(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110799994 B
(45)授权公告日 2022.07.19

(21)申请号 201880042804.4
(22)申请日 2018.06.07
(65)同一申请的已公布的文献号
申请公布号 CN 110799994 A
(43)申请公布日 2020.02.14
(30)优先权数据
15/676,701 2017.08.14 US
(85)PCT国际申请进入国家阶段日
2020.01.03
(86)PCT国际申请的申请数据
PCT/CN2018/090300 2018.06.07
(87)PCT国际申请的公布数据
WO2019/033836 EN 2019.02.21
(73)专利权人 美的集团股份有限公司
地址 528311 广东省佛山市顺德区北滘镇美的大道6号美的总部大楼B区26-28楼
(72)发明人 王奥森 周华 陈昕
(74)专利代理机构 深圳市威世博知识产权代理事务所(普通合伙) 44280
专利代理师 周心志
(51)Int.Cl.
G06N 3/04(2006.01)
(56)对比文件
US 2016328647 A1,2016.11.10
US 2016086078 A1,2016.03.24
CN 106485316 A,2017.03.08
CN 105760933 A,2016.07.13
Xi Chen et al.FxpNet: Training a Deep Convolutional Neural Network in Fixed-Point Representation.《International Joint Conference on Neural Networks》.2017,
审查员 杨杨

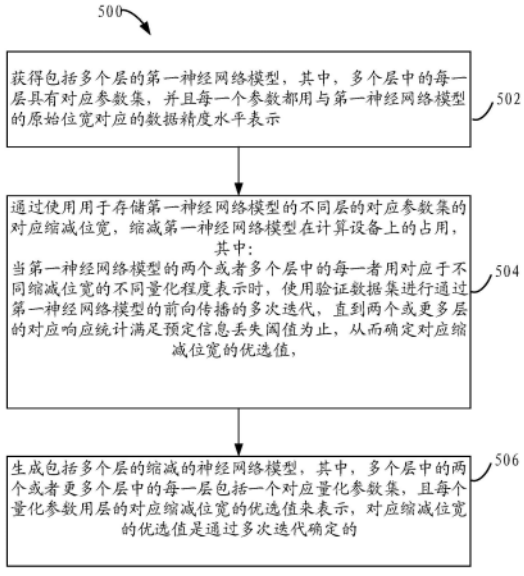
权利要求书2页 说明书15页 附图5页

(54)发明名称

神经网络的自适应位宽缩减

(57)摘要

一种在计算设备上提供自适应位宽神经网络模型的方法,包括:获得第一神经网络模型,其中,第一神经网络模型的每一层具有利用第一神经网络模型的原始位宽表示的对应参数集;通过使用用于存储第一神经网络模型的不同层的对应参数集的对应缩减位宽,缩减第一神经网络模型的占用,其中:当第一神经网络模型的两个或者多个层中的每一层用不同量化程度表示时,使用验证数据集进行通过第一神经网络模型的前向传播的多次迭代,直到满足预定信息丢失阈值为止,从而确定对应缩减位宽的优选值;以及,利用由对应缩减位宽表示的量化参数,生成缩减的神经网络模型。



CN 110799994 B

1.一种在计算设备上提供自适应位宽神经网络模型的方法,包括:

在所述计算设备处,其中,所述计算设备具有一个或多个处理器和内存:

获得包括多个层的第一神经网络模型,其中,所述多个层中的每一层具有对应参数集,并且每一个参数都用与所述第一神经网络模型的原始位宽对应的数据精度水平表示;

通过使用用于存储所述第一神经网络模型的不同层的对应参数集的对应缩减位宽,缩减所述第一神经网络模型在所述计算设备上的占用,其中:

当所述第一神经网络模型的两个或者多个层中的每一层用对应于不同缩减位宽的不同量化程度表示时,使用验证数据集进行通过所述第一神经网络模型的前向传播的多次迭代,直到所述两个或更多层的对应层响应的统计分布满足预定信息丢失阈值为止,从而确定所述对应缩减位宽的优选值;以及,

生成包括所述多个层的缩减的神经网络模型,其中的两个或者更多个层中的每一层包括一个对应量化参数集,且每个量化参数用所述层的所述对应缩减位宽的优选值来表示,所述对应缩减位宽的优选值是通过所述多次迭代确定的,

其中,所述对应层响应的统计分布为所述对应层响应的集合,

所述两个或更多层的对应层响应为所述两个或更多层的对应输出,

所述两个或更多层的对应层响应的统计分布满足预定信息丢失阈值包括所述两个或更多层的对应层响应的统计分布之间的差异的散度低于所述预定信息丢失阈值。

2.根据权利要求1所述的方法,其特征在于:

在所述缩减的神经网络模型的多个层中的第一层具有第一缩减位宽,所述第一缩减位宽小于所述第一神经网络模型的原始位宽,

在所述缩减的神经网络模型的多个层中的第二层具有第二缩减位宽,所述第二缩减位宽小于所述第一神经网络模型的原始位宽;以及,

在所述缩减的神经网络模型中,所述第一缩减位宽与所述第二缩减位宽不同。

3.根据权利要求1所述的方法,其特征在于,所述缩减所述第一神经网络模型在所述计算设备上的占用包括:

对于所述两个或更多层中具有第一参数集的第一层,所述第一参数集用对应于所述第一神经网络模型的原始位宽的数据精度水平表示:

当所述多个层的对应参数集用所述第一神经网络模型的原始位宽表示时,随着所述验证数据集作为输入通过所述第一神经网络模型前向传播,收集所述第一层的层响应的对应基准统计分布;

当所述第一层的对应参数集用小于所述第一神经网络模型的原始位宽的第一缩减位宽表示时,随着所述验证数据集作为输入通过所述第一神经网络模型前向传播,收集所述第一层的层响应的对应修改的统计分布;

确定在所述第一层的层响应的对应修改的统计分布和所述第一层的层响应的对应基准统计分布之间的差异的散度;以及,

确认所述第一缩减位宽的最小值,响应于所述第一缩减位宽的值为所述最小值,由于所述第一层的位宽的进一步缩减引起的所述差异的散度的缩减低于预定阈值,

其中,所述对应基准统计分布为当所述多个层的对应参数集用所述第一神经网络模型的原始位宽表示时,所述第一层的层响应的集合,

所述对应修改的统计分布为当所述第一层的对应参数集用小于所述第一神经网络模型的原始位宽的第一缩减位宽表示时，所述第一层的层响应的集合。

4.根据权利要求3所述的方法，其特征在于：

用所述第一缩减位宽表示所述第一层的对应参数集包括，对所述第一层的对应参数集执行非均匀量化，以生成所述第一层的第一量化参数集，以及，

在通过所述第一层的每一次前向传播期间，基于所述第一层的层响应的基准统计分布，选择所述第一层的非均匀量化的最大边界值。

5.根据权利要求1所述的方法，其特征在于，获得包括所述多个层的第一神经网络模型包括：

在所述第一神经网络的训练期间：

对于所述两个或更多层中具有第一参数集的第一层，所述第一参数集用对应于所述第一神经网络模型的原始位宽的数据精度水平表示：

根据在与所述第一层对应的第一权重集和所述第一权重集的整数部分之间的差异，获得对应于所述第一层的整数正则化项；以及，

在通过所述第一层的前向传播期间，将所述整数正则化项添加至偏差项，使得在通过所述第一层的逆向传播期间，梯度被更改，以将所述第一参数集的值推向整数值。

6.根据权利要求5所述的方法，其特征在于，获得包括所述多个层的第一神经网络模型包括：

在所述第一神经网络的训练期间：

对于所述两个或更多层中具有第一参数集的第一层，所述第一参数集用对应于所述第一神经网络模型的原始位宽的数据精度水平表示，在通过所述第一层的前向传播期间，利用小于所述第一神经网络模型的原始位宽的预定缩减位宽，对所述第一参数集执行均匀量化。

7.根据权利要求6所述的方法，其特征在于，获得包括所述多个层的第一神经网络模型包括：

在所述第一神经网络的训练期间：

对于所述两个或更多层中具有第一参数集的第一层，所述第一参数集用对应于所述第一神经网络模型的原始位宽的数据精度水平表示，在通过所述第一层的逆向传播期间，舍弃利用所述预定缩减位宽对所述第一参数集执行均匀量化。

8.一种计算设备，包括：

一个或多个处理器；以及，

内存，所述内存包括指令，当由所述一个或多个处理器执行所述指令时，导致所述处理器执行如权利要求1至权利要求7中的任一项所述的方法。

9.一种存储指令的非瞬态计算机可读存储介质，当通过一个或多个处理器执行所述指令时，导致所述处理器执行如权利要求1至权利要求7中的任一项所述的方法。

神经网络的自适应位宽缩减

技术领域

[0001] 本公开涉及机器学习实时应用,并且更具体地,涉及通过缩减机器学习模型的模型大小和计算占用(computational footprint)改善用于便携式设备和实时应用的机器学习模型,并且保持相同的精确度。

背景技术

[0002] 机器学习在许多领域具有广泛的适用性,这些领域包括计算机视觉、语音识别、机器翻译、社交网络过滤、棋牌游戏和视频游戏、医疗诊断以及许多其他领域。诸如人工神经网络(ANN)的机器学习模型是由简单单元(神经元)组成的网络,简单单元接收输入,根据该输入改变其内部状态(激活),并且根据输入和激活产生输出。该网络是通过有向加权图将某些神经元的输出连接到其他神经元的输入而形成的。权重以及计算激活的函数根据预定学习规则,通过迭代学习过程逐渐进行修改,直到实现预定的收敛为止。机器学习模型有很多变型,并且有不同的学习范式。卷积神经网络(CNN)是一类前馈网络,其由一个或多个具有全连接层(对应于ANN的层)的卷积层组成。CNN具有关联权重和池化层,且能够利用标准逆向传播训练。深度学习模型,诸如VGG16和不同类型的ResNet,是包含多于一个隐藏层的大型模型,其能够产生与人类专家相媲美的分析结果,这使它们成为许多真实世界应用的有吸引力的候选对象。

[0003] 神经网络模型通常具有几千到几百万个单位和数百万个参数。更深和更高精确度的CNN要求大量的计算资源,这使得它们对于实时应用或在电池寿命、内存和处理能力有限的便携式设备上的部署不太实用。现有的用于为各种应用部署大型神经网络模型(例如,深度学习模型)的最新解决方案专注于两种方法,模型缩减和硬件升级。专注于缩减模型结构的复杂度的模型缩减通常会极大地损害模型的精确度,而实际成本和能量消耗问题限制了硬件升级。因此,需要用于产生有效的轻量级机器学习模型的改进技术。

发明内容

[0004] 本公开描述了一种用于产生具有针对模型的不同层的参数的自适应位宽的高精确度、轻量级机器学习模型的技术。具体而言,修改了常规训练阶段,以将模型的每一层的参数朝8位范围内的整数值推进。在完成训练阶段之后,所获得的浮点精度模型(例如,FP32精度)可选地通过剪枝过程以产生精简的稀疏网络。另外,全精度模型(例如,原始训练模型或剪枝模型)通过非均匀量化转换为缩减的自适应位宽模型,例如,将FP32参数转换为其具有对应缩减位宽的量化对等物,该对应缩减位宽通过使用校准数据集的多轮测试确认。具体地,校准数据集前向传播通过带有用于不同层的不同位宽组合的数据模型的量化版本,直到确认合适的缩减位宽集(例如,最小位宽集)为止,该合适的缩减位宽集产生可接受的模型精确度水平(例如,具有低于阈值信息丢失量的信息丢失量,或者具有最小的信息丢失量)。然后部署缩减的自适应位宽模型(例如,部署在便携式电子设备上)以执行预定任务。

[0005] 一方面,一种在计算设备上提供自适应位宽神经网络模型的方法,包括:获得包括

多个层的第一神经网络模型，其中，多个层中的每一层具有对应参数集，并且每一个参数都用与第一神经网络模型的原始位宽对应的数据精度水平表示；通过使用用于存储第一神经网络模型的不同层的对应参数集的对应缩减位宽，缩减第一神经网络模型在计算设备上的占用，其中：当第一神经网络模型的两个或者多个层中的每一层用对应于不同缩减位宽的不同量化程度表示时，使用验证数据集进行通过第一神经网络模型的前向传播的多次迭代，直到两个或更多层的对应响应统计满足预定信息丢失阈值为止，从而确定对应缩减位宽的优选值；以及，生成包括多个层的缩减的神经网络模型，其中，多个层中的两个或者更多个中的每一层包括对应量化参数集，且每一个量化参数由通过多次迭代确定的用于层的对应缩减位宽的优选值表示。

[0006] 根据一些实施例，一种电子设备包括：一个或多个处理器，以及，存储一个或多个程序的内存；一个或多个程序配置成由一个或多个处理器执行，并且该一个或多个程序包括用于执行本文中所描述的方法中的任一项的操作或者引起操作的执行的指令。根据一些实施例，一种非瞬态计算机可读存储介质具有在其中存储的指令，当由电子设备执行该指令时，导致设备执行本文中所描述的方法中的任一项的操作或者导致操作的执行。根据一些实施例，一种电子设备包括：用于执行本文中所描述的方法中的任一项的操作或者引起其操作的装置。根据一些实施例，用于在电子设备中使用的一种信息处理装置包括用于执行本文中所描述的方法中的任一项的操作或者引起操作的执行的手段。

[0007] 在附图和下文的描述中陈述在该说明书中描述的本主题的一个或多个实施方式的细节。从说明书、附图以及权利要求，本主题的其他特征、方面和优势将变得显而易见。

附图说明

[0008] 图1示出根据一些实施例的示例机器学习系统在其中操作的环境。

[0009] 图2是根据一些实施例的示例模型生成系统的框图。

[0010] 图3是根据一些实施例的示例模型部署系统的框图。

[0011] 图4是根据一些实施例的示例机器学习模型。

[0012] 图5是根据一些实施例的模型缩减过程的流程图。

具体实施方式

[0013] 现在将详细参考实施例，其示例在附图中示出。在下文的详细描述中，阐述了许多具体细节以便提供对所描述的各种实施例的透彻理解。然而，对于本领域普通技术人员将显而易见的是，可以在没有这些具体细节的情况下实践各种所述实施例。在其他情形中，没有详细描述众所周知的方法、步骤、部件、电路和网络，以免不必要地使实施例的方面晦涩难懂。

[0014] 还将理解，尽管在本文中在一些情形中使用术语第一、第二等描述各种元件，但是这些元件不应该被这些术语限制。这些术语仅用于区分一个元件与另一个元件。例如，在不脱离所描述的各种实施例的范围的情况下，第一层可以被称为第二层，并且类似地，一个层可以被称为第一层。第一层和第二层都是模型的层，但是除非上下文另有明确指示，否则它们不是同一层。

[0015] 在本文中对所述各种实施例的描述中使用的术语仅出于描述具体实施例的目的，

而不旨在进行限制。如在所述各种实施例和所附权利要求的描述中使用地，单数形式“一个”、“一种”和“所述”也旨在包括复数形式，除非上下文另外明确指出。还应理解，本文所用的术语“和/或”是指并涵盖一个或多个关联所列项目的任何和所有可能的组合。还将理解，当在本说明书中使用时，术语“包括”和/或“包含”规定了所陈述的特征、整体、步骤、操作、元件和/或部件的存在，但是不预先排除存在或添加一个或多个其他特征、整体、步骤、操作、元件、部件和/或其组合。

[0016] 如本文所用地，取决于上下文，术语“如果”可选地被解释为意思是“何时”或“在……时”或“响应于确定”或“响应于检测”。类似地，取决于上下文，短语“如果确定”或“如果检测到[所陈述的状况或事件]”可选地被解释为意思是“在确定时”或“响应于确定”或“在检测到[所陈述的状况或事件]时”或“响应检测到[所陈述的状况或事件]”。

[0017] 图1是示出示例机器学习系统100的框图，模型生成系统102和模型部署系统104在该示例机器学习系统100中操作。在一些实施例中，模型生成系统102是有一个或多个处理器和内存的服务器系统，该服务器系统能够进行大型计算和数据处理任务。在一些实施例中，模型部署系统104是有一个或多个处理器和内存的便携式电子设备，与服务器系统102相比，便携式电子设备是轻量级的并且具有有限的电池电量，并且具有较小的计算和数据处理能力。在一些实施例中，模型生成系统102和模型部署系统104经由网络(例如，因特网)远程连接。在一些实施例中，模型部署系统104通过网络或通过其他文件或数据传输装置(例如，经由便携式可移动盘驱动器或光盘)从模型生成系统102接收根据本文所述的技术生成的缩减模型。

[0018] 如在图1中示出地，模型生成系统102通过使用训练数据108的语料库(例如，带有对应输出数据的输入数据)训练来生成全精度深度学习模型106(例如，有两个或更多隐藏层的CNN，并且CNN的参数(例如，权重和偏差)用占32位的单精度浮点格式(例如，FP32)表示)。如本文所用，“全精度”是指浮点精度，并且可以包括半精度(16位)、单精度(32位)、双精度(64位)、四精度(128位)、八精度(256位)和其他扩展精度格式(40位或80位)。在一些实施例中，训练包括监督训练、无监督训练或半监督训练。在一些实施例中，训练包括通过深度学习模型106的多个层的前向传播和通过深度学习模型106的多个层的逆向传播。

[0019] 在一些实施例中，在训练期间，应用整数(INT)权重正则化和8位量化技术，以将深度学习模型106的全精度参数的值推向其对应的整数值，并缩减参数的值范围，使参数的值范围落入预定缩减最大位宽(例如，8位)的动态范围内。在本说明书的后面部分中描述关于INT权重正则化和8位量化技术的更多细节。

[0020] 在一些实施例中，使用针对具体应用和输入数据定制的模型结构和训练规则执行训练。例如，如果学习模型用于语音识别系统，则将语音样本(例如，原始频谱图或线性滤波器组特征)和对应的文本用作深度学习模型的训练数据。在一些实施例中，如果学习模型用于计算机视觉，则将图像或特征以及对应类别用作学习模型的训练数据。在一些实施例中，如果学习模型用于垃圾内容检测或内容过滤，则将内容和对应的内容类别用作学习模型的训练数据。本领域技术人员将能够根据应用来选择合适的学习模型结构和合适的训练数据，并且为了简洁起见，在此不详尽列举示例。

[0021] 在一些实施例中，一旦通过迭代训练过程实现了预定的收敛，就生成全精度学习模型106’。模型106’的每一层都具有对应的参数集(例如，用于将本层中的单元连接到与本

层相邻的下一层的权重集，以及应用在连接两层的函数中的可选的一个或多个偏差项的集)。此时，所有参数都利用与学习模型106的原始位宽(例如32位)对应的原始精度水平(例如FP32)表示。请注意，即使在训练过程的前向传递中将8位量化应用于参数和中间结果，也会将浮点精度补偿标量添加到原始偏差项中以改变逆向传递中的梯度，并且所得到的模型106'的所有参数仍然用原始全精度表示，但是这些参数具有更小的动态范围，该更小的动态范围与前向传递中使用的缩减位宽(例如，8位)相对应。此外，在训练过程期间，输入也保持其全精度形式。

[0022] 如在图1中所示，在完成对全精度模型106的训练之后，所得的全精度模型106'可选地经过网络剪枝过程。在一些实施例中，使用阈值执行网络剪枝，并且权重小于预定阈值的连接(例如，弱连接)被移除，并且由这些弱连接链接的单元被从网络中移除，从而导致更轻量级且稀疏的网络106″。在一些实施例中，代替使用阈值作为移除模型106′中较弱的连接的标准，通过附加的增强训练来执行剪枝。例如，验证数据集110用于测试模型106'的修改版本。每次使用验证数据集测试模型106时，都会从模型106中移除不同的连接，或者随机将先前移除的连接添加回模型106。使用验证数据集110评估具有不同连接组合的修改后模型的精确度，并且基于总连接数目和网络精确度的预定剪枝标准(例如，标准是基于指标值的，该指标值是网络精确度的量度和在网络中的总连接数目的倒数的和)用于按照在保存网络精确度和缩减网络复杂度之间的平衡确定连接的最佳组合。

[0023] 如图1所示，一旦完成可选的网络剪枝过程，就生成精简的全精度学习模型106″。该精简的全精度学习模型106″用作后续量化和位宽缩减过程的基础模型，以生成缩减的自适应位宽模型112。在一些实施例中，如果不执行网络剪枝，则将训练后的全精度模型106'用作后续量化和位宽缩减过程的基础模型，以生成缩减的自适应位宽模型112。缩减的自适应位宽模型112是模型生成系统102的输出。缩减的自适应位宽模型112然后被提供给模型部署系统104，用于在应用中处理真实世界输入。在一些实施例中，在模型106的训练过程中不应用整数权重正则化和/或8位前向量化，并且使用常规的训练方法来生成全精度学习模型106'。如果未执行整数权重正则化和/或8位前向量化，则由于没有训练参数向整数值移动，且对于最终缩减模型112中的位宽的大幅缩减而言，值的动态范围可能太大，因此所得到的自适应位宽模型的精确度可以不那么好。但是，本文所述的自适应位宽缩减技术仍然能够实现模型大小的期望缩减，而不会导致棘手的模型精确度数量的丢失。因此，在一些实施例中，在训练阶段期间也可以独立于整数权重正则化和8位前向量化来使用本文所述的自适应位宽缩减，但是如果两种技术组合使用则可以获得更好的结果。缩减位宽如何被选定用于不同层或者模型112的参数如何确定以及如何从全精度基础模型106'或106″根据预定非线性量化方法(例如，对数量化)生成缩减位宽模型112的更多细节在本说明书的后面部分中提供。

[0024] 如图1所示，一旦缩减的自适应位宽模型112被提供给在模型部署系统104上的部署平台116，真实世界输入数据或测试数据114被馈送至缩减的自适应位宽模型112，且响应于输入数据由缩减的自适应位宽模型112生成最终预测结果118。例如，如果模型112被训练用于语音识别任务，则真实世界输入数据可以是语音输入片段(例如，语音输入的波形或记录)，并且输出可以是与语音输入片段对应的文本。在另一示例中，如果模型112被训练用于计算机视觉任务，则真实世界输入数据可以是图像或者图像特征集，并且输出可以是能够

用于分类图像的图像类别。在其他示例中，如果模型112被训练用于内容过滤，真实世界输入数据可以是内容(例如，网络内容或者电子邮件内容)，且输出数据可以是内容类别或者内容过滤动作。根据模型112被针对训练的应用，本领域技术人员将能够将合适的输入数据输入到缩减的自适应位宽模型112中，并适当地获得和利用输出。为了简洁起见，在此未详尽列举示例。

[0025] 图1仅仅是示例，且根据各种实施例，机器学习系统100的操作环境、工作流程和结构的其他配置是可能的。

[0026] 图2是根据一些实施例的模型生成系统200的框图。根据一些实施例，模型生成系统200可选地用作图1中的模型生成系统102。在一些实施例中，模型生成系统200包括一个或多个处理单元(或“处理器”)202、内存204、输入/输出(I/O)接口206和可选的网络通信接口208。这些部件通过一个或多个通信总线或信号线210彼此通信。在一些实施例中，内存204或内存204的计算机可读存储介质、存储程序、模块、指令和数据结构包括以下各项中的全部或者子集：操作系统212、输入/输出(I/O)模块214、通信模块216和模型生成模块218。一个或多个处理器202耦合到内存204，且可操作以执行这些程序、模块和指令以及对数据结构进行读写。

[0027] 在一些实施例中，处理单元202包括一个或多个微处理器，诸如单核微处理器或者多核微处理器。在一些实施例中，处理单元202包括一个或多个通用处理器。在一些实施例中，处理单元202包括一个或多个专用处理器。在一些实施例中，处理单元202包括一个或多个服务器计算机、个人计算机、移动设备、手持计算机，平板计算机或包含一个或多个处理单元并在各种操作系统上运行的多种硬件平台之一。

[0028] 在一些实施例中，内存204包括诸如DRAM、SRAM、DDR RAM的高速随机存取存储器或其他随机存取固态存储设备。在一些实施例中，内存204包括非易失存储器，诸如一个或多个磁盘存储设备、光盘存储设备、闪存存储设备或其他非易失性固态存储设备。在一些实施例中，内存204包括远离处理单元202定位的一个或多个存储设备。内存204或替代地在内存204内的一个或多个非易失存储设备包括计算机可读存储介质。

[0029] 在一些实施例中，I/O接口206将诸如显示器、键盘、触摸屏、扬声器和麦克风的输入/输出设备耦合至语音识别系统200的I/O模块214。I/O接口206与I/O模块214一起接收用户输入(例如，语音输入、键盘输入、接触输入等)，且相应地处理它们。I/O接口206和用户接口模块214还根据在模型生成系统200上实施的各种程序指令将输出(例如，声音、图像、文本等)呈现给用户。

[0030] 在一些实施例中，网络通信接口208包括一个或多个有线通信端口和/或无线发送和接收电路。一个或多个有线通信端口经由一个或多个有线接口，例如，以太网、通用串行总线(USB)、火线等接收和发送通信信号。无线电路从通信网络和其他通信设备接收RF信号和/或光信号并将其发送到通信网络和其他通信设备。无线通信可以使用多种通信标准、协议和技术中的任一种，诸如GSM、EDGE、CDMA、TDMA、蓝牙、Wi-Fi、VoIP、Wi-MAX或任何其他合适的通信协议。网络通信接口208允许在模型生成系统200与诸如因特网、专用网的网络和/或诸如移动电话网络、无线局域网(LAN)的无线网络和/或城域网(MAN)以及其他设备之间的通信。通信模块216便于在模型生成系统200和其他设备(例如，模型部署系统300)之间通过网络通信接口208的通信。

[0031] 在一些实施例中，操作系统202(例如，Darwin、RTXC、LINUX、UNIX、OS X、WINDOWS或诸如VxWorks的嵌入操作系统)包括各种软件组件和/或驱动，这些软件组件和/或驱动用于控制和管理一般系统任务(例如，内存管理、存储设备控制、功率管理等)，并且便于在各种硬件、固件和软件组件之间的通信。

[0032] 在一些实施例中，在独立的计算机系统上实施模型生成系统200。在一些实施例中，模型生成系统200分布在多台计算机上。在一些实施例中，模型生成系统200的一些模块和功能位于第一组计算机上，并且模型生成系统200的一些模块和功能位于与第一组计算机不同的第二组计算机上；并且，两组计算机通过一个或多个网络彼此通信。

[0033] 应当注意，在图2中示出的模型生成系统200仅是模型生成系统的一个示例，并且该模型生成系统200可以具有相比示出的部件更多或更少的部件，可以组合两个或者更多个部件，或者可以具有不同的部件配置或者布置。图2所示的各种部件可以以硬件、软件、固件的方式实施，包括一个或多个信号处理和/或专用集成电路或其组合。

[0034] 如图2所示，模型生成系统200将模型生成模块218存储在内存204中。在一些实施例中，模型生成模块218还包括以下子模块或其子集或超集：训练模块220、整数权重正则化模块222、8位前向量化模块224、网络剪枝模块226、自适应量化模块228和部署模块230。在一些实施例中，部署模块230可选地执行模型部署系统的功能，使得能够在将缩减的自适应位宽模型实际部署到单独的模型部署系统上(诸如在便携式电子设备上之前)，用真实世界数据来测试缩减的自适应位宽模型。

[0035] 此外，如图2所示，这些模块和子模块中的每一个都可以访问以下数据结构和模型或其子集或超集中的一个或多个：训练语料库232(例如，包含图1中的训练数据108)、验证数据集234(例如，包含图1中的验证数据集110)、全精度模型236(例如，启动时为未经训练的全精度模型106，训练之后转换为经训练的全精度模型106')、精简的全精度模型238(例如，图1中的剪枝模型106″)和缩减的自适应位宽模型240(例如，图1中的缩减的自适应位宽模型112)。关于图1、图4和图5以及随附的描述，提供了关于模型生成系统200的子模块和数据结构的结构、功能以及相互作用的更多细节。

[0036] 图3是根据一些实施例的模型部署系统300的框图。根据一些实施例，模型部署系统300可选地用作图1中的模型部署系统104。在一些实施例中，模型部署系统300包括一个或多个处理单元(或“处理器”)302、内存304、输入/输出(I/O)接口306和可选的网络通信接口308。这些部件通过一条或多条通信总线或信号线310彼此通信。在一些实施例中，内存304或内存304的计算机可读存储介质存储包括以下各项的全部或子集的程序、模块、指令和数据结构：操作系统312、I/O模块314、通信模块316以及模型部署模块318。一个或多个处理器302耦合到内存304，并且可操作来执行这些程序、模块和指令，以及对数据结构进行读写。

[0037] 在一些实施例中，处理单元302包括一个或多个微处理器，即单核或多核微处理器。在一些实施例中，处理单元302包括一个或多个通用处理器。在一些实施例中，处理单元302包括一个或多个专用处理器。在一些实施例中，处理单元302包括一个或多个服务器计算机、个人计算机、移动设备、手持计算机，平板计算机或包含一个或多个处理单元并在各种操作系统上运行的多种硬件平台之一。

[0038] 在一些实施例中，内存304包括诸如DRAM、SRAM、DDR RAM的高速随机存取存储器或

其他随机存取固态存储设备。在一些实施例中，内存304包括非易失存储器，诸如一个或多个磁盘存储设备、光盘存储设备、闪存存储设备或其他非易失性固态存储设备。在一些实施例中，内存304包括远离处理单元302定位的一个或多个存储设备。内存304或替代地在内存304内的一个或多个非易失存储设备包括计算机可读存储介质。

[0039] 在一些实施例中，I/O接口306将诸如显示器、键盘、触摸屏、扬声器和麦克风的输入/输出设备耦合至模型部署系统300的I/O模块314。I/O接口306与I/O模块314一起接收用户输入(例如，语音输入、键盘输入、接触输入等)，且相应地处理它们。I/O接口306和用户接口模块314还根据在模型部署系统300上实施的各种程序指令将输出(例如，声音、图像、文本等)呈现给用户。

[0040] 在一些实施例中，网络通信接口308包括一个或多个有线通信端口和/或无线发送和接收电路。一个或多个有线通信端口经由一个或多个有线接口，例如，以太网、通用串行总线(USB)、火线等接收和发送通信信号。无线电路从通信网络和其他通信设备接收RF信号和/或光信号并将其发送到通信网络和其他通信设备。无线通信可以使用多种通信标准、协议和技术中的任一种，诸如GSM、EDGE、CDMA、TDMA、蓝牙、Wi-Fi、VoIP、Wi-MAX或任何其他合适的通信协议。网络通信接口308允许在模型部署系统300与诸如因特网、专用网的网络和/或诸如移动电话网络、无线局域网(LAN)的无线网络和/或城域网(MAN)以及其他设备之间的通信。通信模块316便于在模型部署系统300和其他设备(例如，模型生成系统200)之间通过网络通信接口308的通信。

[0041] 在一些实施例中，操作系统302(例如，Darwin、RTXC、LINUX、UNIX、OS X、WINDOWS或诸如VxWorks的嵌入操作系统)包括各种软件组件和/或驱动，这些软件组件和/或驱动用于控制和管理一般系统任务(例如，内存管理、存储设备控制、功率管理等)，并且便于在各种硬件、固件和软件组件之间的通信。

[0042] 在一些实施例中，在独立的计算机系统上实施模型部署系统300。在一些实施例中，模型部署系统300分布在多台计算机上。在一些实施例中，模型生成系统300的一些模块和功能位于第一组计算机上，并且模型生成系统300的一些模块和功能位于与第一组计算机不同的第二组计算机上;并且，两组计算机通过一个或多个网络彼此通信。应当注意，在图3中示出的模型部署系统300仅是模型部署系统的一个示例，并且该模型部署系统300可以具有相比示出的部件更多或更少的部件，可以组合两个或者更多个部件，或者可以具有不同的部件配置或者布置。图3所示的各种部件可以以硬件、软件、固件的方式实施，包括一个或多个信号处理和/或专用集成电路或其组合。

[0043] 如图3所示，模型部署系统300将模型部署模块318存储在内存304中。在一些实施例中，部署模块318可以访问以下数据结构和模型中的一个或多个或其子集或超集:输入数据320(例如，包含图1中的真实世界数据114)、缩减的自适应位宽模型322(例如，图1中的缩减的自适应位宽模型112和图2中的缩减的自适应位宽模型240)和输出数据324。关于图1、图4和图5以及随附的描述，提供了关于模型部署系统300的子模块和数据结构的结构、功能以及相互作用的更多细节。

[0044] 图4示出全精度深度学习模型(例如，有多个隐藏层的人工神经网络)的结构和训练过程。学习模型包括由圆圈代表的一些单位(或神经元)，在这些单元之间具有连接(或突触)。连接具有关联的权重，每个权重代表先前层(例如，第i层)中的神经元的输出对下一层

(例如,第i+1层)中的神经元的影响。每一层中的神经元都将来自从先前层连接至该神经元的所有神经元的输出相加,并应用激活函数以获得响应值。训练过程是使用在输入层中提供的训练数据集为学习模型的每一层校准所有权重W_i的过程。

[0045] 训练过程通常包括两个步骤,即前向传播和逆向传播,这些步骤被重复多次直到满足预定收敛条件。在前向传播中,用于不同层的权重集被应用于来源于先前层的中间结果和输入数据。在逆向传播中,测量输出的误差幅度,并相应地调整权重以减少误差。激活函数能够是线性的、修正线性单位、Sigmoid函数、双曲正切或其他类型。

[0046] 通常,在应用激活函数之前,将网络偏差项b添加到来自先前层的加权输出的和中。网络偏差提供了必要的干扰,这帮助该网络避免对训练数据的过拟合。以下是常规神经网络的工作流程示例:

[0047] 训练以如下组件开始:

[0048] 输入:图像$I_{n\times m}$,输出:标签$B_{k\times 1}$,以及模型:L层神经网络。

[0049] L层神经网络的每个层i被描述为

[0050] $y_{p\times q}=g(W_{p*q}\times x_{q\times 1}+b_{p\times 1})$,

[0051] 其中,y是层响应向量,W是权重矩阵,x是输入向量,g(x)=max(0,x)是激活函数,并且b是偏差向量。(请注意,p和q是层维度,且在不同的层中将是不同的)。

//Training Phase

While *res*has not converged:

//Network Forward pass

$x_1=I$

For i = 1 to L:

$y_i=g(W_i\times x_i+b_i)$

[0052] $x_{i+1}=y_i$

$res=\frac{1}{2}\|y_L-B\|_2$ //a measure of margin of error//

//Backward pass

For i = L to 1:

$$W_i=W_i-\gamma\frac{\partial \text{res}}{\partial W_{i,j}}$$ // γ is a parameter

$$b_i=b_i-\gamma\frac{\partial \text{res}}{\partial b_{i,j}}$$

[0053] 训练结果包括:每层的网络权重参数W,每层的网络偏差参数b。

//Testing Phase: To classify an unseen image I_{un}

[0054] $x_1=I_{un}$

For i = 1 to L:

$y_i=g(\boldsymbol{W}_i\times\boldsymbol{x}_i+\boldsymbol{b}_i)$

[0055] $x_{i+1}=y_i$

Return y_L

[0056] 在本公开中，根据一些实施例，将整数(INT)权重正则化项添加到每一层的权重W，其中，W_i是层i的权重，并且$\lfloor W_i \rfloor$是权重W_i的按元素(element-wise)的整数部分。利用该正则化，训练将权重的小数部分(decimals)作为惩罚，并且这能够在训练后将网络中的所有全精度(例如，FP32)权重推向其对应的整数值。

[0057] 此外，对通过网络的前向传播中的所有权重和中间结果执行8位均匀量化。选择全精度(例如，FP32)补偿标量以调整权重的值范围，使得权重的值范围由预定最大位宽(例如，8位)很好地表示。一般而言，对于保留权重分布的显著信息来说，8位是一个相对宽裕的位宽。该量化将在逆向传递中改变梯度，以约束不同层的权重值范围。该量化位宽稍后用作缩减的自适应位宽模型的最大允许缩减位宽。

[0058] 以下是添加了整数权重正则化和8位均匀量化的样本训练过程。

[0059] 在此训练过程中，我们还将用以下组件启动：

[0060] 输入：图像$I_{n\times m}$，输出：标签$B_{k\times 1}$，以及模型：L层神经网络。

[0061] (x)代表均匀量化，其中$(x) = \lfloor x/X_{max} \rfloor * 127 * (X_{max}/127.0)$，其中，$X_{max}$是x(整个层响应)的当前最大边界值，其将在训练阶段中进行更新。

//Training Phase

While *res*has not converged:

//Network Forwarding

[0062] $x_1 = I$

For i = 1 to L:

$y_i = g(\boldsymbol{Qu}(\boldsymbol{W_i}) \times \boldsymbol{x_i} + \boldsymbol{Qu}(\boldsymbol{b_i}) + \boldsymbol{RI_i})$

$x_{i+1} = y_i$

$res=\frac{1}{2}\|y_L-B\|_2$ //a measure of margin of error//

//Backward pass

For i = L to 1:

[0063] $$W_i = W_i - \gamma \frac{\partial \text{res}}{\partial W_{i,j}} \quad // \gamma \text{ is a parameter}$$

$$b_i = b_i - \gamma \frac{\partial \text{res}}{\partial b_{i,j}}$$

[0064] 训练的结果包括:用全精度(例如,单精度浮点)格式表示的每一层的网络权重参数W,每一层的网络偏差参数b。这些结果与前面所述的常规训练的结果不同。将全精度参数(例如,权重W和偏差b)推向其最接近的整数,而对模型精确度的影响最小。为了避免大的信息丢失,通过8位前向量化来约束整数值的值范围。在均匀量化之前,卷积层和全连接层中的偏差项对网络的紧凑性有害。本文提出,卷积神经网络中的偏差项不会降低整个网络的精确度。当对权重参数应用8位均匀量化时,权重参数仍用全精度格式表示,但是此类响应级别(level)的总数受2^8-1约束,其中,总数中的一半为正数,且一半为负数。在训练阶段的每次迭代中,会在前向传递的每个层中应用该量化功能。逆向传播仍然使用全精度数字进行可学习的参数的更新。因此,利用补偿参数X_{max},有效地约束了权重整数值的范围。

[0065] 在满足收敛条件之后(例如,当res已收敛时),获得了全精度训练后学习模型(例如,模型106')。该模型具有高精确度和高复杂度,并且在测试真实世界数据时具有较大的占用(例如,计算、功耗消耗、内存使用等)。

[0066] 在一些实施例中,网络剪枝被用于降低网络复杂度。通常,设置阈值权重。仅高于阈值的权重保持不变,而低于阈值权重的权重被设置为零,并从网络中移除与零权重相对应的连接。未与任何其他神经元连接的神经元(例如,由于一个或多个连接的移除)被实际上从网络中移除,从而导致更紧凑和稀疏的学习模型。传统的剪枝技术是强制性的,并且会导致大量信息丢失,而且极大损害了模型精确度。

[0067] 在本公开中,验证数据集用于执行增强学习,该增强学习使用在层之间的连接的不同子集组合来测试原始的全精确训练模型(例如,模型106')的修改版本的精确度。该问题被视为权重选择游戏,并且应用增强学习以寻找兼顾对更好的模型精确度和模型紧凑性的期望的最佳解决方案(例如,连接的最佳组合)。在一些实施例中,剪枝有效性Q的量度是网络精确度(例如,如通过在原始模型与具有缩减连接的模型之间的层响应的Jensen-Shannon散度所测量的网络精确度)和总连接数的倒数之和。增强学习的结果是网络中更有价值的权重的子集。在一些实施例中,在增强学习的每次迭代期间,从网络移除一个连接,或者将先前移除的一个连接添加回网络。在每次迭代之后评估网络精确度。在训练被执行一段时间之后,将出现一个带有稀疏权重(和神经元)的精简网络。

[0068] 网络剪枝过程的结果包括:用全精度浮点格式表示,每一层的网络权重参数Wr,每一层的网络偏差参数br。

[0069] 在一些实施例中,在剪枝之后,精简的全精度模型(例如,模型106″)经过自适应量化过程以产生缩减的自适应位宽的精简INT模型(例如,模型112)。模型的自适应位宽是指以下特性:对于模型的每一层,用于存储参数集(例如,权重和偏差)的对应位宽是针对该参

数集特别选择的(例如,根据参数的分布和范围)。具体地,将验证数据集用作通过剪枝的精简全精度网络(例如,模型106″)的前向传递中的输入,并且收集每一层中的响应值的统计分布。然后,准备不同的位宽和层组合配置,作为用于评估的候选项。对于每个候选模型,将验证数据集用作通过候选模型的前向传递的输入,并且收集每一层中响应值的统计分布。然后,基于从将量化应用到候选模型所导致的信息丢失的数量,评估候选模型。在一些实施例中,每一层(或用于模型整体)的两个统计分布之间的Jensen-Shannon散度用于确认具有用于该层的最小信息丢失的最佳位宽。在一些实施例中,通过对所有层使用不同位宽组合,生成量化候选;作为替代,基于权重的值,对来自不同层的权重进行聚类,并且通过对所有聚类使用不同的位宽组合生成量化候选。

[0070] 在一些实施例中,代替对模型的每一层的模型参数使用线性或均匀量化,将非线性量化应用于不同层的全精度参数。常规的线性量化不考虑参数值的分布,并且导致大量的信息丢失。在全精度参数上进行非均匀量化(例如,对数量化)时,更多的量化级别被赋予具有较大值的子级别,并且导致量化误差减少。能够以如下公式表示对数量化:

[0071] $$y(x) = X_{max} * \frac{log\left(1+\left\lfloor R*\frac{x}{X_{max}}\right\rfloor\right)}{log(1+R)},$$

[0072] 该公式以具有R级别的间隔[0,X_{max}]量化x。注意符号$\lfloor . \rfloor$表示找到不超过符号内数字(例如,具有FP32精度的数字)的最大整数。例如,如果我们设置X_{max}=105.84且R=15,则我们将具有在{0,26.4600,41.9381,52.9200,61.4382,68.3981,74.2826,79.3800,83.8762,87.8982,91.5366,94.8581,97.9136,100.7426,103.3763,105.8400}中的y(x)。与均匀量化相比较,这种非均匀方案将为具有更大值的子间隔分配更多的量化级别。在自适应量化中,X_{max}是根据预定信息丢失标准而不是根据间隔的实际最大值来学习的。在实践中采取的另一步骤是,实际值范围可以在[X_{min},X_{max}]中,从X_{max}减去X_{min}以对其范围进行归一化,以与上述讨论一致。

[0073] 如上所述,在一些实施例中,信息丢失的预定量度是测量两个统计分布之间的差异的Jensen-Shannon散度。在这种情况下,统计分布是全精度训练模型(例如,模型106'或106″)以及具有用于量化候选模型的层的特定位宽组合的量化候选模型中的所有层(或对应层)的全层响应的集合。Jensen-Shannon散度用以下公式表示:

[0074] $$\boldsymbol{JSD(P\|Q) = \frac{1}{2}D(P\|M) + \frac{1}{2}D(Q\|M)}$$

[0075] 其中,$\boldsymbol{M = \frac{1}{2}(P+Q)}$,注意,P和Q是两个独立的数据分布(例如,原始模型和候选模型中同一层(或所有层)的层响应分布)。D(P‖Q)是从Q到P的Kullback-Leibler散度,其能够根据如下公式计算:

[0076] $$\boldsymbol{D(P\|Q) = \sum_i P(i) log\left(\frac{P(i)}{Q(i)}\right)}。$$

[0077] 注意,Kullback-Leibler散度是不对称的。JSD值越小,对应的信息丢失就越小。能够基于将信息丢失约束在预定阈值以下或者找到产生最小信息丢失的位宽组合来选择候选模型。

[0078] 在一些实施例中,校准数据集被用作通过不同候选模型(例如,不同候选模型有用

于不同层的参数(例如,权重和偏差)和中间结果(例如,层响应)的不同位宽组合)的前向传播传递中的输入。

[0079] 在下文的选择不同层的最佳位宽组合的样本过程中,S是校准数据集,Statistics_i是第i层响应的统计分布。Q_{non}(x,qb)是非均匀量化函数,R是量化级别数。qb是用于量化的位宽。

[0080] $Q_{non}(x) = X_{max} * \frac{log\left(1+\left\lfloor R*\frac{x}{X_{max}}\right\rfloor\right)}{log(1+R)}$,其中,$R=2^{qb}-1$

[0081] 在样本自适应位宽非均匀量化过程中,基础模型是全精度训练模型106'或剪枝精简全精度模型106″,其中,用于全精度模型的每一层i的对应权重集是W_i(或Wr_i)且偏差是b_i(或br_i)。

For $Iz \in S$: // S is the calibration data set

For i = 1 to L:

$\boldsymbol{y_i}=g(\boldsymbol{W_i}\times\boldsymbol{x_i}+\boldsymbol{b_i})$

$\boldsymbol{x_{i+1}}=\boldsymbol{y_i}$

$\boldsymbol{Statistics_i}=\boldsymbol{Statistics_i} \cup \boldsymbol{y_i}$

[0082] **For** i = 1 to L:

For $qb1$ = 1 to 8: // quantization bit-width candidates for weights

For $qb2$ = 1 to 8: // quantization bit-width for layer response

$\boldsymbol{W_{q,}}=\boldsymbol{Q_{Non}}(\boldsymbol{W_i}, qb1), \boldsymbol{b_{q,i}}=\boldsymbol{Q_{Non}}(\boldsymbol{b_i}, qb1)$

$\boldsymbol{Stat_{q,}}=\boldsymbol{Q_{Non}}(\boldsymbol{W_{q,i}}\times\boldsymbol{x_i}+\boldsymbol{b_i}, qb2)$

Inf_tmp = **InformationLoss**($\boldsymbol{Stat_{q,i}}$,$\boldsymbol{Statistics_i}$)

If Inf_tmp < Inf_min:

$\boldsymbol{W_{opt,i}}=\boldsymbol{W_{q,i}}, \boldsymbol{b_{opt,i}}=\boldsymbol{b_{q,i}}$

[0083] $\boldsymbol{qb_{opt,}}=qb2$

Inf_min = Inf_tmp

[0084] 上述过程的结果是具有每一层i的一个或多个最佳量化位宽的量化权重集$W_{opt,i}$和具有每一层i的一个或多个最佳量化位宽的量化偏差集$b_{opt,i}$。因此获得自适应位宽模型

112。此外，还获得层响应qb2的最佳量化位宽。

[0085] 在模型部署阶段，根据上文所述的方法获得的缩减的自适应位宽模型(例如，模型112)在模型部署系统(例如，便携式电子设备)上使用，以产生对应于真实世界输入(例如，测试数据114)的输出(例如，结果118)。在测试阶段，模型参数以量化格式保存，且根据在量化过程期间确认的最佳量化位宽qb2量化中间结果(并且中间结果被提供给带有缩减模型的模型部署系统)。(更具体地)，如下面的示例过程所示：

//Testing Phase: To classify an unseen image I_{un}

$x_1=I_{un}$

For i = 1 to L:

[0086]

$y_i=(Q_{Non}(W_{opt,i}\times x_i+b_{opt,i}, qb2_{opt}))$

$x_{i+1}=y_i$

Return y_L

[0087] 上述模型比原始全精度训练模型(例如模型116')紧凑得多，并且计算使用整数而不是浮点值进行，这进一步缩减了计算占用并改善了计算速度。此外，能够利用某些硬件功能，以缩减的位宽和使用整数表示来进一步加速矩阵处理/计算。在一些实施例中，能够进一步约束位宽选择(例如，甚至只有位宽数)，以与在部署系统上所使用的硬件(例如，内存结构)更兼容。

[0088] 图5是根据一些实施例由模型生成系统(例如，模型生成系统102或200)实施的示例过程500的流程图。在一些实施例中，在机器学习系统100的服务器部件上实施示例过程500。

[0089] 过程500在计算设备上提供自适应位宽神经网络模型。在计算设备处，其中，该计算设备具有一个或多个处理器和内存：设备获取(502)包括多个层的第一神经网络模型(例如，训练后的全精度模型160'或剪枝后的全精度模型160")，其中，多个层中的每一层(例如，一个或多个卷积层、池化层、激活层等)具有对应参数集(例如，用于耦合该层及其下一个层的权重集，用于该层的网络偏差参数集)，并且，每一个参数以与第一神经网络模型的原始位宽(例如，32位或者其他硬件特有的位宽)对应的数据精度水平(例如，作为单精度浮点值)表示(例如，在计算设备的内存中，每一个参数占第一位数(例如，作为32位浮点数，占32位))。

[0090] 通过使用用于存储第一神经网络模型的不同层的对应参数集的对应缩减位宽，设备缩减(504)第一神经网络模型在计算设备(例如，在存储期间以及可选地在模型的部署期间两者)上的占用(例如，内存和计算成本)，其中：当第一神经网络模型的两个或者多个层中的每一层用对应于不同缩减位宽的不同量化程度表示时，使用验证数据集进行通过第一神经网络模型的前向传播的多次迭代，直到两个或更多层的对应响应统计满足预定信息丢失阈值为止(例如，通过本文中所描述的Jensen-Shannon散度测量)，从而确定对应缩减位宽的优选值(例如，使用本文中所描述的技术确认最佳位宽值)。

[0091] 设备生成(506)包括多个层的缩减的神经网络模型(例如,模型112),其中,多个层中的两个或更多的每一层包括对应量化参数集(例如,量化权重和偏差参数),且每一个量化参数利用通过多次迭代所确定的用于该层的对应缩减位宽的优选值表示。在一些实施例中,缩减的神经网络模型被部署在便携式电子设备上,其中,便携式电子设备根据缩减模型处理真实世界数据以生成预测结果,以及其中,在数据处理期间产生的中间结果被通过计算设备根据被提供给便携式电子设备的最佳缩减位宽量化。

[0092] 在一些实施例中,在缩减的神经网络模型(例如,模型112)的多个层的第一层(例如,i＝2)具有小于第一神经网络模型的原始位宽(例如,32位)的第一缩减位宽(例如,4位),在缩减的神经网络模型(例如,模型112)的多个层中的第二层(例如,i＝3)具有小于第一神经网络模型的原始位宽的第二缩减位宽(例如,6位),以及,在缩减的神经网络模型中,第一缩减位宽与第二缩减位宽不同。

[0093] 在一些实施例中,缩减第一神经网络的占用包括:对于两个或更多层中具有用对应于第一神经网络模型的原始位宽的数据精度水平表示的第一参数集(例如,权重集和一个或多个偏差)的第一层:当多个层的对应参数集以第一神经网络模型的原始位宽(例如,32位)表示时,随着验证数据集作为输入通过第一神经网络模型前向传播,计算设备收集第一层的激活值的对应基准统计分布(例如,statistics_i);当第一层的对应参数集用小于第一神经网络模型的原始位宽的第一缩减位宽(例如,$W_{q,i}$和$b_{q,i}$)表示时,随着验证数据集作为输入通过第一神经网络模型前向传播,计算设备收集第一层的激活值的对应修改统计分布(例如,$stat_{q,i}$);确定在第一层的激活值的对应修改统计分布和第一层的激活值的对应基准统计分布之间的预定散度(例如,Inf_tmp＝InformationLoss($stat_{q,I}$,statistics_i));以及,计算设备确认第一缩减位宽的最小值,对于该最小值,由于第一层的位宽的进一步缩减引起的预定散度的缩减低于预定阈值。

[0094] 在一些实施例中,以第一缩减位宽表示第一层的对应参数集包括对第一层的对应参数集执行非均匀量化(例如,对数量化$Q_{non}(\cdots)$),以生成第一层的第一量化参数集,以及,在通过第一层的每一次前向传播期间,基于第一层的激活值的基准统计分布,选择用于第一层的非均匀量化的最大边界值(例如,X_{max})。

[0095] 在一些实施例中,获得包括多个层的第一神经网络模型包括:在第一神经网络的训练期间:对于两个或更多层中具有用对应于第一神经网络模型的原始位宽的数据精度水平表示的第一参数集的第一层:根据在与第一层对应的第一权重集(例如,W_i)和第一权重集的整数部分(例如,$\lfloor W_i \rfloor$)之间的差异,获得对应于第一层(例如,层i)的整数正则化项(例如,RI_i)(例如,$RI_i = \frac{1}{2} \| W_i - \lfloor W_i \rfloor \|_2$);以及,在通过第一层的前向传播期间将整数正则化项(例如,RI_i)添加至偏差项(其中,8位均匀量化被应用到权重和偏差项),使得在通过第一层的逆向传播期间,梯度被更改以将第一参数集的值推向整数值。

[0096] 在一些实施例中,获得包括多个层的第一神经网络模型包括:在第一神经网络的训练期间:对于两个或更多层中具有用对应于第一神经网络模型的原始位宽的数据精度水平表示的第一参数集的第一层,在通过第一层的前向传播期间,利用小于第一神经网络模型的原始位宽的预定缩减位宽(例如,8位),对第一参数集执行均匀量化。

[0097] 在一些实施例中,获得包括多个层的第一神经网络模型包括:在第一神经网络的

训练期间:对于两个或更多层中具有用对应于第一神经网络模型的原始位宽的数据精度水平表示的第一参数集的第一层,在通过第一层的逆向传播期间,舍弃(forgoing)利用预定缩减位宽对第一参数集执行均匀量化。

[0098] 示例过程500仅仅涵盖本文中所描述的方法和技术中的一些方面。在该说明书的其他部分中提供其他细节和组合。为了简洁起见,这里不再重复或详尽列举这些细节和组合。

[0099] 应当理解,已经描述的操作的特定次序仅仅是示例,并且并不旨在指示所描述的次序是可以执行操作的唯一次序。本领域普通技术人员将认识到对本文中描述的操作进行重新排序的各种方式。通过在诸如通用处理器或专用芯片的信息处理设备中运行一个或多个功能模块,可选地实施在上文中描述的信息处理方法中的操作。

[0100] 为了解释的目的,已经参考特定实施例描述了前文。然而,以上说明性讨论并非旨在穷举本发明或将本发明限制为所公开的精确形式。鉴于以上教导,许多修改和变化是可能的。选择和描述实施例以便最佳解释本发明的原理和其实际应用,以由此使本领域技术人员能够最佳使用本发明和使用具有适合于所考虑的特定用途的不同修改的所述各种实施例。

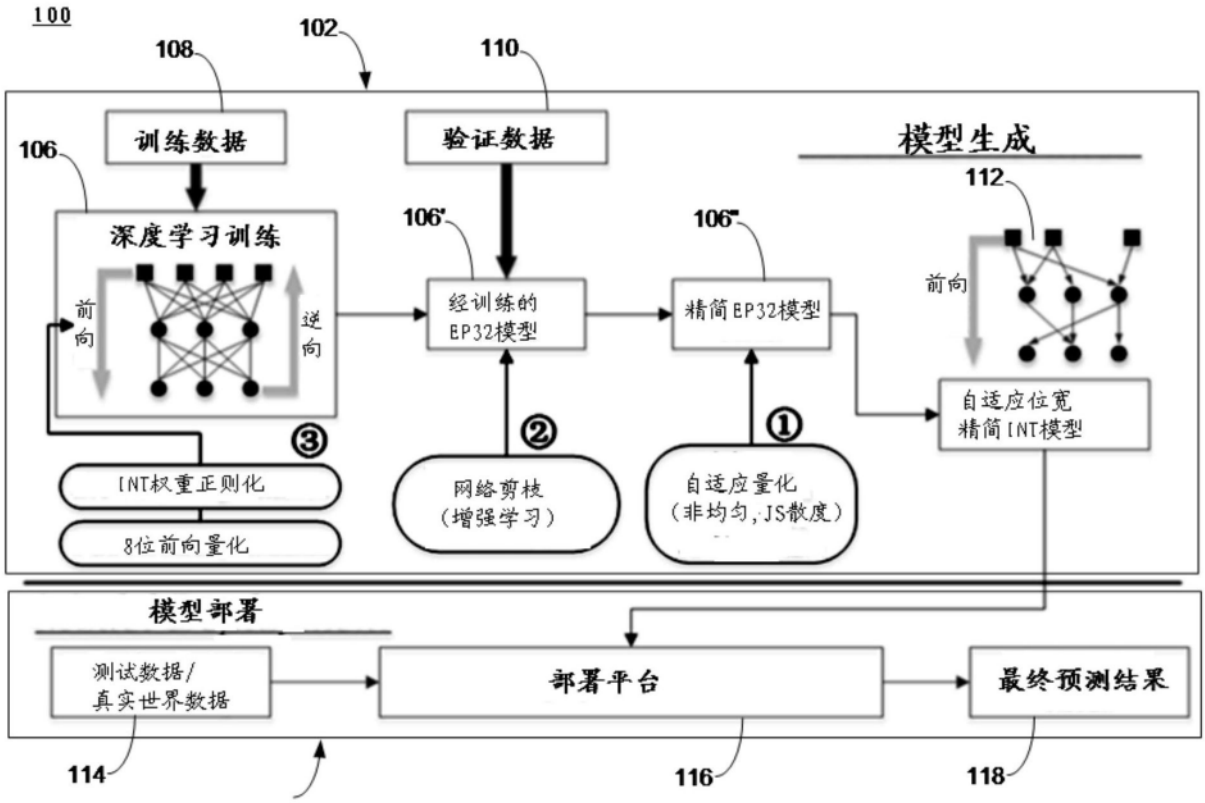

100
102
108
110
训练数据
验证数据
模型生成
106
深度学习训练
前向
逆向
106'
经训练的
EP32模型
106"
精简EP32模型
112
前向
自适应位宽
精简INT模型
③
INT权重正则化
8位前向量化
②
网络剪枝
(增强学习)
①
自适应量化
(非均匀,JS散度)
模型部署
测试数据/
真实世界数据
部署平台
最终预测结果
114
116
118

图1

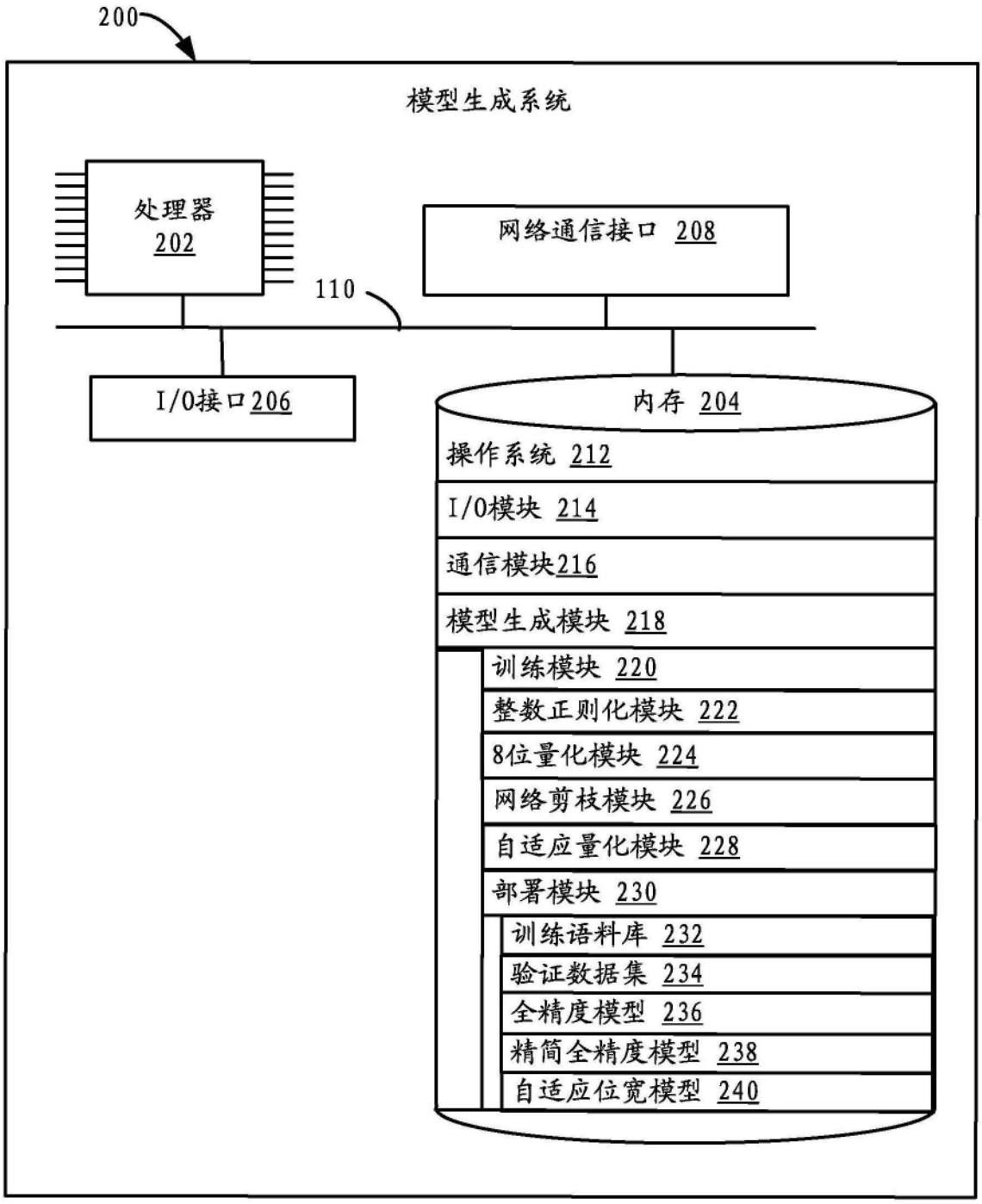

200
模型生成系统
处理器
202
网络通信接口 208
110
I/O接口206
内存 204
操作系统 212
I/O模块 214
通信模块216
模型生成模块 218
训练模块 220
整数正则化模块 222
8位量化模块 224
网络剪枝模块 226
自适应量化模块 228
部署模块 230
训练语料库 232
验证数据集 234
全精度模型 236
精简全精度模型 238
自适应位宽模型 240

图2

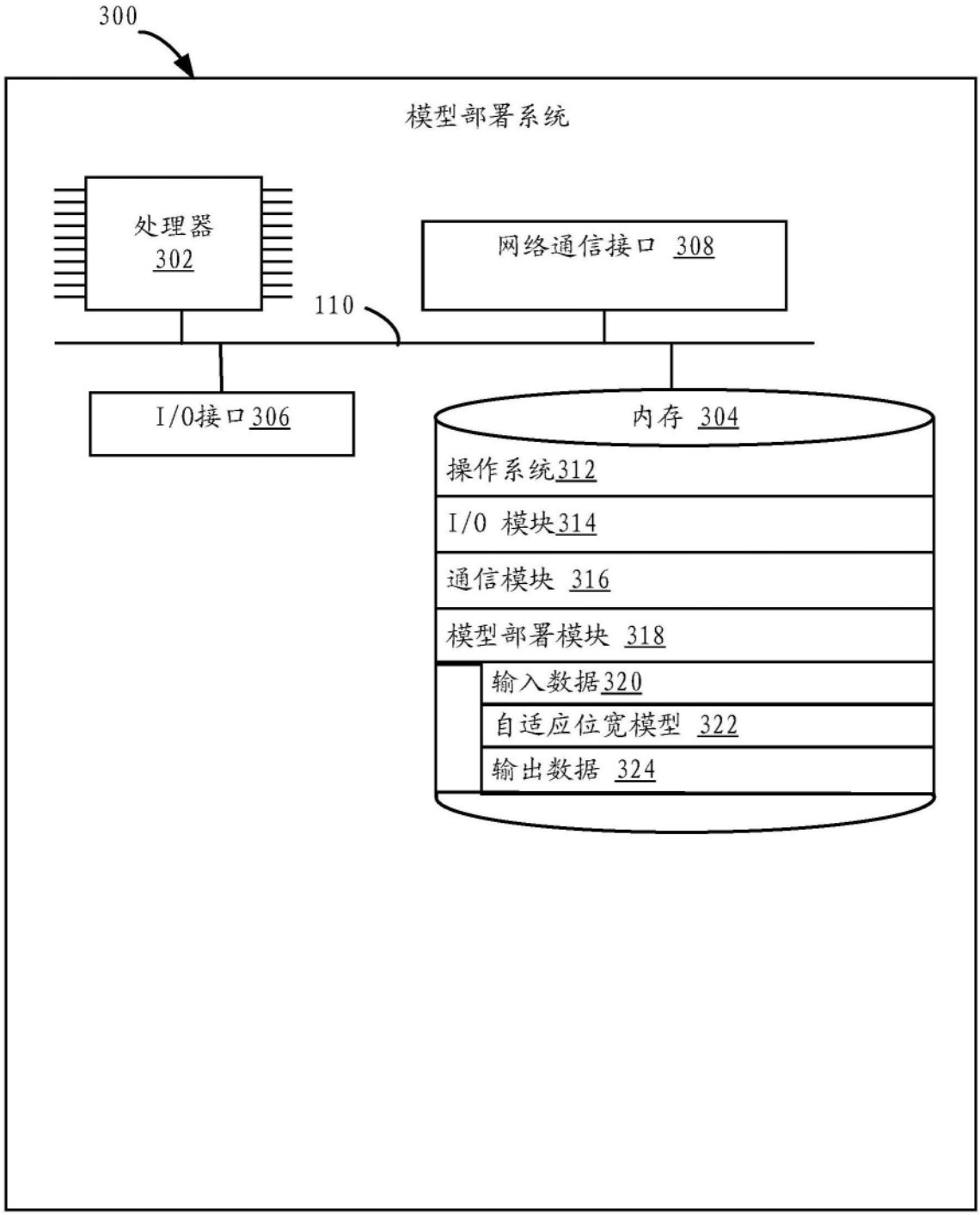

300
模型部署系统
处理器
302
网络通信接口 308
110
I/O接口306
内存 304
操作系统312
I/O 模块314
通信模块 316
模型部署模块 318
输入数据320
自适应位宽模型 322
输出数据 324

图3

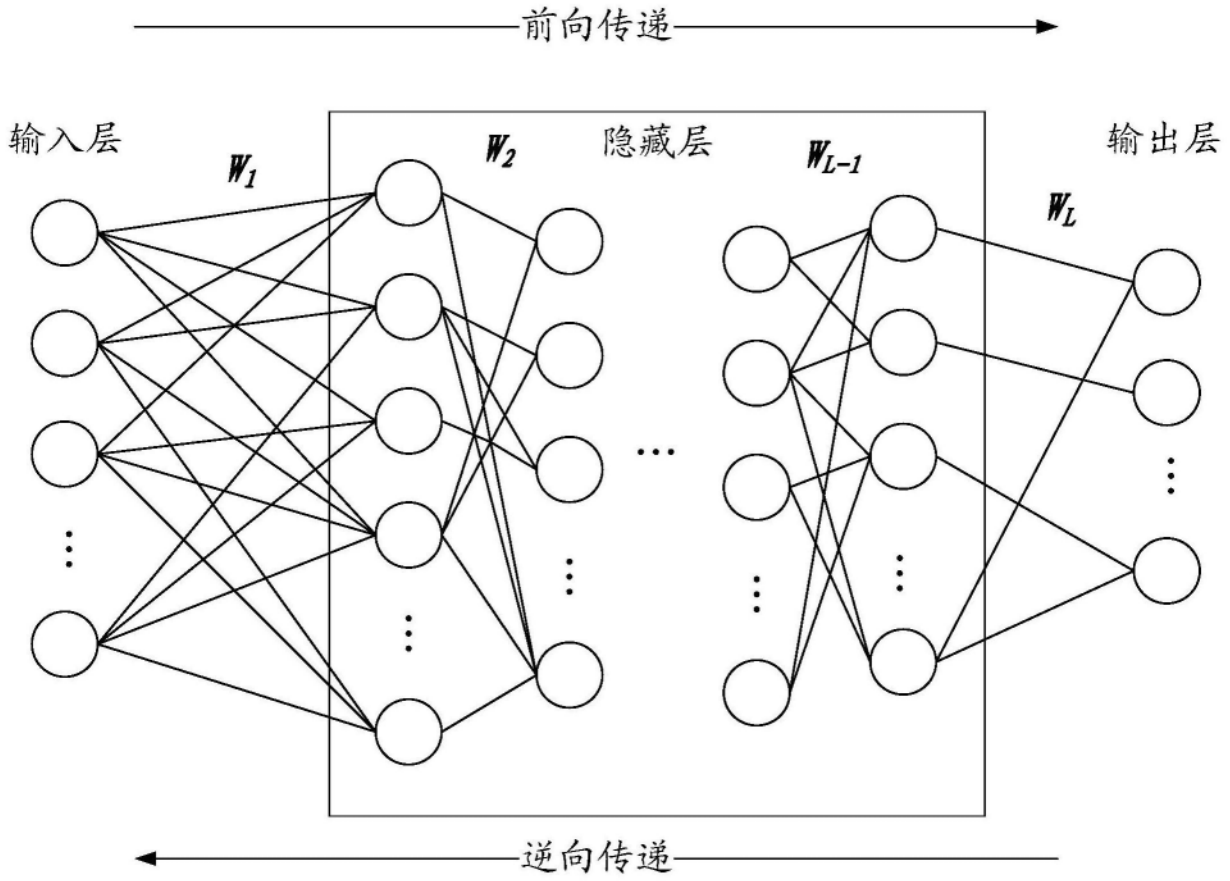


图4

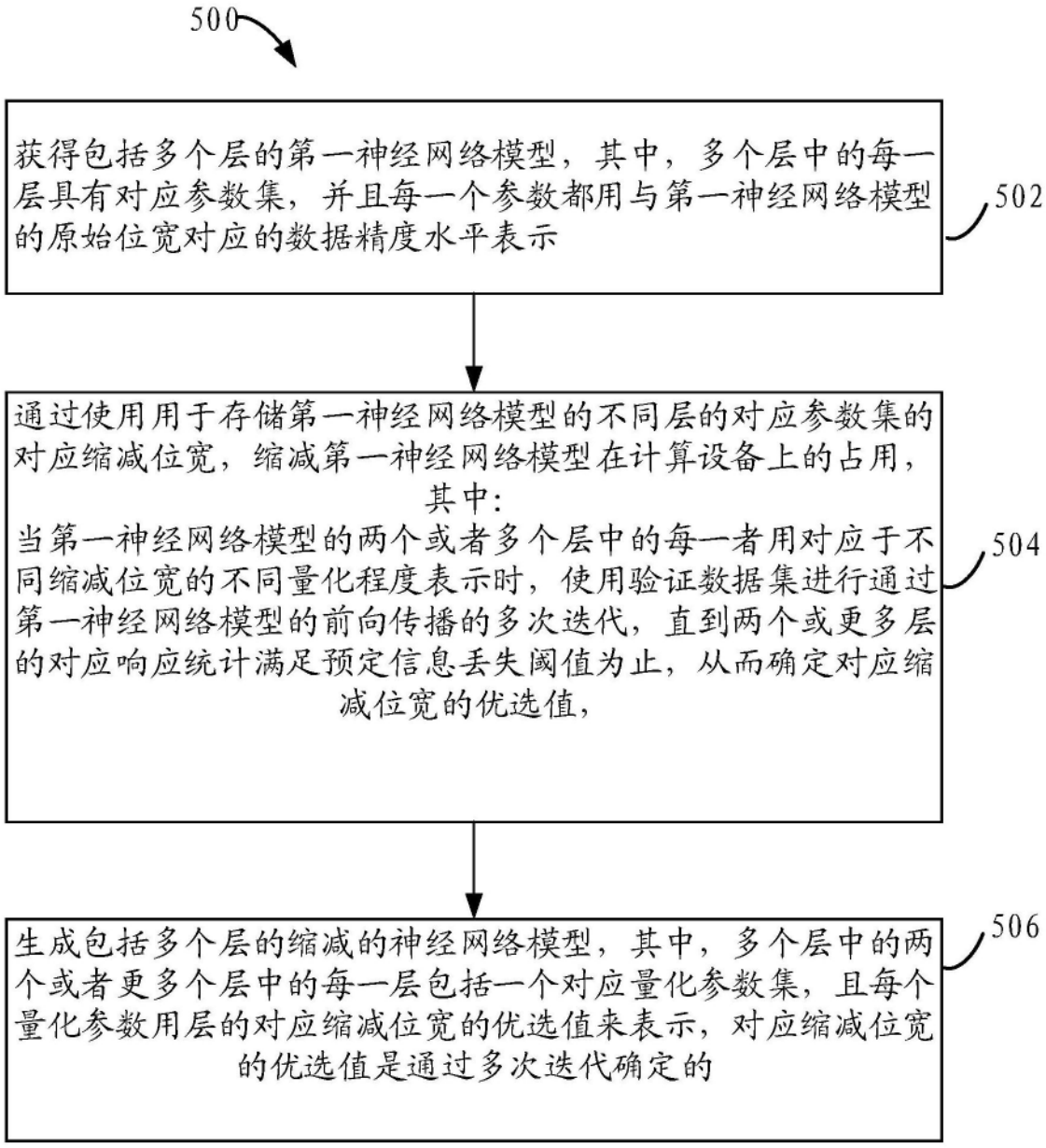

500
获得包括多个层的第一神经网络模型，其中，多个层中的每一层具有对应参数集，并且每一个参数都用与第一神经网络模型的原始位宽对应的数据精度水平表示
502
通过使用用于存储第一神经网络模型的不同层的对应参数集的对应缩减位宽，缩减第一神经网络模型在计算设备上的占用，其中：
当第一神经网络模型的两个或者多个层中的每一者用对应于不同缩减位宽的不同量化程度表示时，使用验证数据集进行通过第一神经网络模型的前向传播的多次迭代，直到两个或更多层的对应响应统计满足预定信息丢失阈值为止，从而确定对应缩减位宽的优选值，
504
生成包括多个层的缩减的神经网络模型，其中，多个层中的两个或者更多个层中的每一层包括一个对应量化参数集，且每个量化参数用层的对应缩减位宽的优选值来表示，对应缩减位宽的优选值是通过多次迭代确定的
506

图5