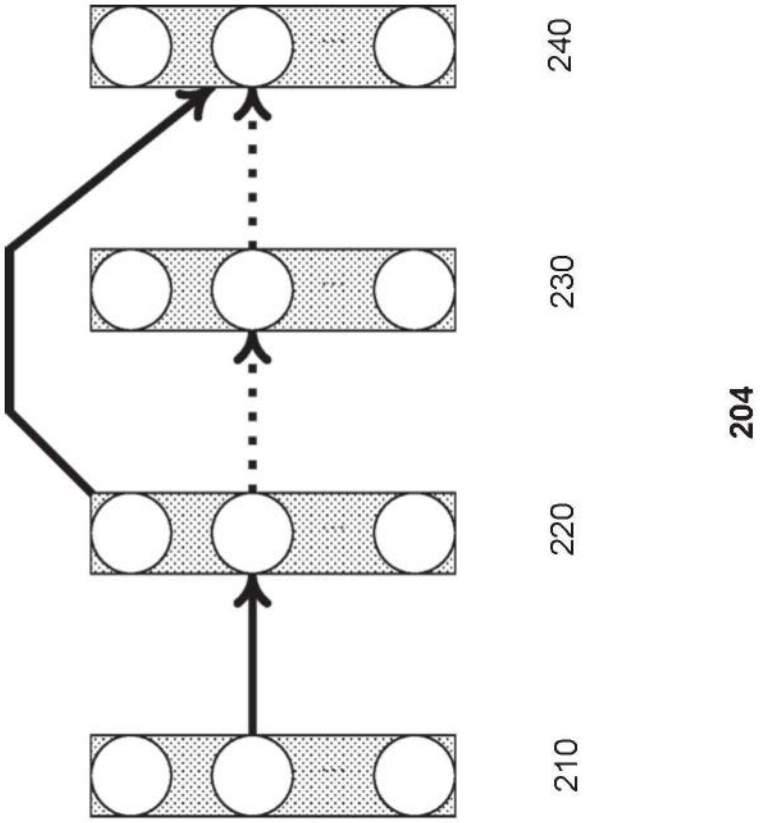


图2D

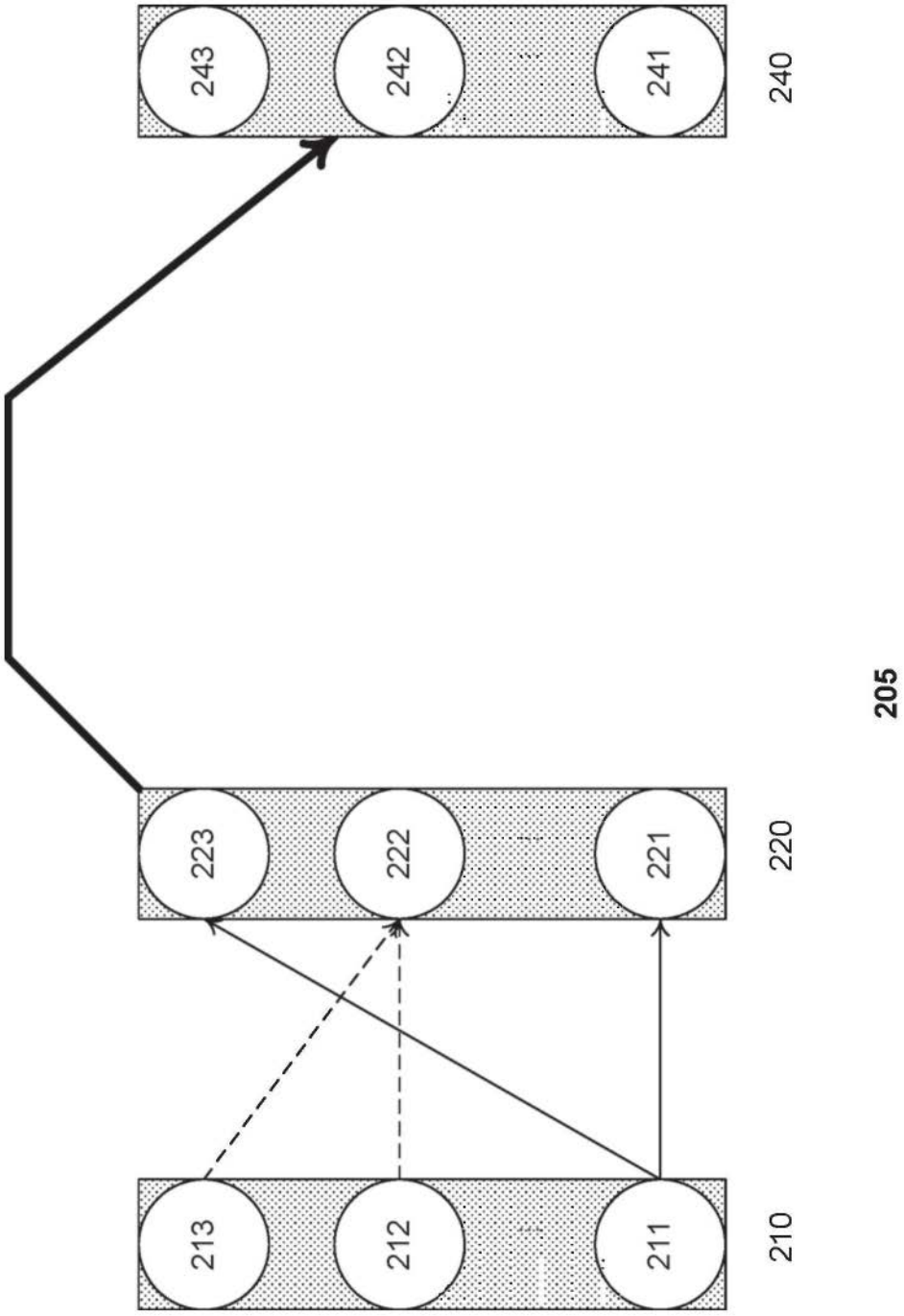


图2E

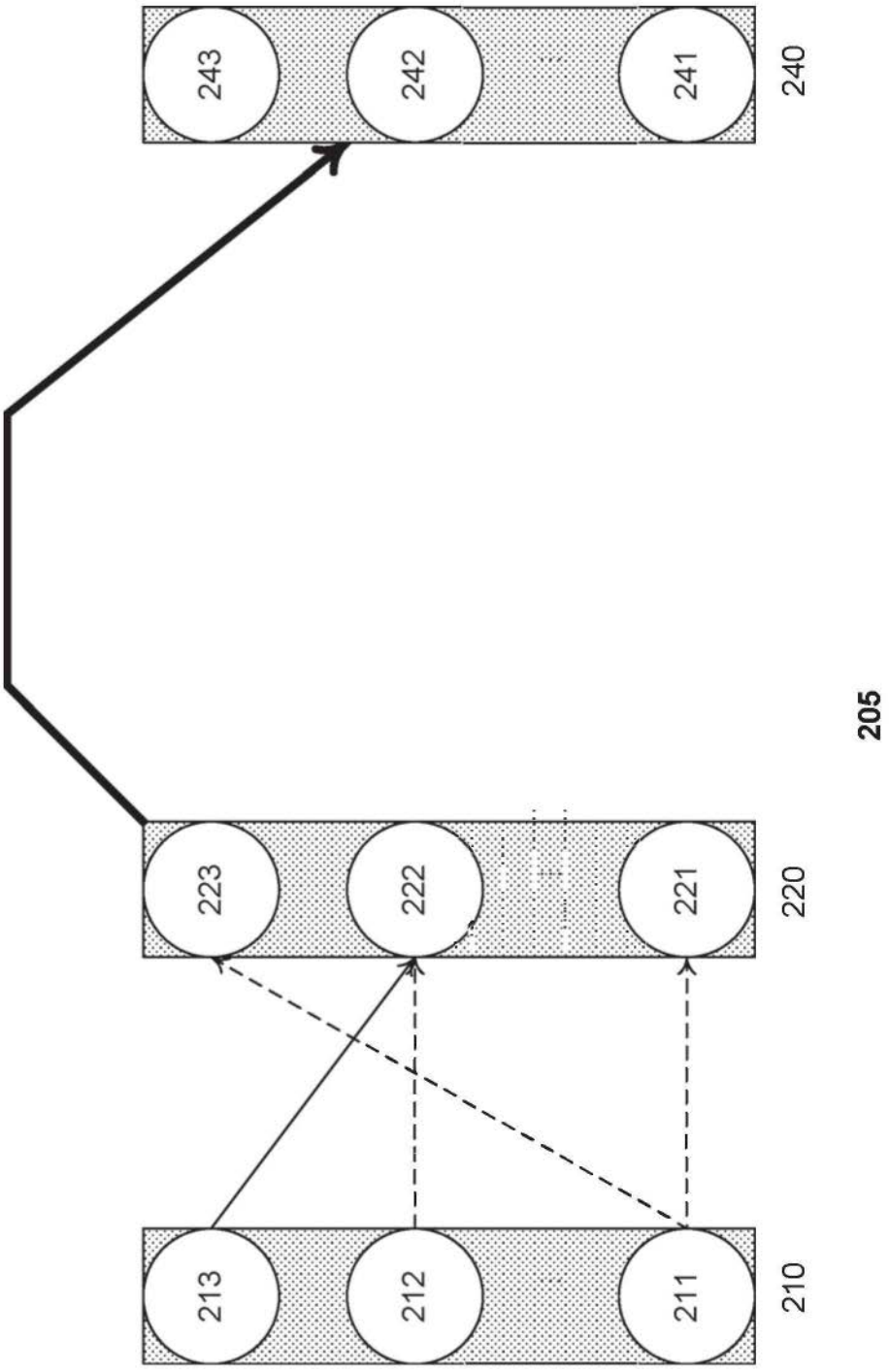


图2F

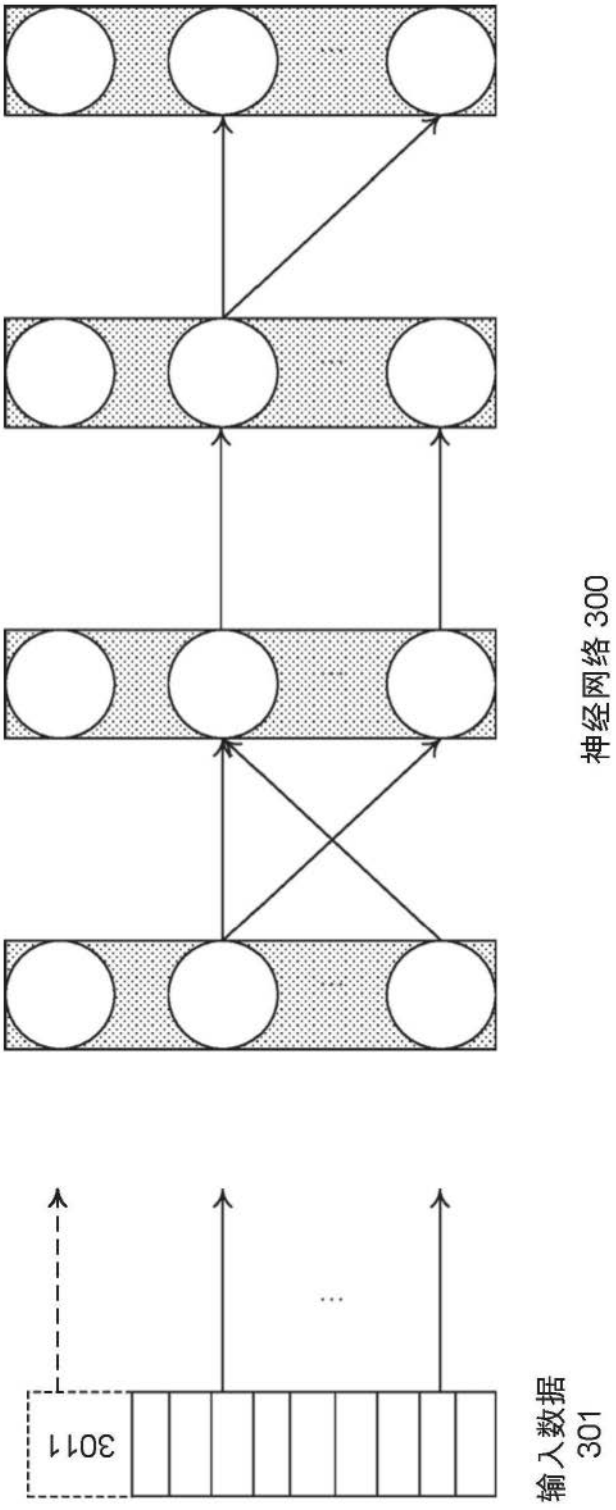


图3

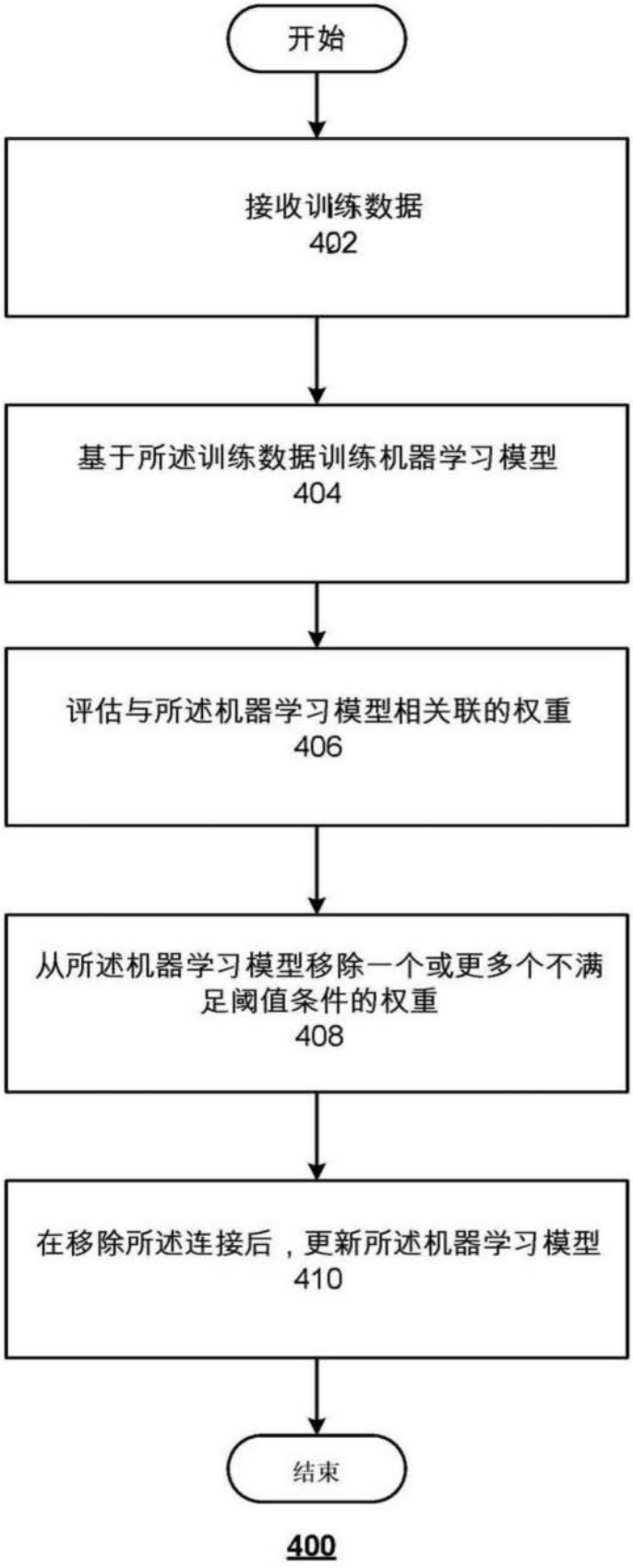

开始
接收训练数据
402
基于所述训练数据训练机器学习模型
404
评估与所述机器学习模型相关联的权重
406
从所述机器学习模型移除一个或更多个不满足阈值条件的权重
408
在移除所述连接后，更新所述机器学习模型
410
结束
400

图4

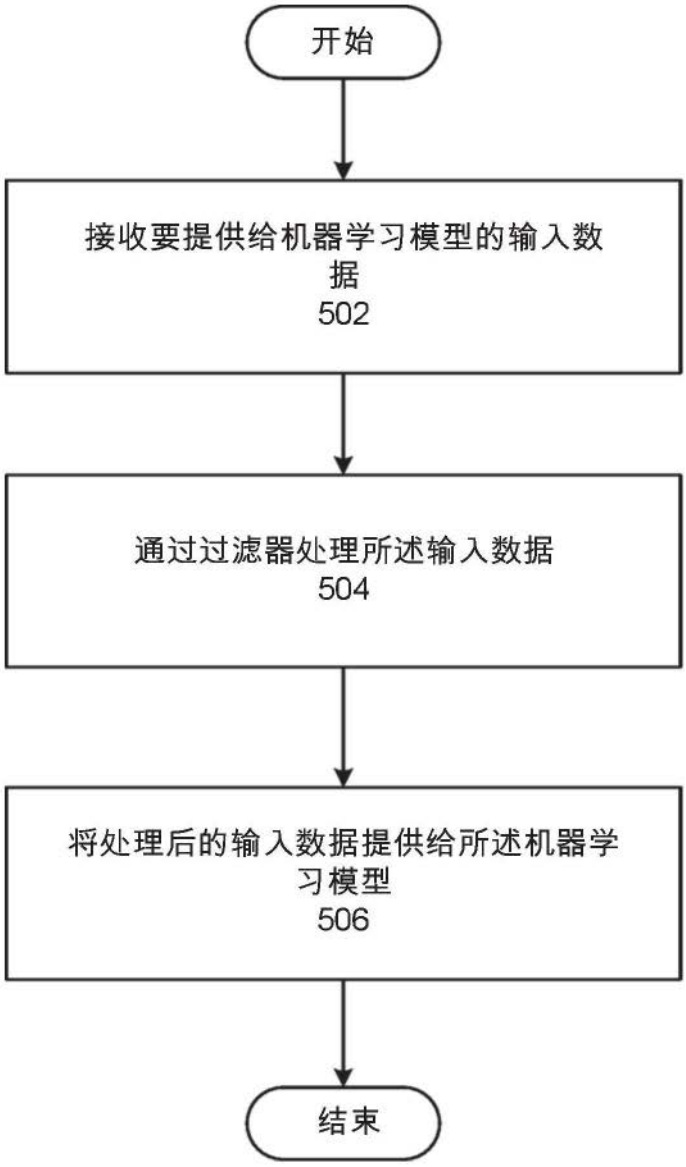

开始
接收要提供给机器学习模型的输入数据
502
通过过滤器处理所述输入数据
504
将处理后的输入数据提供给所述机器学习模型
506
结束
500

图5