(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107004157 A
(43)申请公布日 2017.08.01

(21)申请号 201580065132.5

(22)申请日 2015.12.15

(30)优先权数据
62/106,608 2015.01.22 US
14/846,579 2015.09.04 US

(85)PCT国际申请进入国家阶段日
2017.05.31

(86)PCT国际申请的申请数据
PCT/US2015/065783 2015.12.15

(87)PCT国际申请的公布数据
WO2016/118257 EN 2016.07.28

(71)申请人 高通股份有限公司
地址 美国加利福尼亚州

(72)发明人 V·S·R·安纳普莱蒂
D·H·F·德克曼 D·J·朱利安

(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 周敏

(51)Int.Cl.
G06N 3/04(2006.01)
G06N 3/08(2006.01)

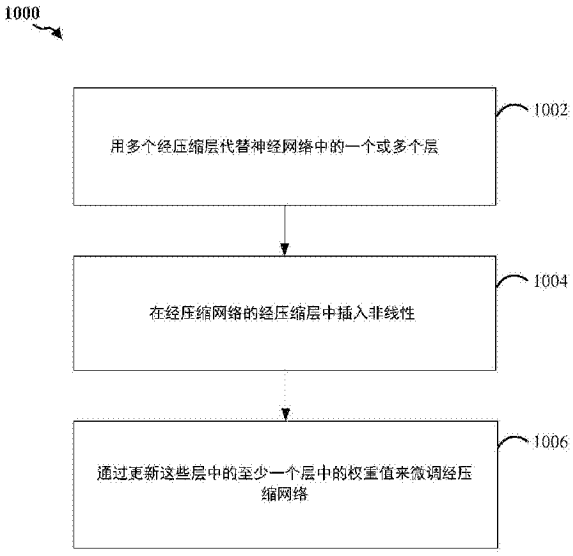
权利要求书3页 说明书18页 附图13页

(54)发明名称

模型压缩和微调

(57)摘要

压缩机器学习网络(诸如神经网络)包括用经压缩层代替神经网络中的一个层以产生经压缩网络。经压缩网络可通过更新经压缩层中的权重值来微调。



CN 107004157 A

1.一种压缩神经网络的方法,包括:

用多个经压缩层代替所述神经网络中的至少一个层以产生经压缩神经网络;

在所述经压缩网络的经压缩层之间插入非线性;以及

通过更新所述经压缩层中的至少一个经压缩层中的权重值来微调所述经压缩网络。

2.如权利要求1所述的方法,其特征在于,所述插入非线性包括向所述经压缩层的神经元应用非线性激活函数。

3.如权利要求2所述的方法,其特征在于,所述非线性激活函数是矫正器、绝对值函数、双曲正切函数或S形函数。

4.如权利要求1所述的方法,其特征在于,所述微调是通过更新所述经压缩神经网络中的所述权重值来执行的。

5.如权利要求4所述的方法,其特征在于,所述微调包括更新所述经压缩层的子集或者未压缩层的子集中的至少一者中的权重值。

6.如权利要求4所述的方法,其特征在于,所述微调是使用训练示例来执行的,所述训练示例包括用来训练未压缩网络的第一示例集合或新的示例集合中的至少一者。

7.如权利要求1所述的方法,其特征在于,进一步包括:

通过重复地应用压缩、插入非线性层、以及微调作为用于初始化较深神经网络的方法来初始化所述神经网络。

8.一种压缩神经网络的方法,包括:

用多个经压缩层代替所述神经网络中的至少一个层以产生经压缩神经网络,从而组合的经压缩层的感受野大小与未压缩层的感受野相匹配;以及

通过更新至少一个经压缩层中的权重值来微调所述经压缩网络。

9.如权利要求8所述的方法,其特征在于,未压缩层的内核大小等于所述感受野大小。

10.如权利要求8所述的方法,其特征在于,所述代替包括:用具有内核大小$k_{1x}\times k_{1y}$、$k_{2x}\times k_{2y}\cdots k_{Lx}\times k_{Ly}$的相同类型的多个经压缩层来代替所述神经网络中具有内核大小$k_x\times k_y$的至少一个层以产生所述经压缩网络,其中满足性质$(k_{1x}-1)+(k_{2x}-1)+\cdots=(k_x-1)$和$(k_{1y}-1)+(k_{2y}-1)+\cdots=(k_y-1)$。

11.如权利要求10所述的方法,其特征在于,具有内核大小$k_x\times k_y$的卷积层用分别具有内核大小1×1、$k_x\times k_y$和1×1的三个卷积层代替。

12.一种压缩神经网络的方法,包括:

用多个经压缩层代替所述神经网络中的至少一个层以产生经压缩神经网络;以及

通过应用交替最小化过程来确定经压缩层的权重矩阵。

13.如权利要求12所述的方法,其特征在于,进一步包括通过更新所述经压缩层中的至少一个经压缩层中的权重值来微调所述经压缩网络。

14.如权利要求13所述的方法,其特征在于,所述微调包括更新所述经压缩层的子集或者未压缩层的子集中的至少一者中的权重值。

15.如权利要求13所述的方法,其特征在于,所述微调是在多个阶段中执行的,其中在第一阶段中针对经压缩层的子集执行所述微调,并且在第二阶段中针对经压缩层和未压缩层的子集执行所述微调。

16.一种用于压缩神经网络的装置,包括:

存储器;以及

耦合至所述存储器的至少一个处理器,所述至少一个处理器被配置成:

用多个经压缩层代替所述神经网络中的至少一个层以产生经压缩神经网络;

在所述经压缩网络的经压缩层之间插入非线性;以及

通过更新至少一个经压缩层中的权重值来微调所述经压缩网络。

17.如权利要求16所述的装置,其特征在于,所述至少一个处理器被进一步配置成通过向所述经压缩层的神经元应用非线性激活函数来插入非线性。

18.如权利要求17所述的装置,其特征在于,所述非线性激活函数是矫正器、绝对值函数、双曲正切函数或S形函数。

19.如权利要求16所述的装置,其特征在于,所述至少一个处理器被进一步配置成通过更新所述经压缩神经网络中的所述权重值来执行所述微调。

20.如权利要求19所述的装置,其特征在于,所述至少一个处理器被进一步配置成通过更新经压缩层的子集或者未压缩层的子集中的至少一者中的权重值来执行所述微调。

21.如权利要求19所述的装置,其特征在于,所述至少一个处理器被进一步配置成通过使用训练示例来执行所述微调,所述训练示例包括用来训练未压缩网络的第一示例集合或新的示例集合中的至少一者。

22.如权利要求16所述的装置,其特征在于,所述至少一个处理器被进一步配置成通过重复地应用压缩、插入非线性层、以及微调作为用于初始化较深神经网络的方法来初始化所述神经网络。

23.一种用于压缩神经网络的装置,包括:

存储器;以及

耦合至所述存储器的至少一个处理器,所述至少一个处理器被配置成:

用多个经压缩层代替所述神经网络中的至少一个层以产生经压缩神经网络,从而组合的经压缩层的感受野大小与未压缩层的感受野大小相匹配;以及

通过更新至少一个经压缩层中的权重值来微调所述经压缩网络。

24.如权利要求23所述的装置,其特征在于,未压缩层的内核大小等于所述感受野大小。

25.如权利要求23所述的装置,其特征在于,所述至少一个处理器被进一步配置成:用具有内核大小$k_{1x}\times k_{1y}$、$k_{2x}\times k_{2y}\cdots k_{Lx}\times k_{Ly}$的相同类型的多个经压缩层来代替所述神经网络中具有内核大小$k_x\times k_y$的至少一个层以产生所述经压缩网络,其中满足性质$(k_{1x}-1)+(k_{2x}-1)+\cdots=(k_x-1)$和$(k_{1y}-1)+(k_{2y}-1)+\cdots=(k_y-1)$。

26.如权利要求25所述的装置,其特征在于,具有内核大小$k_x\times k_y$的卷积层用分别具有内核大小1×1、$k_x\times k_y$和1×1的三个卷积层代替。

27.一种用于压缩神经网络的装置,包括:

存储器;以及

耦合至所述存储器的至少一个处理器,所述至少一个处理器被配置成:

用多个经压缩层代替所述神经网络中的至少一个层以产生经压缩神经网络;以及

通过应用交替最小化过程来确定经压缩层的权重矩阵。

28.如权利要求27所述的装置,其特征在于,所述至少一个处理器被进一步配置成通过

更新所述经压缩层中的至少一个经压缩层中的权重值来微调所述经压缩网络。

29.如权利要求28所述的装置，其特征在于，所述至少一个处理器被配置成通过更新所述经压缩层的子集或者未压缩层的子集中的至少一者中的权重值来执行所述微调。

30.如权利要求28所述的装置，其特征在于，所述至少一个处理器被进一步配置成在多个阶段中执行所述微调，其中在第一阶段中针对经压缩层的子集执行所述微调，并且在第二阶段中针对经压缩层和未压缩层的子集执行所述微调。

31.一种用于压缩神经网络的设备，包括：

用于用多个经压缩层代替所述神经网络中的至少一个层以产生经压缩神经网络的装置；

用于在所述经压缩网络的经压缩层之间插入非线性的装置；以及

用于通过更新所述经压缩层中的至少一个经压缩层中的权重值来微调所述经压缩网络的装置。

32.一种用于压缩神经网络的设备，包括：

用于用多个经压缩层代替所述神经网络中的至少一个层以产生经压缩神经网络，从而组合的经压缩层的感受野大小与未压缩层的感受野相匹配的装置；以及

用于通过更新至少一个经压缩层中的权重值来微调所述经压缩网络的装置。

33.一种用于压缩神经网络的设备，包括：

用于用多个经压缩层代替所述神经网络中的至少一个层以产生经压缩神经网络的装置；以及

用于通过应用交替最小化过程来确定经压缩层的权重矩阵的装置。

34.一种其上编码有用于压缩神经网络的程序代码的非瞬态计算机可读介质，所述程序代码被处理器执行并且包括：

用于用多个经压缩层代替所述神经网络中的至少一个层以产生经压缩神经网络的程序代码；

用于在所述经压缩网络的经压缩层之间插入非线性的程序代码；以及

用于通过更新至少一个经压缩层中的权重值来微调所述经压缩网络的程序代码。

35.一种其上编码有用于压缩神经网络的程序代码的非瞬态计算机可读介质，所述程序代码被处理器执行并且包括：

用于用多个经压缩层代替所述神经网络中的至少一个层以产生经压缩神经网络，从而组合的经压缩层的感受野大小与未压缩层的感受野大小相匹配的程序代码；以及

用于通过更新至少一个经压缩层中的权重值来微调所述经压缩网络的程序代码。

36.一种其上编码有用于压缩神经网络的程序代码的非瞬态计算机可读介质，所述程序代码被处理器执行并且包括：

用于用多个经压缩层代替所述神经网络中的至少一个层以产生经压缩神经网络的程序代码；以及

用于通过应用交替最小化过程来确定经压缩层的权重矩阵的程序代码。

模型压缩和微调

[0001] 相关申请的交叉引用

[0002] 本申请要求于2015年1月22日提交的题为“MODEL COMPRESSION AND FINE-TUNING(模型压缩和微调)”的美国临时专利申请No.62/106,608的权益,其公开内容通过援引全部明确纳入于此。

[0003] 背景

[0004] 领域

[0005] 本公开的某些方面一般涉及神经系统工程,并且尤其涉及用于压缩神经网络的系统和方法。

背景技术

[0006] 可包括一群互连的人工神经元(例如,神经元模型)的人工神经网络是一种计算设备或者表示将由计算设备执行的方法。

[0007] 卷积神经网络是一种前馈人工神经网络。卷积神经网络可包括神经元集合,其中每个神经元具有感受野并且合而铺覆一个输入空间。卷积神经网络(CNN)具有众多应用。具体地,CNN已被广泛使用于模式识别和分类领域。

[0008] 深度学习架构(诸如,深度置信网络和深度卷积网络)是分层神经网络架构,其中第一层神经元的输出变成第二层神经元的输入,第二层神经元的输出变成第三层神经元的输入,以此类推。深度神经网络可被训练以识别特征阶层并且因此它们被越来越多地使用于对象识别应用。类似于卷积神经网络,这些深度学习架构中的计算可分布在处理节点群体上,其可被配置在一个或多个计算链中。这些多层架构可每次训练一层并可使用反向传播来进行微调。

[0009] 其他模型也可用于对象识别。例如,支持向量机(SVM)是可被应用于分类的学习工具。支持向量机包括分类数据的分离超平面(例如,决策边界)。该超平面由监督式学习来定义。期望的超平面增加训练数据的余量。换言之,该超平面应该具有到训练示例的最大的最小距离。

[0010] 尽管这些解决方案在数个分类基准上取得了优异的结果,但它们的计算复杂度可能极其高。另外,模型的训练可能是有挑战性的。

[0011] 概述

[0012] 在一个方面,公开了一种压缩机器学习网络(诸如神经网络)的方法。该方法包括用多个经压缩层代替神经网络中的至少一个层以产生经压缩神经网络。该方法还包括在经压缩网络的经压缩层之间插入非线性。进一步,该方法包括通过更新这些经压缩层中的至少一个经压缩层中的权重值来微调经压缩网络。

[0013] 另一方面公开了一种用于压缩机器学习网络(诸如神经网络)的设备。该设备包括用于用多个经压缩层代替神经网络中的至少一个层以产生经压缩神经网络的装置。该设备还包括用于在经压缩网络的经压缩层之间插入非线性的装置。进一步,该设备包括用于通过更新这些经压缩层中的至少一个经压缩层中的权重值来微调经压缩网络的装置。

[0014] 另一方面公开了一种用于压缩机器学习网络(诸如神经网络)的装置。该装置包括存储器以及耦合至该存储器的至少一个处理器。该处理器被配置成用多个经压缩层代替神经网络中的至少一个层以产生经压缩神经网络。该处理器还被配置成在经压缩网络的经压缩层之间插入非线性。进一步,该处理器还被配置成通过更新至少一个经压缩层中的权重值来微调经压缩网络。

[0015] 另一方面公开了一种非瞬态计算机可读介质。该计算机可读介质上记录有用于压缩机器学习网络(诸如神经网络)的非瞬态程序代码。该程序代码在由处理器执行时使处理器用多个经压缩层代替神经网络中的至少一个层以产生经压缩神经网络。该程序代码还使处理器在经压缩网络的经压缩层之间插入非线性。进一步,该程序代码还使处理器通过更新至少一个经压缩层中的权重值来微调经压缩网络。

[0016] 在另一方面,公开了一种用于压缩机器学习网络(诸如神经网络)的方法。该方法包括用多个经压缩层代替神经网络中的至少一个层以产生经压缩神经网络,从而组合的经压缩层的感受野大小与未压缩层的感受野相匹配。该方法还包括通过更新这些经压缩层中的至少一个经压缩层中的权重值来微调经压缩网络。

[0017] 另一方面公开了一种用于压缩机器学习网络(诸如神经网络)的设备。该设备包括用于用多个经压缩层代替神经网络中的至少一个层以产生经压缩神经网络,从而组合的经压缩层的感受野大小与未压缩层的感受野相匹配的装置。该设备还包括用于通过更新这些经压缩层中的至少一个经压缩层中的权重值来微调经压缩网络的装置。

[0018] 另一方面公开了一种用于压缩机器学习网络(诸如神经网络)的装置。该装置包括存储器以及耦合至该存储器的至少一个处理器。该处理器被配置成用多个经压缩层代替神经网络中的至少一个层以产生经压缩神经网络,从而组合的经压缩层的感受野大小与未压缩层的感受野相匹配。该处理器还被配置成通过更新至少一个经压缩层中的权重值来微调经压缩网络。

[0019] 另一方面公开了一种非瞬态计算机可读介质。该计算机可读介质上记录有用于压缩机器学习网络(诸如神经网络)的非瞬态程序代码。该程序代码在由处理器执行时使处理器用多个经压缩层代替神经网络中的至少一个层以产生经压缩神经网络,从而组合的经压缩层的感受野大小与未压缩层的感受野相匹配。该程序代码还使处理器通过更新至少一个经压缩层中的权重值来微调经压缩网络。

[0020] 在另一方面,公开了一种压缩机器学习网络(诸如神经网络)的方法。该方法包括用多个经压缩层代替神经网络中的至少一个层以产生经压缩神经网络。该方法还包括通过应用交替最小化过程来确定经压缩层的权重矩阵。

[0021] 另一方面公开了一种用于压缩机器学习网络(诸如神经网络)的设备。该设备包括用于用多个经压缩层代替神经网络中的至少一个层以产生经压缩神经网络的装置。该设备还包括用于通过应用交替最小化过程来确定经压缩层的权重矩阵的装置。

[0022] 另一方面公开了一种用于压缩机器学习网络(诸如神经网络)的装置。该装置包括存储器以及耦合至该存储器的至少一个处理器。该处理器被配置成用多个经压缩层代替神经网络中的至少一个层以产生经压缩神经网络。该处理器还被配置成通过应用交替最小化过程来确定经压缩层的权重矩阵。

[0023] 另一方面公开了一种非瞬态计算机可读介质。该计算机可读介质上记录有用于压

缩机器学习网络(诸如神经网络)的程序代码。该程序代码在由处理器执行时使处理器用多个经压缩层代替神经网络中的至少一个层以产生经压缩神经网络。该程序代码还使处理器通过应用交替最小化过程来确定经压缩层的权重矩阵。

[0024] 本公开的附加特征和优点将在下文描述。本领域技术人员应该领会,本公开可容易地被用作修改或设计用于实施与本公开相同的目的的其他结构的基础。本领域技术人员还应认识到,这样的等效构造并不脱离所附权利要求中所阐述的本公开的教导。被认为是本公开的特性的新颖特征在其组织和操作方法两方面连同进一步的目的和优点在结合附图来考虑以下描述时将被更好地理解。然而,要清楚理解的是,提供每一幅附图均仅用于解说和描述目的,且无意作为对本公开的限定的定义。

[0025] 附图简述

[0026] 在结合附图理解下面阐述的详细描述时,本公开的特征、本质和优点将变得更加明显,在附图中,相同附图标记始终作相应标识。

[0027] 图1解说了根据本公开的一些方面的使用片上系统(SOC)(包括通用处理器)来设计神经网络的示例实现。

[0028] 图2解说了根据本公开的各方面的系统的示例实现。

[0029] 图3A是解说根据本公开的各方面的神经网络的示图。

[0030] 图3B是解说根据本公开的各方面的示例性深度卷积网络(DCN)的框图。

[0031] 图4A是解说根据本公开的各方面的可将人工智能(AI)功能模块化的示例性软件架构的框图。

[0032] 图4B是解说根据本公开的各方面的智能手机上的AI应用的运行时操作的框图。

[0033] 图5A-B和6A-B是解说根据本公开的各方面的全连接层和经压缩全连接层的框图。

[0034] 图7是解说根据本公开的各方面的示例性卷积层的框图。

[0035] 图8A-B和9A-B解说了根据本公开的各方面的卷积层的示例压缩。

[0036] 图10-13是解说根据本公开的各方面的用于压缩神经网络的方法的流程图。

[0037] 详细描述

[0038] 以下结合附图阐述的详细描述旨在作为各种配置的描述,而无意表示可实践本文中所描述的概念的仅有的配置。本详细描述包括具体细节以便提供对各种概念的透彻理解。然而,对于本领域技术人员将显而易见的是,没有这些具体细节也可实践这些概念。在一些实例中,以框图形式示出众所周知的结构和组件以避免湮没此类概念。

[0039] 基于本教导,本领域技术人员应领会,本公开的范围旨在覆盖本公开的任何方面,不论其是与本公开的任何其他方面相独立地还是组合地实现的。例如,可以使用所阐述的任何数目的方面来实现装置或实践方法。另外,本公开的范围旨在覆盖使用作为所阐述的本公开的各个方面的补充或者与之不同的其他结构、功能性、或者结构及功能性来实践的此类装置或方法。应当理解,所披露的本公开的任何方面可由权利要求的一个或多个元素来实施。

[0040] 措辞“示例性”在本文中用于表示“用作示例、实例或解说”。本文中描述为“示例性”的任何方面不必被解释为优于或胜过其他方面。

[0041] 尽管本文描述了特定方面,但这些方面的众多变体和置换落在本公开的范围之内。虽然提到了优选方面的一些益处和优点,但本公开的范围并非旨在被限定于特定益处、

用途或目标。相反，本公开的各方面旨在能宽泛地应用于不同的技术、系统配置、网络和协议，其中一些作为示例在附图以及以下对优选方面的描述中解说。详细描述和附图仅仅解说本公开而非限定本公开，本公开的范围由所附权利要求及其等效技术方案来定义。

[0042] 模型压缩和微调

[0043] 深度神经网络执行若干人工智能任务中的现有技术，诸如图像/视频分类、语音识别、面部识别等。神经网络模型从大型训练示例数据库进行训练，并且通常而言，较大模型趋向于达成较佳的性能。为了在移动设备(诸如智能电话、机器人和汽车)上部署这些神经网络模型的目的，期望尽可能地减小或最小化计算复杂度、存储器占用和功耗。这些也是云应用(诸如，每天处理数百万图像/视频的数据中心)的期望属性。

[0044] 本公开的各方面涉及一种压缩神经网络模型的方法。经压缩模型具有显著更少数目的参数，并且由此具有较小的存储器占用。另外，经压缩模型具有显著更少的操作，并且由此导致较快的推断运行时。

[0045] 卷积神经网络模型可被划分成层序列。每一层可变换从网络中的一个或多个在前层接收的输入并且可产生可被提供给网络的后续层的输出。例如，卷积神经网络可包括全连接层、卷积层、局部连接层以及其它层。这些不同层中的每一层可执行不同类型的变换。

[0046] 根据本公开的各方面，全连接层、卷积层和局部连接层可被压缩。所有这些层可对输入值应用线性变换，其中该线性变换通过权重值来参数化。这些权重值可被表示为二维矩阵或较高维张量。线性变换结构对于不同层类型可以是不同的。

[0047] 在一些方面，层可通过用相同类型的多个层代替它来被压缩。例如，卷积层可通过用多个卷积层代替它来被压缩，并且全连接层可通过用多个全连接层代替它来被压缩。经压缩层中的神经元可被配置有相同激活函数。尽管一层可用多个经压缩层来代替，但所有经压缩层一起的总复杂度(和存储器占用)可小于单个未压缩层。

[0048] 然而，模型压缩可能以神经网络的性能下降为代价。例如，如果神经网络被用于分类问题，则分类准确度可能因压缩而下降。此准确度下降限制了压缩程度，因为较高压缩导致较高的准确度下降。为了对抗准确度下降，在一些方面，经压缩模型可使用训练示例来微调。经压缩模型的微调可收回准确度下降并且由此可在没有过多分类准确度下降的情况下产生改善的压缩。经压缩模型的微调可在逐层基础上进行。

[0049] 因此，本公开的各方面提供了其中神经网络的层用更多此类层来代替的压缩技术。因此，结果所得的经压缩模型是另一卷积神经网络模型。这是显著的优势，因为可避免开发新的且明确的用于实现经压缩模型的技术。也就是说，如果任何神经网络库与原始未压缩模型兼容，则它也将与经压缩模型兼容。此外，由于结果所得的经压缩模型是卷积神经网络，因此可执行全栈微调(例如，反向传播)，而非将微调限于特定层。

[0050] 神经网络由若干层构成。每一层取来自一个或多个先前层的激活向量作为输入，对组合输入向量应用线性/非线性变换，以及输出将由后续层使用的激活向量。一些层用权重来参数化，而一些层并非如此。本公开的各方面关注于以下权重层：1)全连接层，2)卷积层，以及3)局部连接层。所有三个层都执行线性变换，但在输出神经元如何连接至输入神经元方面有所不同。

[0051] 图1解说了根据本公开的某些方面使用片上系统(SOC)100进行前述压缩神经网络的示例实现，SOC 100可包括通用处理器(CPU)或多核通用处理器(CPU)102。变量(例如，神

经信号和突触权重)、与计算设备相关联的系统参数(例如,带有权重的神经网络)、延迟、频率槽信息、以及任务信息可被存储在与神经处理单元(NPU)108相关联的存储器块、与CPU 102相关联的存储器块、与图形处理单元(GPU)104相关联的存储器块、与数字信号处理器(DSP)106相关联的存储器块、专用存储器块118中,或可跨多个块分布。在通用处理器102处执行的指令可从与CPU 102相关联的程序存储器加载或可从专用存储器块118加载。

[0052] SOC 100还可包括为具体功能定制的额外处理块(诸如GPU 104、DSP 106、连通性块110)以及例如可检测和识别姿势的多媒体处理器112,这些具体功能可包括第四代长期演进(4G LTE)连通性、无执照Wi-Fi连通性、USB连通性、蓝牙连通性等。在一种实现中,NPU实现在CPU、DSP、和/或GPU中。SOC 100还可包括传感器处理器114、图像信号处理器(ISP)、和/或导航120(其可包括全球定位系统)。

[0053] SOC可基于ARM指令集。在本公开的一方面,加载到通用处理器102中的指令可包括用于以下操作的代码:用多个经压缩层代替机器学习网络(诸如神经网络)中的至少一个层以产生经压缩网络,在经压缩网络的经压缩层之间插入非线性和/或通过更新这些经压缩层中的至少一个经压缩层中的权重值来微调经压缩网络。

[0054] 在本公开的另一方面,加载到通用处理器102中的指令可包括用于以下操作的代码:用多个经压缩层代替机器学习网络(诸如神经网络)中的至少一个层以产生经压缩网络,从而组合的经压缩层的感受野大小与未压缩层的感受野相匹配。还可存在用于通过更新这些经压缩层中的至少一个经压缩层中的权重值来微调经压缩网络的代码。

[0055] 在本公开的又一方面,加载到通用处理器102中的指令可包括用于以下操作的代码:用多个经压缩层代替机器学习网络(诸如神经网络)中的至少一个层以产生经压缩网络。还可存在用于通过应用交替最小化过程来确定经压缩层的权重矩阵的代码。

[0056] 图2解说了根据本公开的某些方面的系统200的示例实现。如图2中所解说的,系统200可具有多个可执行本文所描述的方法的各种操作的局部处理单元202。每个局部处理单元202可包括局部状态存储器204和可存储神经网络的参数的局部参数存储器206。另外,局部处理单元202可具有用于存储局部模型程序的局部(神经元)模型程序(LMP)存储器208、用于存储局部学习程序的局部学习程序(LLP)存储器210、以及局部连接存储器212。此外,如图2中所解说的,每个局部处理单元202可与用于为该局部处理单元的各局部存储器提供配置的配置处理器单元214对接,并且与提供各局部处理单元202之间的路由的路由连接处理单元216对接。

[0057] 深度学习架构可通过学习在每一层中以逐次更高的抽象程度来表示输入、藉此构建输入数据的有用特征表示来执行对象识别任务。以此方式,深度学习解决了传统机器学习的主要瓶颈。在深度学习出现之前,用于对象识别问题的机器学习办法可能严重依赖人类工程设计的特征,或许与浅分类器相结合。浅分类器可以是两类线性分类器,例如,其中可将特征向量分量的加权和与阈值作比较以预测输入属于哪一类。人类工程设计的特征可以是由拥有领域专业知识的工程师针对具体问题领域定制的模版或内核。相反,深度学习架构可学习去表示与人类工程师可能会设计的相似的特征,但它是通过训练来学习的。另外,深度网络可以学习去表示和识别人类可能还没有考虑过的新类型的特征。

[0058] 深度学习架构可以学习特征阶层。例如,如果向第一层呈递视觉数据,则第一层可学习去识别输入流中的简单特征(诸如边)。如果向第一层呈递听觉数据,则第一层可学习

去识别特定频率中的频谱功率。取第一层的输出作为输入的第二层可以学习去识别特征组合，诸如对于视觉数据去识别简单形状或对于听觉数据去识别声音组合。更高层可学习去表示视觉数据中的复杂形状或听觉数据中的词语。再高层可学习去识别常见视觉对象或口语短语。

[0059] 深度学习架构在被应用于具有自然阶层结构的问题时可能表现特别好。例如，机动交通工具的分类可受益于首先学习去识别轮子、挡风玻璃、以及其他特征。这些特征可在更高层以不同方式被组合以识别轿车、卡车和飞机。

[0060] 神经网络可被设计成具有各种连通性模式。在前馈网络中，信息从较低层被传递到较高层，其中给定层中的每个神经元向更高层中的神经元进行传达。如上所述，可在前馈网络的相继层中构建阶层式表示。神经网络还可具有回流或反馈(也被称为自顶向下(top-down))连接。在回流连接中，来自给定层中的神经元的输出被传达给相同层中的另一神经元。回流架构可有助于识别跨越不止一个按顺序递送给该神经网络的输入数据组块的模式。从给定层中的神经元到较低层中的神经元的连接被称为反馈(或自顶向下)连接。当高层级概念的识别可辅助辨别输入的特定低层级特征时，具有许多反馈连接的网络可能是有助益的。

[0061] 参照图3A，神经网络的各层之间的连接可以是全连接的(302)或局部连接的(304)。在全连接网络302中，第一层中的神经元可将它的输出传达给第二层中的每个神经元，从而第二层中的每个神经元将从第一层中的每个神经元接收输入。替换地，在局部连接网络304中，第一层中的神经元可连接至第二层中有限数目的神经元。卷积网络306可以是局部连接的，并且被进一步配置成使得与针对第二层中每个神经元的输入相关联的连接强度被共享(例如，308)。更一般化地，网络的局部连接层可被配置成使得一层中的每个神经元将具有相同或相似的连通性模式，但其连接强度可具有不同的值(例如，310、312、314和316)。局部连接的连通性模式可能在更高层中产生空间上相异的感受野，这是由于给定区域中的更高层神经元可接收到通过训练被调谐为到网络的总输入的受限部分的性质的输入。

[0062] 局部连接的神经网络可能非常适合于其中输入的空间位置有意义的问题。例如，被设计成识别来自车载相机的视觉特征的网络300可发展具有不同性质的高层神经元，这取决于它们与图像下部关联还是与图像上部关联。例如，与图像下部相关联的神经元可学习去识别车道标记，而与图像上部相关联的神经元可学习去识别交通信号灯、交通标志等。

[0063] DCN可以用受监督式学习来训练。在训练期间，DCN可被呈递图像326(诸如限速标志的经裁剪图像)，并且可随后计算“前向传递(forward pass)”以产生输出322。输出322可以是对应于特征(诸如“标志”、“60”、和“100”)的值向量。网络设计者可能希望DCN在输出特征向量中针对其中一些神经元输出高得分，例如对应于经训练的网络300的输出322中所示的“标志”和“60”的那些神经元。在训练之前，DCN产生的输出很可能是不正确的，并且由此可计算实际输出与目标输出之间的误差。DCN的权重可随后被调整以使得DCN的输出得分与目标更紧密地对准。

[0064] 为了调整权重，学习算法可为权重计算梯度向量。该梯度可指示如果权重被略微调整则误差将增加或减少的量。在顶层，该梯度可直接对应于连接倒数第二层中的活化神经元与输出层中的神经元的权重的值。在较低层中，该梯度可取决于权重的值以及所计算

出的较高层的误差梯度。权重可随后被调整以减小误差。这种调整权重的方式可被称为“反向传播”，因为其涉及在神经网络中的“反向传递(backward pass)”。

[0065] 在实践中，权重的误差梯度可能是在少量示例上计算的，从而计算出的梯度近似于真实误差梯度。这种近似方法可被称为随机梯度下降法。随机梯度下降法可被重复，直到整个系统可达成的误差率已停止下降或直到误差率已达到目标水平。

[0066] 在学习之后，DCN可被呈递新图像326并且在网络中的前向传递可产生输出322，其可被认为是该DCN的推断或预测。

[0067] 深度置信网络(DBN)是包括多层隐藏节点的概率性模型。DBN可被用于提取训练数据集的阶层式表示。DBN可通过堆叠多层受限波尔兹曼机(RBM)来获得。RBM是一类可在输入集上学习概率分布的人工神经网络。由于RBM可在没有关于每个输入应该被分类到哪个类的信息的情况下学习概率分布的，因此RBM经常被用于无监督式学习。使用混合无监督式和受监督式范式，DBN的底部RBM可按无监督方式被训练并且可以用作特征提取器，而顶部RBM可按受监督方式(在来自先前层的输入和目标类的联合分布上)被训练并且可用作分类器。

[0068] 深度卷积网络(DCN)是卷积网络的网络，其配置有附加的池化和归一化层。DCN已在许多任务上达成现有最先进的性能。DCN可使用受监督式学习来训练，其中输入和输出目标两者对于许多典范是已知的并被用于通过使用梯度下降法来修改网络的权重。

[0069] DCN可以是前馈网络。另外，如上所述，从DCN的第一层中的神经元到下一更高层中的神经元群的连接跨第一层中的神经元被共享。DCN的前馈和共享连接可被利用于进行快速处理。DCN的计算负担可比例如类似大小的包括回流或反馈连接的神经网络小得多。

[0070] 卷积网络的每一层的处理可被认为是空间不变模版或基础投影。如果输入首先被分解成多个通道，诸如彩色图像的红色、绿色和蓝色通道，那么在该输入上训练的卷积网络可被认为是三维的，其具有沿着该图像的轴的两个空间维度以及捕捉颜色信息的第三个维度。卷积连接的输出可被认为在后续层318和320中形成特征图，该特征图(例如，320)中的每个元素从先前层(例如，318)中一定范围的神经元以及从该多个通道中的每一个通道接收输入。特征图中的值可以用非线性(诸如矫正)max(0,x)进一步处理。来自毗邻神经元的值可被进一步池化(这对应于降采样)并可提供附加的局部不变性以及维度缩减。还可通过特征图中神经元之间的侧向抑制来应用归一化，其对应于白化。

[0071] 深度学习架构的性能可随着有更多被标记的数据点变为可用或随着计算能力提高而提高。现代深度神经网络用比仅仅十五年前可供典型研究者使用的计算资源多数千倍的计算资源来例行地训练。新的架构和训练范式可进一步推升深度学习的性能。经矫正的线性单元可减少被称为梯度消失的训练问题。新的训练技术可减少过度拟合(over-fitting)并因此使更大的模型能够达成更好的推广。封装技术可抽象出给定的感受野中的数据并进一步提升总体性能。

[0072] 图3B是解说示例性深度卷积网络350的框图。深度卷积网络350可包括多个基于连通性和权重共享的不同类型的层。如图3B所示，该示例性深度卷积网络350可包括多个卷积块(例如，C1和C2)。每个卷积块可配置有卷积层、归一化层(LNorm)、和池化层。卷积层可包括一个或多个卷积滤波器，其可被应用于输入数据以生成特征图。尽管仅示出了两个卷积块，但本公开不限于此，相反，根据设计偏好，任何数目的卷积块可被包括在深度卷积网络350中。归一化层可被用于对卷积滤波器的输出进行归一化。例如，归一化层可提供白化或

侧向抑制。池化层可提供在空间上的降采样聚集以实现局部不变性和维度缩减。

[0073] 例如，深度卷积网络的平行滤波器组可任选地基于ARM指令集被加载到SOC 100的CPU 102或GPU 104上以达成高性能和低功耗。在替换实施例中，平行滤波器组可被加载到SOC 100的DSP 106或ISP 116上。另外，DCN可访问其他可存在于SOC上的处理块，诸如专用于传感器114和导航120的处理块。

[0074] 深度卷积网络350还可包括一个或多个全连接层(例如，FC1和FC2)。深度卷积网络350可进一步包括逻辑回归(LR)层。深度卷积网络350的每一层之间是要被更新的权重(未示出)。每一层的输出可以用作深度卷积网络350中后续层的输入以从在第一卷积块C1处提供的输入数据(例如，图像、音频、视频、传感器数据和/或其他输入数据)学习阶层式特征表示。

[0075] 图4A是解说可使人工智能(AI)功能模块化的示例性软件架构400的框图。使用该架构，应用402可被设计成可使得SOC 420的各种处理块(例如CPU 422、DSP 424、GPU 426和/或NPU 428)在该应用402的运行时操作期间执行支持计算。

[0076] AI应用402可配置成调用在用户空间404中定义的功能，例如，这些功能可提供对指示该设备当前操作位置的场景的检测和识别。例如，AI应用402可取决于识别出的场景是否为办公室、报告厅、餐馆、或室外环境(诸如湖泊)而以不同方式配置话筒和相机。AI应用402可向与在场景检测应用编程接口(API)406中定义的库相关联的经编译程序代码作出请求以提供对当前场景的估计。该请求可最终依赖于配置成基于例如视频和定位数据来提供场景估计的深度神经网络的输出。

[0077] 运行时引擎408(其可以是运行时框架的经编译代码)可进一步可由AI应用402访问。例如，AI应用402可使得运行时引擎请求特定时间间隔的场景估计或由应用的用户接口检测到的事件触发的场景估计。在使得运行时引擎估计场景时，运行时引擎可进而发送信号给在SOC 420上运行的操作系统410(诸如Linux内核412)。操作系统410进而可使得在CPU 422、DSP 424、GPU 426、NPU 428、或其某种组合上执行计算。CPU 422可被操作系统直接访问，而其他处理块可通过驱动器(诸如用于DSP 424、GPU 426、或NPU 428的驱动器414-418)被访问。在示例性示例中，深度神经网络可被配置成在处理块的组合(诸如CPU 422和GPU 426)上运行，或可在NPU 428(如果存在的话)上运行。

[0078] 图4B是解说智能手机452上的AI应用的运行时操作450的框图。AI应用可包括预处理模块454，该预处理模块454可被配置(例如，使用JAVA编程语言被配置)成转换图像456的格式并随后对图像458进行剪裁和/或调整大小。经预处理的图像可接着被传达给分类应用460，该分类应用460包含场景检测后端引擎462，该场景检测后端引擎462可被配置(例如，使用C编程语言被配置)成基于视觉输入来检测和分类场景。场景检测后端引擎462可被配置成通过缩放(466)和剪裁(468)来进一步预处理(464)该图像。例如，该图像可被缩放和剪裁以使所得到的图像是224像素×224像素。这些维度可映射到神经网络的输入维度。神经网络可由深度神经网络块470配置以使得SOC 100的各种处理块进一步借助深度神经网络来处理图像像素。深度神经网络的结果可随后被取阈(472)并被传递通过分类应用460中的指数平滑块474。经平滑的结果可接着使得智能手机452的设置和/或显示改变。

[0079] 在一种配置中，机器学习模型(诸如神经网络)被配置成用于代替网络中的一个或多个层、在经压缩网络的经压缩层之间插入非线性和/或微调经压缩网络。该模型包括代替

装置、插入装置和调谐装置。在一个方面，该代替装置、插入装置和/或调谐装置可以是被配置成执行所叙述的功能的通用处理器102、与通用处理器102相关联的程序存储器、存储器块118、局部处理单元202、和/或路由连接处理单元216。在另一配置中，前述装置可以是被配置成执行由前述装置所叙述的功能的任何模块或任何设备。

[0080] 在另一配置中，机器学习模型(诸如神经网络)被配置成用于用多个经压缩层代替网络中的一个或多个层以产生经压缩网络，从而组合的经压缩层的感受野大小与未压缩层的感受野相匹配。该模型还被配置成用于通过更新这些经压缩层中的一个或多个经压缩层中的权重值来微调经压缩网络。该模型包括代替装置和调谐装置。在一个方面，该代替装置和/或调谐装置可以是被配置成执行所叙述的功能的通用处理器102、与通用处理器102相关联的程序存储器、存储器块118、局部处理单元202、和/或路由连接处理单元216。在另一配置中，前述装置可以是被配置成执行由前述装置所叙述的功能的任何模块或任何设备。

[0081] 在又一配置中，机器学习模型(诸如神经网络)被配置成用于代替网络中的一个或多个层和/或用于通过应用交替最小化过程来确定经压缩层的权重矩阵。该模型包括代替装置和确定装置。在一个方面，该代替装置和/或确定装置可以是被配置成执行所叙述的功能的通用处理器102、与通用处理器102相关联的程序存储器、存储器块118、局部处理单元202、和/或路由连接处理单元216。在另一配置中，前述装置可以是被配置成执行由前述装置所叙述的功能的任何模块或任何设备。

[0082] 根据本公开的某些方面，每个局部处理单元202可被配置成基于网络的一个或多个期望功能性特征来确定网络的参数，以及随着所确定的参数被进一步适配、调谐和更新来使这一个或多个功能性特征朝着期望的功能性特征发展。

[0083] 全连接层的压缩

[0084] 图5A-B是解说根据本公开的各方面的全连接层和经压缩全连接层的框图。如图5A所示，全连接层(fc)500的每一个输出神经元504以全体到全体的方式经由突触连接至每一个输入神经元502(出于解说方便起见示出全体到全体连接的子集)。fc层具有n个输入神经元和m个输出神经元。将第i个输入神经元连接至第j个输出神经元的突触的权重由w_{ij}标示。连接n个输入神经元和m个输出神经元的突触的所有权重可使用矩阵W来共同表示。输出激活向量y可通过将输入激活向量x乘以权重矩阵W并加上偏置向量b来获得。

[0085] 图5B解说了示例性经压缩全连接层550。如图5B所示，单个全连接层用两个全连接层代替。也就是说，图5A中名为‘fc’的全连接层用图5B中所示的两个全连接层‘fc1’和‘fc2’代替。虽然仅示出一个附加全连接层，但本公开并不被如此限定，并且可使用附加全连接层来压缩全连接层fc。例如，图6A-B解说了全连接层以及具有三个全连接层的经压缩全连接层。如图6A-B所示，未压缩全连接层‘fc’用三个全连接层‘fc1’、‘fc2’和‘fc3’代替。居间层中的神经元数目r_1和r_2可基于它们达成的压缩以及结果所得的经压缩网络的性能来确定。

[0086] 可通过将压缩前和压缩后的参数总数进行比较来理解使用压缩的优势。压缩前的参数数目可以等于权重矩阵W中元素数目＝nm、加上偏置向量中的元素数目(其等于m)。因此，压缩前的参数总数等于nm+m。然而，图5B中压缩后的参数数目等于层‘fc1’和‘fc2’中的参数数目之和，其等于nr+rm。取决于r的值，可达成参数数目的显著减少。

[0087] 经压缩网络所达成的有效变换由下式给出：

[0088] $y=W_2W_1x+b$, (1)

[0089] 其中W_1和W_2分别是经压缩全连接层fc1和fc2的权重矩阵。

[0090] 经压缩层的权重矩阵可被确定以使得有效变换是原始全连接层fc所达成的变换的紧密近似：

[0091] $y=Wx+b$。 (2)

[0092] 权重矩阵W_1和W_2可被选取成减小近似误差，近似误差由下式给出：

[0093] $|W-W_2W_1|^2$ (3)

[0094] 从而由经压缩层fc1和fc2达成的式1的有效变换近似式2中指定的原始全连接层的变换。

[0095] 类似地，图6B中压缩后的参数数目等于$nr_1+r_1r_2+r_2m+m$。再次，取决于r_1和r_2的值，可达成参数数目的显著减少。

[0096] 图6B中所示的经压缩网络中的经压缩层‘fc1’、‘fc2’和‘fc3’一起达成的有效变换为$y=W_3W_2W_1x+b$。权重矩阵W_1、W_2和W_3可被确定以使得有效变换是原始‘fc’层(图6A中所示)所达成的变换$y=Wx+b$的近似。

[0097] 一种用于确定经压缩层的权重矩阵的方法是重复地应用奇异值分解(SVD)。SVD可被用来计算W的秩近似，可从该秩近似确定权重矩阵W_1和W_2。

[0098] 用于确定经压缩层的权重矩阵W_1和W_2的另一种方法是应用交替最小化过程。根据交替最小化过程，使用随机值来初始化矩阵并且使用最小二乘法一次一个地交替地更新经压缩矩阵，以减小或最小化式3的目标函数并由此改善该近似。

[0099] 在一些方面，可通过训练经压缩网络来确定权重矩阵。例如，可针对一组训练示例记录未压缩层的输入和输出。随后可使用例如梯度下降来初始化经压缩层以产生对应于未压缩层的输出。

[0100] 由于经压缩网络是原始网络的近似，因此经压缩网络的端对端行为可能与原始网络相比会稍微或显著不同。结果，对于该网络被设计成完成的任务，新的经压缩网络可能不如原始网络表现那样好。为了对抗性能下降，可执行对经压缩网络的权重的微调和修改。例如，可应用通过所有经压缩层的反向传播以修改权重。替换地，可执行通过这些层的子集的反向传播，或者可执行贯穿从输出回到输入的整个模型通过经压缩层和未压缩层两者的反向传播。可通过用一些训练示例教导神经网络来达成微调。取决于可用性，可采用用于训练原始网络的相同训练示例、或不同训练示例集合、或旧训练示例和新训练示例的子集。

[0101] 在微调期间，可选择仅微调新的经压缩层或者所有层一起(其可被称为全栈微调)、或经压缩层和未压缩层的子集。如关于图5B和6B中的示例所示的经压缩层分别是图5A和6A中所示的全连接层类型的另一实例。因此，可规避用于训练经压缩层或用于使用经压缩网络执行推断的新技术的设计或实现。取而代之，可以重用可用于原始网络的任何训练和推断平台。

[0102] 卷积层的压缩

[0103] 神经网络的卷积层也可被压缩。用于压缩卷积层的办法在一些方面类似于以上关于全连接层所描述的。例如，如同全连接层的情形，经压缩卷积层的权重矩阵可被选择以使得经压缩层所达成的有效变换是对原始未压缩卷积层所达成的变换的良好近似。然而，因全连接层与卷积层之间的架构差异而存在一些差异。

[0104] 图7是解说根据本公开的各方面的示例性卷积层的框图。卷积层将n个输入图连接到m个输出图。每个图可包括对应于图像的空间维度或音频信号中的时频维度等等的神经元集合(例如,二维(2-D)网格或3-D图)。例如,在一些方面,该神经元集合可对应于视频数据或体素数据(例如,磁共振成像(MRI)扫描)。该卷积层将卷积内核W^{ij}应用于第i个输入图X^i,,并且将结果加至第j个输出图Y^j,,其中索引i从1到n,并且索引j从1到m。换言之,第j个输出图Y^j可被表达为:

[0105] $$Y^j = \sum_{i=1}^{n} X^i * W^{ij} + B^j。 \qquad (4)$$

[0106] 图8A-B解说了根据本公开的各方面的卷积层的示例压缩。参照图8A-B,名为'conv'的卷积层用两个卷积层'conv1'和'conv2'代替。

[0107] 如以上关于全连接层所指出的,经压缩层(代替层)的数目不限于图8B中所示的两个,并且可根据设计偏好代替地使用任何数目的经压缩层。例如,图9A-B解说了用三个卷积层'conv1'、'conv2'和'conv3'代替'conv'层的示例压缩。在该情形(图9B)中,居间层中的图数目r_1和r_2可基于所达成的压缩以及结果所得的经压缩网络的性能来确定。

[0108] 用两个卷积层代替

[0109] 图8B中描述的经压缩网络中的经压缩层'conv1'和'conv2'一起达成的有效变换可被表达为:

[0110] $$Y^j = \sum_{k=1}^{r_1} Z^k * W_2^{kj} + B^j = \sum_{i=1}^{n} X^i * \left(\sum_{k=1}^{r_1} W_1^{ik} * W_2^{kj}\right) + B^j, \quad (5)$$

[0111] 其中Bj为偏置项,并且Z^k为添加的神经元层的激活向量,并且在此示例中$r_1 = r$。实质上,原始网络中的'conv'层的权重矩阵W^{ij}在经压缩网络中通过'conv1'层的权重矩阵W_1^{ik}和'conv2'层的权重矩阵W_2^{kj}近似为:

[0112] $$W^{ij} \approx \sum_{k=1}^{r_1} W_1^{ik} * W_2^{kj}, \ i = 1,2\ldots,n, \ j = 1,2,\ldots,m。 \quad (6)$$

[0113] 为了确保经压缩卷积层的有效变换(式5)是原始卷积层的变换的紧密近似,经压缩层的权重矩阵可被选择以使得减小或最小化近似误差,该近似误差可被指定为:

[0114] $$\sum_{i,j=1,1}^{n,m} \left(W^{ij} - \sum_{k=1}^{r_1} W_1^{ik} * W_2^{kj} \right)^2 \qquad (7)$$

[0115] 再次,SVD方法、交替最小化过程或类似方法可被用来确定减小或最小化式7的目标函数的经压缩层的权重矩阵。

[0116] 经压缩卷积层的内核大小

[0117] 为了使式7中提及的近似误差有意义,矩阵W^{ij}应当具有与矩阵$W_1^{ik} * W_2^{kj}$相同的维度。出于此目的,经压缩卷积层'conv1'和'conv2'的内核大小可被恰适地选择。例如,如果$k_x \times k_y$标示未压缩'conv'层的内核大小,则经压缩卷积层'conv1'和'conv2'的内核大小可被选择为使得它们满足:

[0118] $(k_{1x}-1)+(k_{2x}-1)=(k_x-1)$

[0119] 以及 (8)

[0120] $(k_{1y}-1)+(k_{2y}-1)=(k_y-1)$,

[0121] 其中$k_{1x} \times k_{1y}$表示'conv1'层的内核大小,并且$k_{2x} \times k_{2y}$表示'conv2'层的内核大小。

[0122] 因此，具有内核大小$k_x \times k_y$的卷积层可用分别具有内核大小$k_x \times k_y$和1×1的两个卷积层来代替，或反之。

[0123] 此外，在一些方面，具有内核大小$k_x \times k_y$的卷积层可用分别具有内核大小1×1、$k_x \times k_y$和1×1的三个卷积层来代替。

[0124] 在一些方面，经压缩卷积层‘conv1’和‘conv2’的相应内核大小k_1和k_2可被选择以满足：

[0125] $$k_1+k_2-1=k, \quad (9)$$

[0126] 其中k是未压缩卷积层‘conv’的内核大小。下表1提供了不同的可能内核大小k_1和k_2的示例，其中未压缩卷积层‘conv’的内核大小为$k=5$。

[0127]

k_1	k_2	k_1+k_2-1
5	1	5
4	2	5
3	3	5
2	4	5
1	5	5

[0128] 表1

[0129] 用多个卷积层代替

[0130] 卷积层也可用不止两个卷积层来代替。为了说明方便起见，假定L个卷积层被用来代替一个卷积层。那么，第1层的权重值W_1被选择为减小或最小化目标函数

[0131] $$\sum_{i,j=1,1}^{n,m}\left(W^{ij}-\sum_{k_1,k_2\ldots k_L=1,1,\ldots 1}^{r_1\ r_2,\cdots,r_L} W_1^{ik_1} * W_2^{k_1k_2} * \cdots * W_L^{k_{L-1}j}\right)^2, \quad (10)$$

[0132] 其中变量r_1标示第1层的输出图的数目。再次，为了使目标函数有意义，第1层的内核大小$k_{1x} \times k_{1y}$可被选择为使得它们满足：

[0133] $$(k_{1x}-1)+(k_{2x}-1)+\cdots+(k_{Lx}-1)=(k_x-1)$$

[0134] 以及 (11)

[0135] $$(k_{1y}-1)+(k_{2y}-1)+\cdots+(K_{Lx}-1)=(k_y-1)。$$

[0136] 相应地，在一些方面，具有内核大小$k_x \times k_y$的卷积层用具有内核大小$k_{1x} \times k_{1y}$、$k_{2x} \times k_{2y}$等的多个卷积层代替，以使得满足性质$(k_{1x}-1)+(k_{2x}-1)+\cdots=(k_x-1)$和$(k_{1y}-1)+(k_{2y}-1)+\cdots=(k_y-1)$。

[0137] 交替最小化过程

[0138] 在特殊情形中，其中经压缩卷积层之一具有内核大小1×1，并且其它经压缩层具有与未压缩层相同的内核大小，则SVD方法可被用来确定优化例如在式7和式10中陈述的目标函数的经压缩权重矩阵。在一个方面，此特殊情形对应于$(k_{1x} \times k_{1y}=1 \times 1$和$k_{2x}=k_{2y}=k_x \times k_y)$或$(k_{1x} \times k_{1y}=k_x \times k_y$和$k_{2x} \times k_{2y}=1 \times 1)$。

[0139] 在一般情形中，其中经压缩层均不具有内核大小1×1，则交替最小化方法可被用来确定优化式7和式10的经压缩权重矩阵。交替最小化方法包括以下步骤：

[0140] (1)用随机值来初始化‘conv1’层的权重矩阵W_1^{ik}和‘conv2’层的权重矩阵W_2^{kj}

[0141] (2)以交替方式求解每个经压缩层的权重：

[0142] (a)固定‘conv2’层权重，并且求解最小化式7的‘conv1’层权重。

[0143] (b)固定‘conv1’层权重，并且求解最小化式7的‘conv2’层权重。

[0144] (3)重复步骤2直至收敛、或者达预定数目的迭代。

[0145] 步骤2a和2b可使用标准最小二乘法过程以闭合形式求解。

[0146] 注意，在使用SVD方法或交替最小化方法确定经压缩权重之后，经压缩网络可被微调以在一定程度上收回分类性能损耗。

[0147] 局部连接层的压缩

[0148] 在卷积层中，贯穿输入图像图应用相同卷积内核，而在局部连接层中，在不同空间位置应用不同权重值。相应地，关于卷积层解释的压缩办法可通过在不同空间位置应用压缩方法(参见以上式4-11)来类似地应用于局部连接层。

[0149] 设计参数的选择

[0150] 压缩中涉及的若干设计参数(例如，全连接层情形中居间神经元的数目；卷积层情形中输出图的数目和内核大小)可例如通过经验式地测量所达成的压缩以及经压缩网络的对应功能性能来选择。

[0151] 在一个示例中，居间层中的神经元数目r可使用参数扫掠来选择。r的值可以按16的增量从16扫掠到min(n,m)。对于每个值，可确定经压缩网络对于验证数据集合的分类准确度。可选择分类准确度下降可接受(例如，低于阈值)的最低值r。

[0152] 在一些方面，可使用‘贪婪搜索’方法来确定设计参数。例如，这些层中展示最大压缩希望的一个或多个层(例如，具有较大计算数目的层)可被压缩。随后可选择性地压缩供压缩的附加层。

[0153] 在一些方面，可通过个体地压缩每一层来确定良好的设计参数。例如，每一层可被压缩到可能的最大程度，使得功能性能不降到低于阈值。对于每一层个体地学习的这些设计参数随后可被用来将网络中的多个层一起压缩。在一起压缩多个层之后，可通过更新经压缩层和未压缩层中的权重值来微调网络以收回功能性能下降。

[0154] 在一些方面，经压缩网络可针对每个所确定的参数值(例如，r)进行微调或者可仅针对最终选择的参数值进行微调。以下伪代码提供了用于选择压缩参数的示例性启发。

[0155] 对于每个全连接层：

[0156] 最大_秩＝min(n,m)

[0157] 对于r＝1:最大_秩

[0158] 压缩所选择的全连接层(留下其它层未压缩)

[0159] 确定分类准确度

[0160] 选择使得准确度下降低于阈值的最低r

[0161] 使用对应的所选r值来压缩所有全连接层

[0162] 微调经压缩网络。

[0163] 在经压缩层之间插入非线性层

[0164] 在用多个层代替一层之后，经压缩层中的神经元可配置有相同激活。然而，可在经压缩层之间添加非线性层以改善网络的表示容量。此外，通过添加非线性层，可达成较高的功能性能。

[0165] 在一个示例性方面，可插入矫正器(rectifier)非线性。在此示例中，可使用max(0,x)非线性来向经压缩层的输出激活应用阈值，随后将它们传递给下一层。在一些方面，矫正器非线性可以是矫正线性单元(ReLU)。也可使用其它种类的非线性，包括abs(x)、sigmoid(x)和双曲正切(tanh(x))或诸如此类。

[0166] 在一些方面，可压缩神经网络、可插入非线性层和/或可应用微调。例如，可训练具有9×9卷积层的神经网络(例如，使用第一训练示例集合)。该9×9卷积层随后可被压缩，将其代替为两个5×5卷积层。可在这两个5×5卷积层之间添加非线性层。可执行微调。在一些方面，随后可通过将每个5×5卷积层代替为两个3×3卷积层来进一步压缩该网络。进一步，可在每个3×3卷积层之间添加非线性层。

[0167] 随后可重复附加的微调和压缩直至获得期望性能。在以上示例中，与原始网络相比，最终的经压缩网络具有四个3×3卷积层及其间的非线性层，而非一个9×9卷积层。具有三个附加卷积层和非线性层的较深网络可达成更佳的功能性能。

[0168] 图10解说了用于压缩神经网络的方法1000。在框1002，该过程用多个经压缩层代替神经网络中的一个或多个层以产生经压缩网络。在一些方面，经压缩层具有与初始未压缩网络中的层相同的类型。例如，全连接层可由多个全连接层代替，并且卷积层可由多个卷积层代替。类似地，局部连接层可由多个局部连接层代替。

[0169] 在框1004，该过程在经压缩网络的经压缩层之间插入非线性。在一些方面，该过程通过向经压缩层的神经元应用非线性激活函数来插入非线性。非线性激活函数可包括矫正器、绝对值函数、双曲正切函数、sigmoid(S形)函数或其它非线性激活函数。

[0170] 此外，在框1006，该过程通过更新这些层中的至少一个层中的权重值来微调经压缩网络。可使用反向传播或其它标准过程来执行微调。可通过更新经压缩神经网络中的所有权重值或者通过更新经压缩层的子集、未压缩层的子集、或两者的混合来执行微调。可使用训练示例来执行微调。该训练示例可以与用于教导原始神经网络的那些训练示例相同，或者可包括新的示例集合，或者为两者的混合。

[0171] 在一些方面，训练示例可包括来自智能手机或其它移动设备的数据，包括例如照片图库。

[0172] 在一些方面，该过程可进一步包括通过以下操作来初始化神经网络：重复地应用压缩、插入非线性层、以及微调作为用于初始化较深神经网络的方法。

[0173] 图11解说了用于压缩神经网络的方法1100。在框1102，该过程用多个经压缩层代替神经网络中的至少一个层以产生经压缩网络，从而组合的经压缩层的感受野大小与未压缩层的感受野相匹配。在一些方面，未压缩层的内核大小等于感受野大小。

[0174] 此外，在框1104，该过程通过更新这些经压缩层中的至少一个经压缩层中的权重值来微调经压缩网络。

[0175] 图12解说了用于压缩神经网络的方法1200。在框1202，该过程用多个经压缩层代替神经网络中的一个或多个层以产生经压缩网络。在一些方面，经压缩层具有与初始未压缩网络中的层相同的类型。例如，全连接层可由多个全连接层代替，并且卷积层可由多个卷积层代替。类似地，局部连接层可由多个局部连接层代替。

[0176] 此外，在框1204，该过程通过应用交替最小化过程来确定该多个经压缩层的权重矩阵。

[0177] 在一些方面，该过程还通过更新这些经压缩层中的至少一个经压缩层中的权重值来微调经压缩网络。可通过更新经压缩神经网络中的所有权重值或者通过更新经压缩层的子集、未压缩层的子集、或两者的混合来执行微调。可使用训练示例来执行微调。该训练示例可以与用于教导原始神经网络的那些训练示例相同，或者可包括新的示例集合，或者为两者的混合。

[0178] 可在单个阶段中执行微调或者可在多个阶段中执行微调。例如，当在多个阶段中执行微调时，在第一阶段，仅对经压缩层的子集执行微调。随后，在第二阶段，针对经压缩层和未压缩层的子集执行微调。

[0179] 图13解说了根据本公开的各方面的用于压缩神经网络的方法1300。在框1302，该过程表征机器学习网络，诸如沿多个维度的神经网络。在一些方面，神经网络的任务性能(P)可被表征。如果神经网络是对象分类器，则任务性能可以是测试集合上被正确分类的对象的百分比。存储器占用(M)和/或推断时间(T)也可被表征。存储器占用可以例如对应于神经网络的权重和其它模型参数的存储器存储要求。推断时间(T)可以是在新对象被呈现给神经网络之后在该新对象的分类期间逝去的时间。

[0180] 在框1304，该过程用多个经压缩层代替神经网络中的一个或多个层以产生经压缩网络。在一些方面，经压缩层可具有与初始未压缩网络中的层相同的类型。例如，全连接层可由多个全连接层代替，并且卷积层可由多个卷积层代替。类似地，局部连接层可由多个局部连接层代替。在示例性配置中，该过程用第二层和第三层代替神经网络的第一层，以使得第三层的大小等于第一层的大小。在第三层被如此指定的情况下，经压缩神经网络的第三层的输出可模仿第一层的输出。该过程可进一步指定第二层的大小(其可被标示为(r))，以使得经压缩神经网络(其包括第二层和第三层)的组合存储器占用小于未压缩神经网络的存储器占用(M)。作为存储器占用减小的替换或补充，第二层的大小可被选择以使得经压缩神经网络的推断时间可以小于未压缩神经网络的推断时间(T)。

[0181] 在框1306，该过程初始化神经网络的代替层的参数。在一种示例性配置中，投射到经压缩神经网络的第二层的权重矩阵(W1)以及从第二层投射到经压缩神经网络的第三层的权重矩阵(W2)可被初始化为随机值。在一些方面，权重参数的初始化可涉及交替最小化过程。在一些方面，参数的初始化可涉及奇异值分解。权重的初始设置可被称为初始化，因为根据该方法的附加方面，权重可经历进一步修改。

[0182] 在框1308，该过程在经压缩网络的经压缩层之间插入非线性。例如，该过程可在经压缩神经网络的第二层中插入非线性。在一些方面，该过程通过向经压缩层的神经元应用非线性激活函数来插入非线性。非线性激活函数可包括矫正器、绝对值函数、双曲正切函数、sigmoid函数或其它非线性激活函数。

[0183] 在框1310，该过程通过更新这些层中的一个或多个层中的权重值来微调经压缩网络。可使用反向传播或其它标准过程来执行微调。可通过更新经压缩神经网络中的所有权重值或者通过更新经压缩层的子集、未压缩层的子集、或两者的混合来执行微调。可使用训练示例来执行微调。该训练示例可以与用于教导原始神经网络的那些训练示例相同，或者可包括新的示例集合，或者为两者的混合。在示例性配置中，可对经压缩神经网络的第二层和第三层执行微调。进而，神经网络中的所有权重可在框1312被调整，从而可达成较高水平的任务性能。

[0184] 经微调的经压缩神经网络可沿若干维度被表征，包括例如任务性能(P)。如果经压缩神经网络的任务性能高于可接受阈值，则该过程可终止。如果经压缩神经网络的任务性能低于可接受阈值(1314:否)，则可在框1316增大(r)的值(其可以是经压缩神经网络中的层的大小)。否则(1314:是)，该过程在框1318结束。(r)的新值可被选择以使得存储器占用(M)和/或推断时间(T)小于未压缩神经网络。初始化参数、插入非线性、微调和表征性能的步骤可被重复直至达到可接受任务性能。

[0185] 尽管方法1300描绘了经压缩层的大小(r)可从较小值增加到较大值直至获得可接受任务性能的过程，但本公开并不被如此限定。根据本公开的各方面，(r)的值可初始被保守地选择以使得经压缩神经网络的任务性能很可能获得可接受任务性能。该方法由此可生成可代替未压缩神经网络使用的经压缩神经网络，从而可实现存储器占用或推断时间优势。该方法随后可继续以(r)的此类越来越小的值进行操作，直至不能获得可接受任务性能。以此方式，该方法可在执行该方法的过程期间提供越来越高的压缩水平。

[0186] 以上所描述的方法的各种操作可由能够执行相应功能的任何合适的装置来执行。这些装置可包括各种硬件和/或软件组件和/或模块，包括但不限于电路、专用集成电路(ASIC)、或处理器。一般而言，在附图中有解说的操作的场合，那些操作可具有带相似编号的相应配对装置加功能组件。

[0187] 如本文所使用的，术语“确定”涵盖各种各样的动作。例如，“确定”可包括演算、计算、处理、推导、研究、查找(例如，在表、数据库或其他数据结构中查找)、探知及诸如此类。另外，“确定”可包括接收(例如接收信息)、访问(例如访问存储器中的数据)、及类似动作。而且，“确定”可包括解析、选择、选取、确立及类似动作。

[0188] 如本文所使用的，引述一列项目中的“至少一个”的短语是指这些项目的任何组合，包括单个成员。作为示例，“a、b或c中的至少一者”旨在涵盖:a、b、c、a-b、a-c、b-c、以及a-b-c。

[0189] 结合本公开所描述的各种解说性逻辑框、模块、以及电路可用设计成执行本文所描述功能的通用处理器、数字信号处理器(DSP)、专用集成电路(ASIC)、现场可编程门阵列信号(FPGA)或其他可编程逻辑器件(PLD)、分立的门或晶体管逻辑、分立的硬件组件或其任何组合来实现或执行。通用处理器可以是微处理器，但在替换方案中，处理器可以是任何市售的处理器、控制器、微控制器、或状态机。处理器还可以被实现为计算设备的组合，例如DSP与微处理器的组合、多个微处理器、与DSP核心协同的一个或多个微处理器、或任何其它此类配置。

[0190] 结合本公开所描述的方法或算法的步骤可直接在硬件中、在由处理器执行的软件模块中、或在这两者的组合中体现。软件模块可驻留在本领域所知的任何形式的存储介质中。可使用的存储介质的一些示例包括随机存取存储器(RAM)、只读存储器(ROM)、闪存、可擦除可编程只读存储器(EPROM)、电可擦除可编程只读存储器(EEPROM)、寄存器、硬盘、可移动盘、CD-ROM，等等。软件模块可包括单条指令、或许多条指令，且可分布在若干不同的代码段上，分布在不同的程序间以及跨多个存储介质分布。存储介质可被耦合到处理器以使得该处理器能从/向该存储介质读写信息。替换地，存储介质可以被整合到处理器。

[0191] 本文所公开的方法包括用于实现所描述的方法的一个或多个步骤或动作。这些方法步骤和/或动作可以彼此互换而不会脱离权利要求的范围。换言之，除非指定了步骤或动

作的特定次序，否则具体步骤和/或动作的次序和/或使用可以改动而不会脱离权利要求的范围。

[0192] 所描述的功能可在硬件、软件、固件或其任何组合中实现。如果以硬件实现，则示例硬件配置可包括设备中的处理系统。处理系统可以用总线架构来实现。取决于处理系统的具体应用和整体设计约束，总线可包括任何数目的互连总线和桥接器。总线可将包括处理器、机器可读介质、以及总线接口的各种电路链接在一起。总线接口可用于尤其将网络适配器等经由总线连接至处理系统。网络适配器可用于实现信号处理功能。对于某些方面，用户接口(例如，按键板、显示器、鼠标、操纵杆，等等)也可以被连接到总线。总线还可以链接各种其他电路，诸如定时源、外围设备、稳压器、功率管理电路以及类似电路，它们在本领域中是众所周知的，因此将不再进一步描述。

[0193] 处理器可负责管理总线和一般处理，包括执行存储在机器可读介质上的软件。处理器可用一个或多个通用和/或专用处理器来实现。示例包括微处理器、微控制器、DSP处理器、以及其他能执行软件的电路系统。软件应当被宽泛地解释成意指指令、数据、或其任何组合，无论是被称作软件、固件、中间件、微代码、硬件描述语言、或其他。作为示例，机器可读介质可包括随机存取存储器(RAM)、闪存、只读存储器(ROM)、可编程只读存储器(PROM)、可擦式可编程只读存储器(EPROM)、电可擦式可编程只读存储器(EEPROM)、寄存器、磁盘、光盘、硬驱动器、或者任何其他合适的存储介质、或其任何组合。机器可读介质可被实施在计算机程序产品中。该计算机程序产品可以包括包装材料。

[0194] 在硬件实现中，机器可读介质可以是处理系统中与处理器分开的一部分。然而，如本领域技术人员将容易领会的，机器可读介质或其任何部分可在处理系统外部。作为示例，机器可读介质可包括传输线、由数据调制的载波、和/或与设备分开的计算机产品，所有这些都可由处理器通过总线接口来访问。替换地或补充地，机器可读介质或其任何部分可被集成到处理器中，诸如高速缓存和/或通用寄存器文件可能就是这种情形。虽然所讨论的各种组件可被描述为具有特定位置，诸如局部组件，但它们也可按各种方式来配置，诸如某些组件被配置成分布式计算系统的一部分。

[0195] 处理系统可以被配置为通用处理系统，该通用处理系统具有一个或多个提供处理器功能性的微处理器、以及提供机器可读介质中的至少一部分的外部存储器，它们都通过外部总线架构与其他支持电路系统链接在一起。替换地，该处理系统可以包括一个或多个神经元形态处理器以用于实现本文所述的神经元模型和神经系统模型。作为另一替换方案，处理系统可以用带有集成在单块芯片中的处理器、总线接口、用户接口、支持电路系统、和至少一部分机器可读介质的专用集成电路(ASIC)来实现，或者用一个或多个现场可编程门阵列(FPGA)、可编程逻辑器件(PLD)、控制器、状态机、门控逻辑、分立硬件组件、或者任何其他合适的电路系统、或者能执行本公开通篇所描述的各种功能性的电路的任何组合来实现。取决于具体应用和加诸于整体系统上的总设计约束，本领域技术人员将认识到如何最佳地实现关于处理系统所描述的功能性。

[0196] 机器可读介质可包括数个软件模块。这些软件模块包括当由处理器执行时使处理系统执行各种功能的指令。这些软件模块可包括传送模块和接收模块。每个软件模块可以驻留在单个存储设备中或者跨多个存储设备分布。作为示例，当触发事件发生时，可以从硬驱动器中将软件模块加载到RAM中。在软件模块执行期间，处理器可以将一些指令加载到高

速缓存中以提高访问速度。随后可将一个或多个高速缓存行加载到通用寄存器文件中以供处理器执行。在以下述及软件模块的功能性时，将理解此类功能性是在处理器执行来自该软件模块的指令时由该处理器来实现的。此外，应领会，本公开的各方面导致对实现此类方面的处理器、计算机、机器、或其它系统的机能的改进。

[0197] 如果以软件实现，则各功能可作为一条或多条指令或代码存储在计算机可读介质上或藉其进行传送。计算机可读介质包括计算机存储介质和通信介质两者，这些介质包括促成计算机程序从一地向另一地转移的任何介质。存储介质可以是能被计算机访问的任何可用介质。作为示例而非限定，此类计算机可读介质可包括RAM、ROM、EEPROM、CD-ROM或其他光盘存储、磁盘存储或其他磁存储设备、或能用于携带或存储指令或数据结构形式的期望程序代码且能被计算机访问的任何其他介质。另外，任何连接也被正当地称为计算机可读介质。例如，如果软件是使用同轴电缆、光纤电缆、双绞线、数字订户线(DSL)、或无线技术(诸如红外(IR)、无线电、以及微波)从web网站、服务器、或其他远程源传送而来，则该同轴电缆、光纤电缆、双绞线、DSL或无线技术(诸如红外、无线电、以及微波)就被包括在介质的定义之中。如本文中所使用的盘(disk)和碟(disc)包括压缩碟(CD)、激光碟、光碟、数字多用碟(DVD)、软盘、和蓝光®碟，其中盘(disk)常常磁性地再现数据，而碟(disc)用激光来光学地再现数据。因此，在一些方面，计算机可读介质可包括非瞬态计算机可读介质(例如，有形介质)。另外，对于其他方面，计算机可读介质可包括瞬态计算机可读介质(例如，信号)。上述的组合应当也被包括在计算机可读介质的范围内。

[0198] 因此，某些方面可包括用于执行本文中给出的操作的计算机程序产品。例如，此类计算机程序产品可包括其上存储(和/或编码)有指令的计算机可读介质，这些指令能由一个或多个处理器执行以执行本文中所描述的操作。对于某些方面，计算机程序产品可包括包装材料。

[0199] 此外，应当领会，用于执行本文中所描述的方法和技术的模块和/或其它恰适装置能由用户终端和/或基站在适用的场合下载和/或以其他方式获得。例如，此类设备能被耦合至服务器以促成用于执行本文中所描述的方法的装置的转移。替换地，本文所述的各种方法能经由存储装置(例如，RAM、ROM、诸如压缩碟(CD)或软盘等物理存储介质等)来提供，以使得一旦将该存储装置耦合至或提供给用户终端和/或基站，该设备就能获得各种方法。此外，可利用适于向设备提供本文所描述的方法和技术的任何其他合适的技术。

[0200] 将理解，权利要求并不被限定于以上所解说的精确配置和组件。可在以上所描述的方法和装置的布局、操作和细节上作出各种改动、更换和变形而不会脱离权利要求的范围。

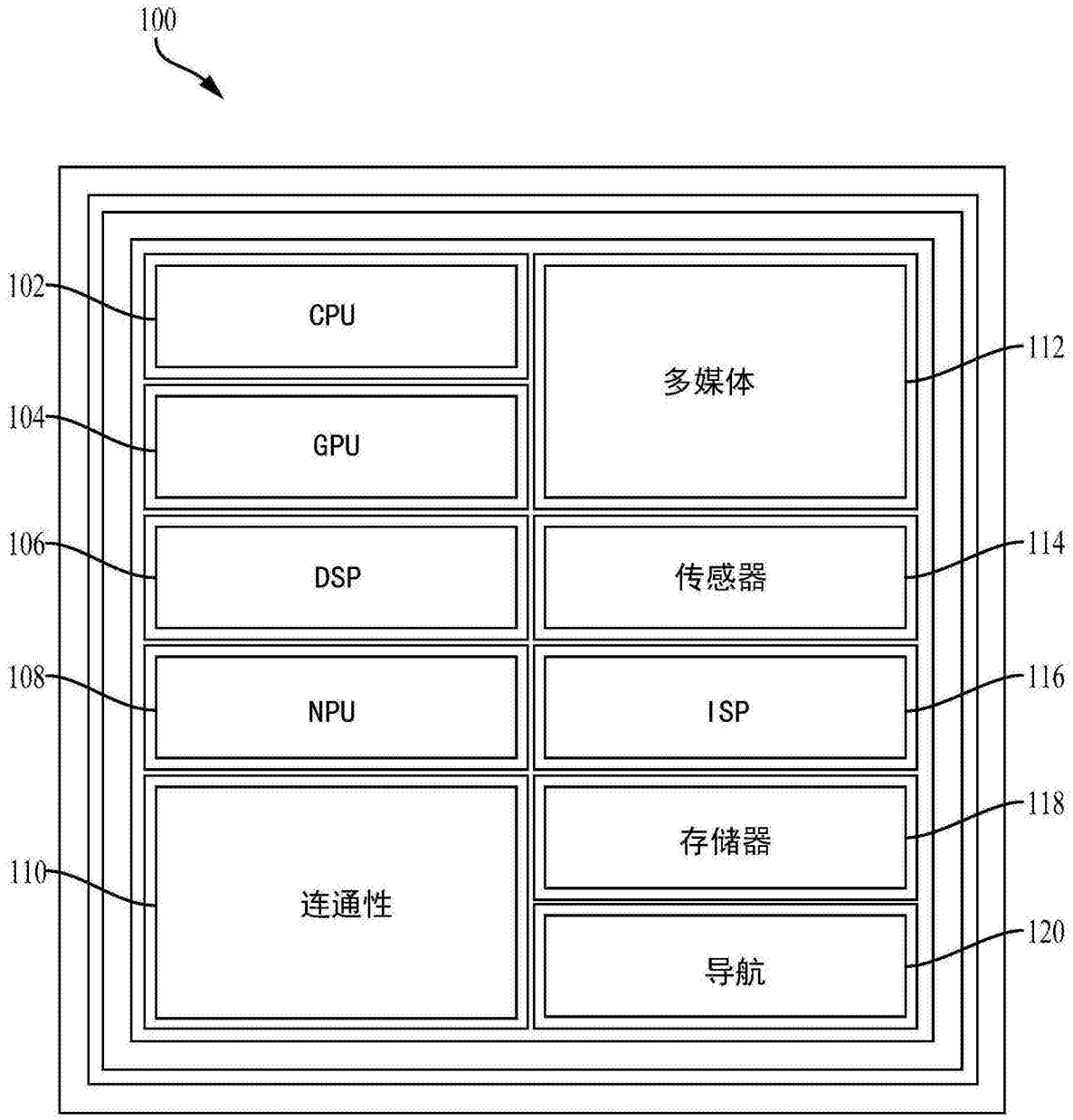

100
102
CPU
104
GPU
106
DSP
108
NPU
110
连通性
112
多媒体
114
传感器
116
ISP
118
存储器
120
导航

图1

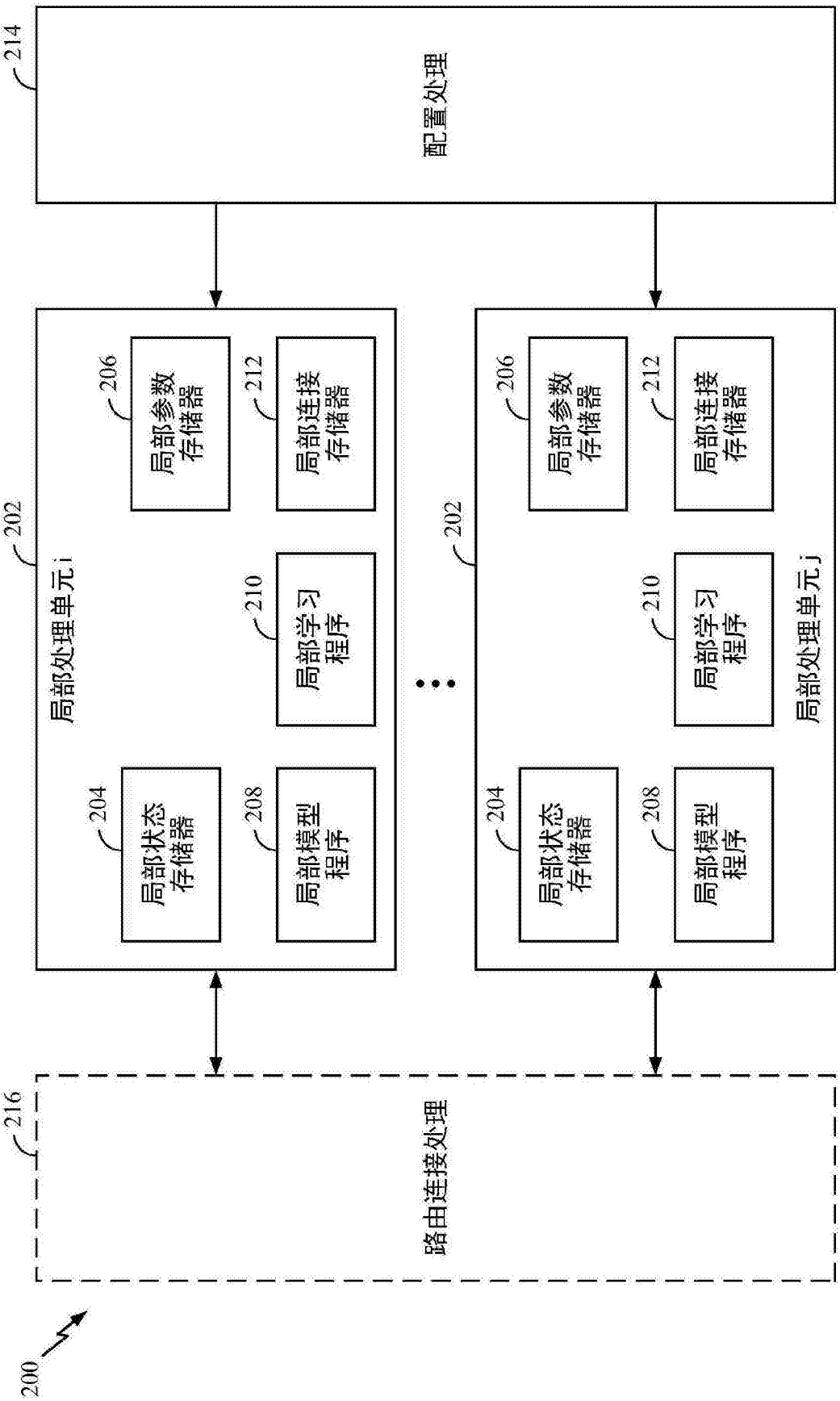

214
配置处理
202
局部处理单元 i
206
局部参数存储器
212
局部连接存储器
210
局部学习程序
204
局部状态存储器
208
局部模型程序
202
局部处理单元 j
206
局部参数存储器
212
局部连接存储器
210
局部学习程序
204
局部状态存储器
208
局部模型程序
216
路由连接处理
200

图2

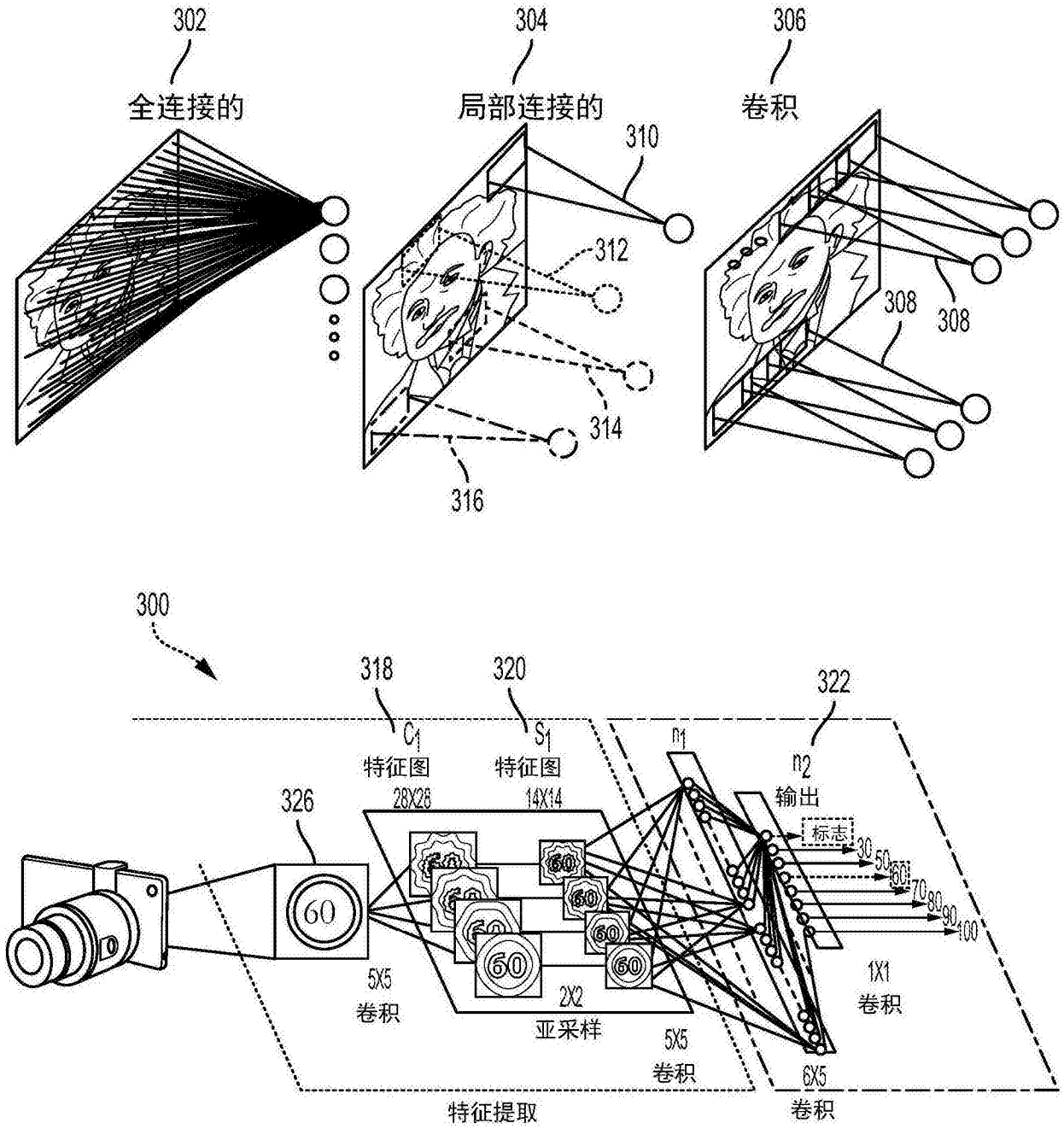

302
全连接的
304
局部连接的
306
卷积
310
312
314
316
308
308
300
318
320
322
C1
特征图
28X28
S1
特征图
14X14
n1
n2
输出
标志
326
60
5X5
卷积
2X2
亚采样
5X5
卷积
1X1
卷积
6X5
卷积
特征提取

图3A

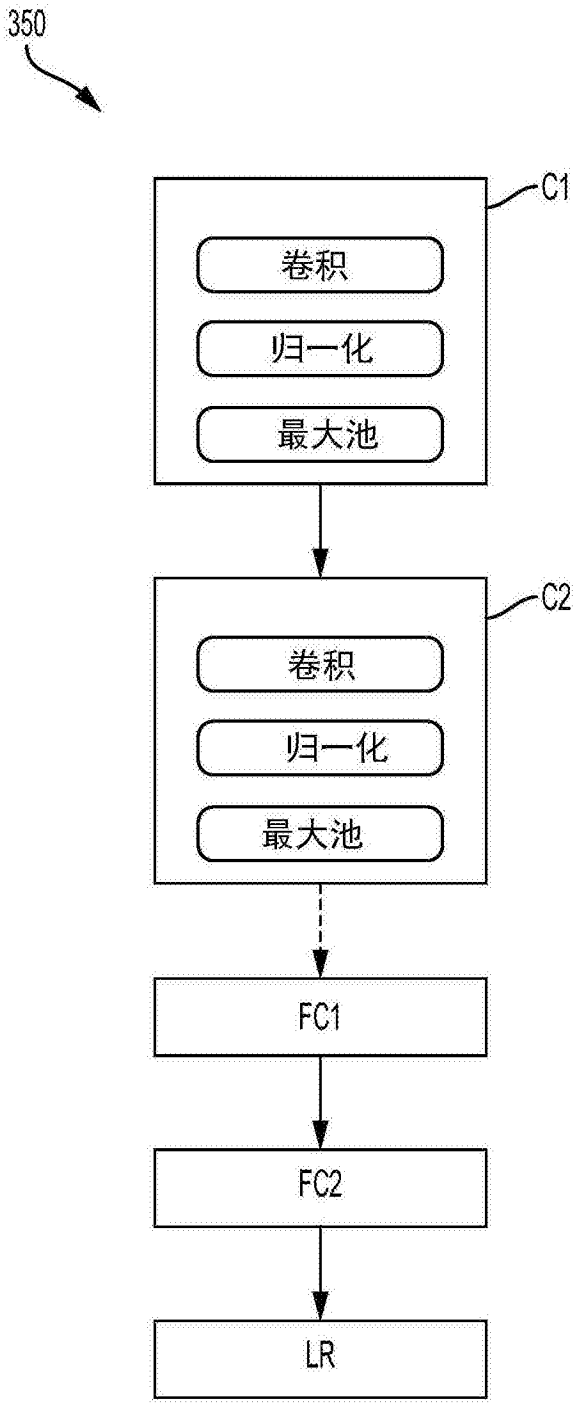


图3B

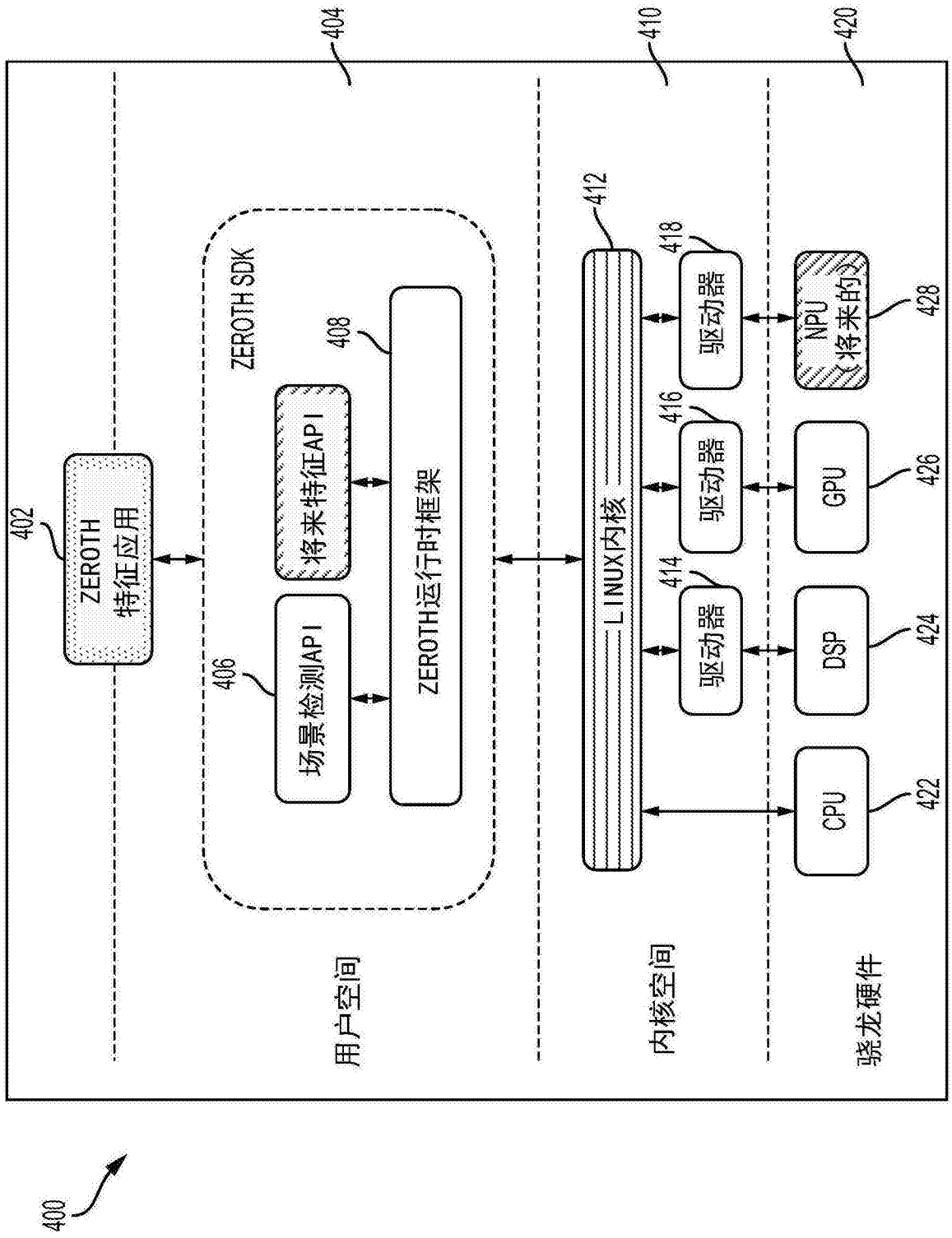

400
404
410
420
402
ZEROTH
特征应用
ZEROTH SDK
408
将来特征API
406
场景检测API
ZEROTH运行时框架
412
LINUX内核
418
驱动器
416
驱动器
414
驱动器
NPU
(将来的)
428
GPU
426
DSP
424
CPU
422
用户空间
内核空间
骁龙硬件

图4A

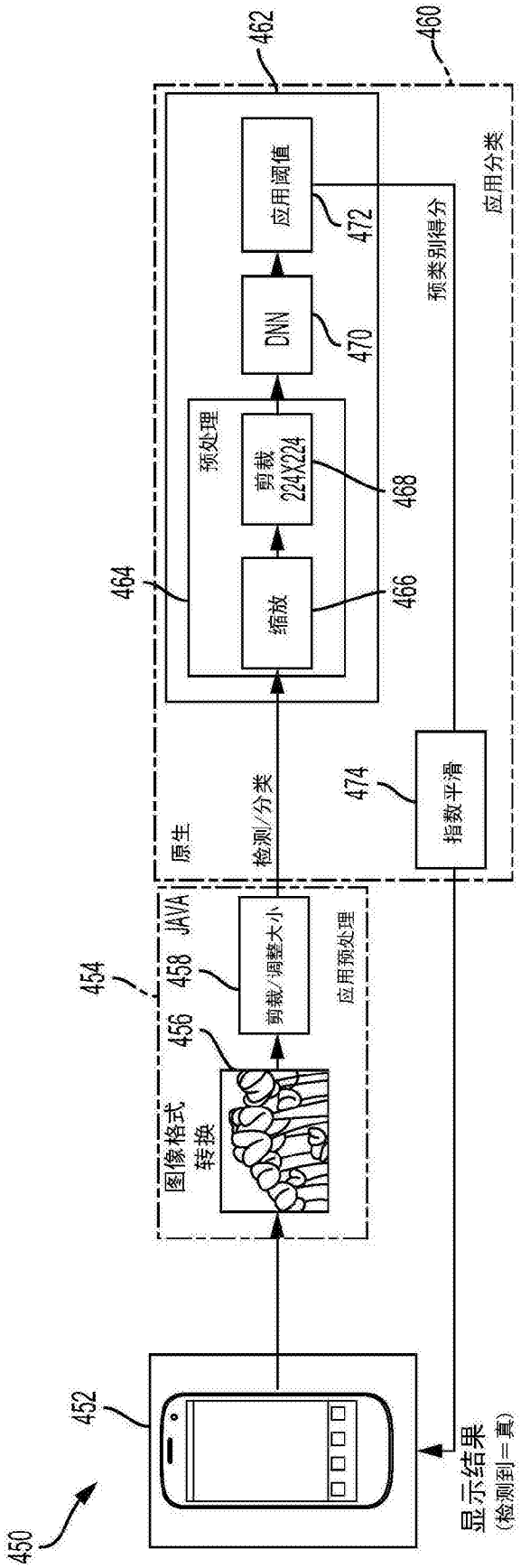

462
460
应用阈值
472
应用分类
预类别得分
DNN
470
预处理
剪裁
224X224
468
464
缩放
466
474
指数平滑
原生
检测/分类
454
JAVA
458
剪裁/调整大小
应用预处理
456
图像格式
转换
452
450
显示结果
(检测到=真)

图4B

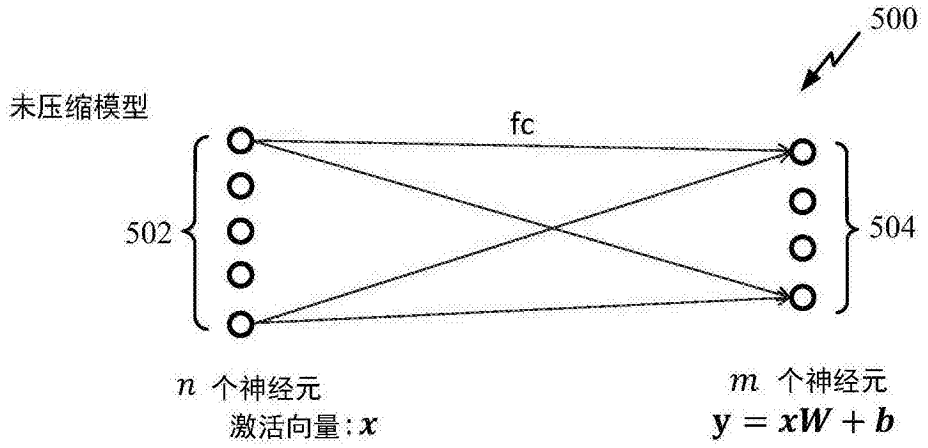

500
未压缩模型
fc
502
504
n 个神经元
激活向量：$\boldsymbol{x}$
m 个神经元
$\mathbf{y} = \boldsymbol{xW} + \boldsymbol{b}$

图5A

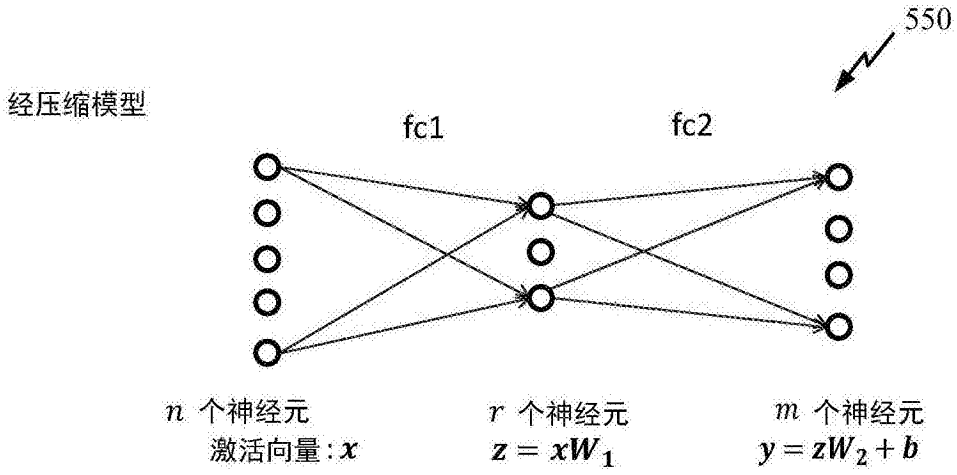

550
经压缩模型
fc1
fc2
n 个神经元
激活向量：$\boldsymbol{x}$
r 个神经元
$\boldsymbol{z} = \boldsymbol{xW_1}$
m 个神经元
$\boldsymbol{y} = \boldsymbol{zW_2} + \boldsymbol{b}$

图5B

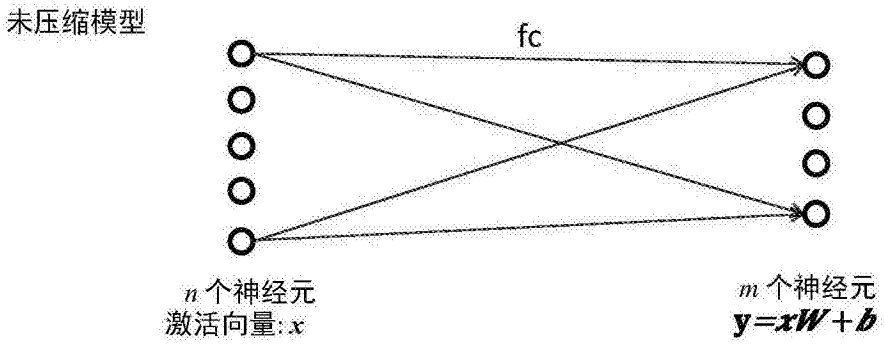

未压缩模型
fc
n 个神经元
激活向量：x
m 个神经元
$\mathbf{y}=\boldsymbol{xW}+\boldsymbol{b}$

图6A

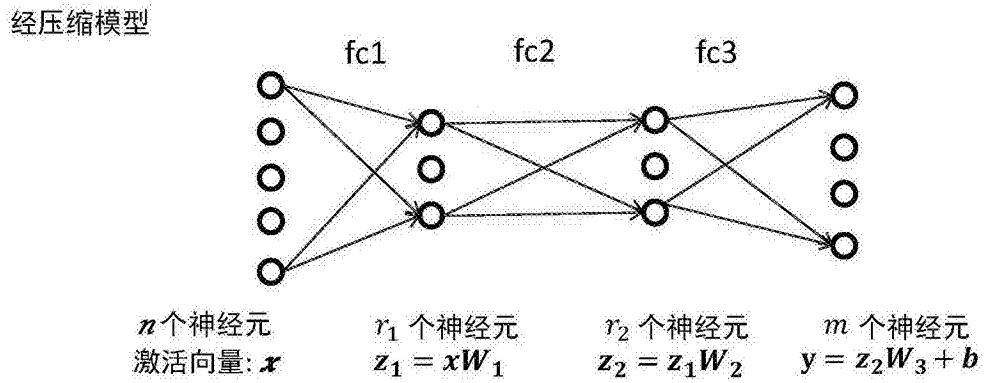

经压缩模型
fc1
fc2
fc3
n个神经元
激活向量: $\boldsymbol{x}$
r_1 个神经元
$\boldsymbol{z_1 = xW_1}$
r_2 个神经元
$\boldsymbol{z_2 = z_1 W_2}$
m 个神经元
$\mathbf{y = z_2 W_3 + b}$

图6B

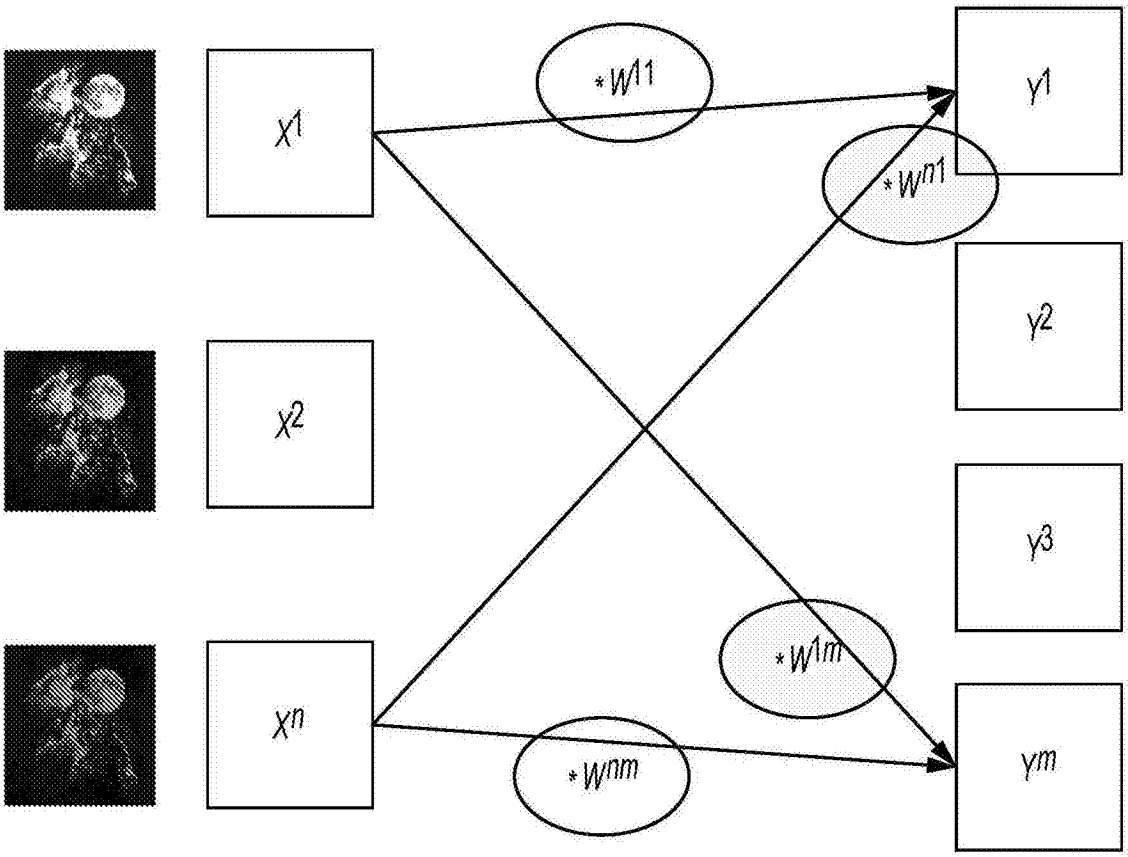

X1
X2
Xn
*W11
*Wn1
*W1m
*Wnm
Y1
Y2
Y3
Ym

图7

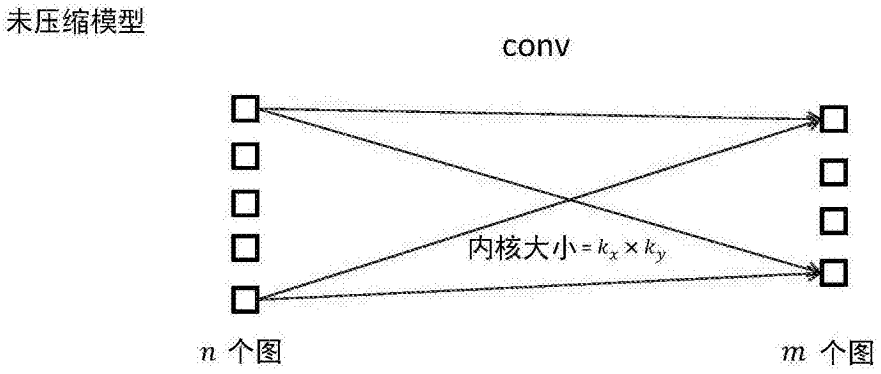

未压缩模型
conv
内核大小 = $k_x \times k_y$
n 个图
m 个图

图8A

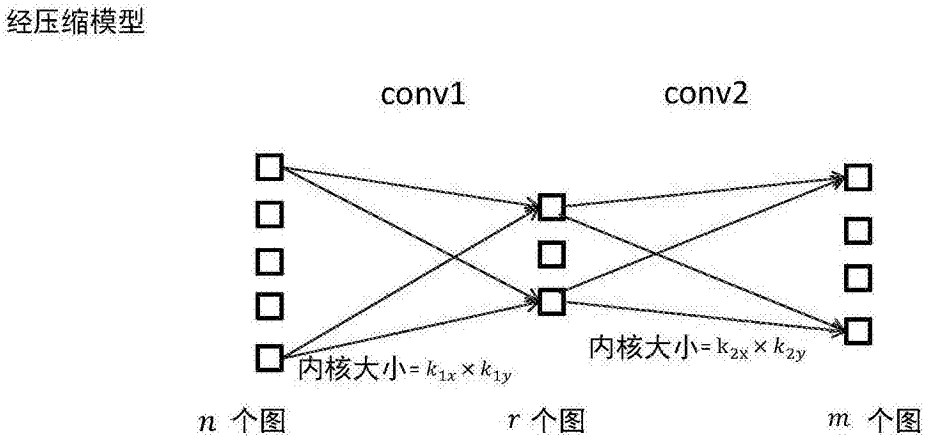

经压缩模型
conv1
conv2
内核大小 = $k_{1x} \times k_{1y}$
内核大小 = $k_{2x} \times k_{2y}$
n 个图
r 个图
m 个图

图8B

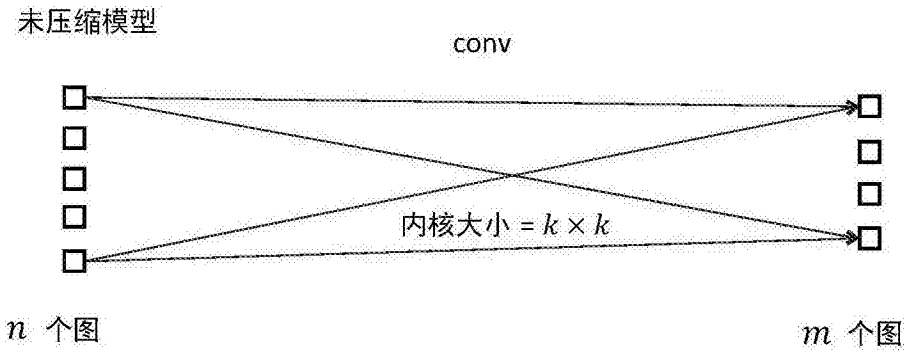

未压缩模型
conv
内核大小 = $k \times k$
n 个图
m 个图

图9A

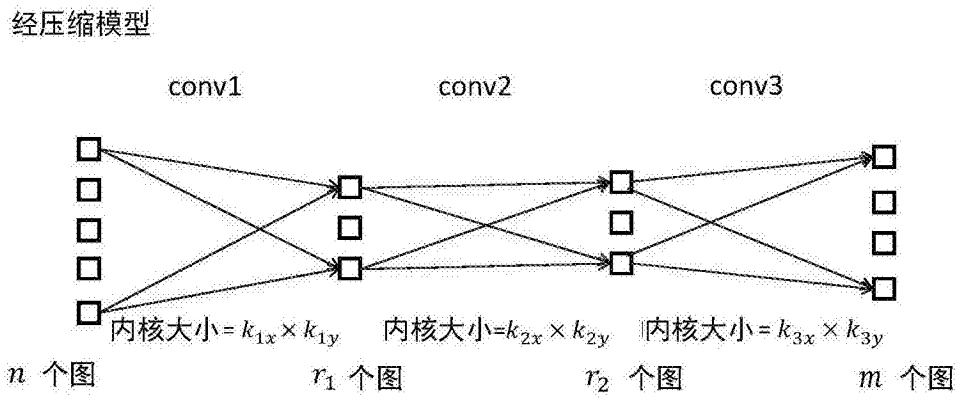

经压缩模型
conv1
conv2
conv3
内核大小 = $k_{1x} \times k_{1y}$
内核大小=$k_{2x} \times k_{2y}$
内核大小 = $k_{3x} \times k_{3y}$
n 个图
r_1 个图
r_2 个图
m 个图

图9B

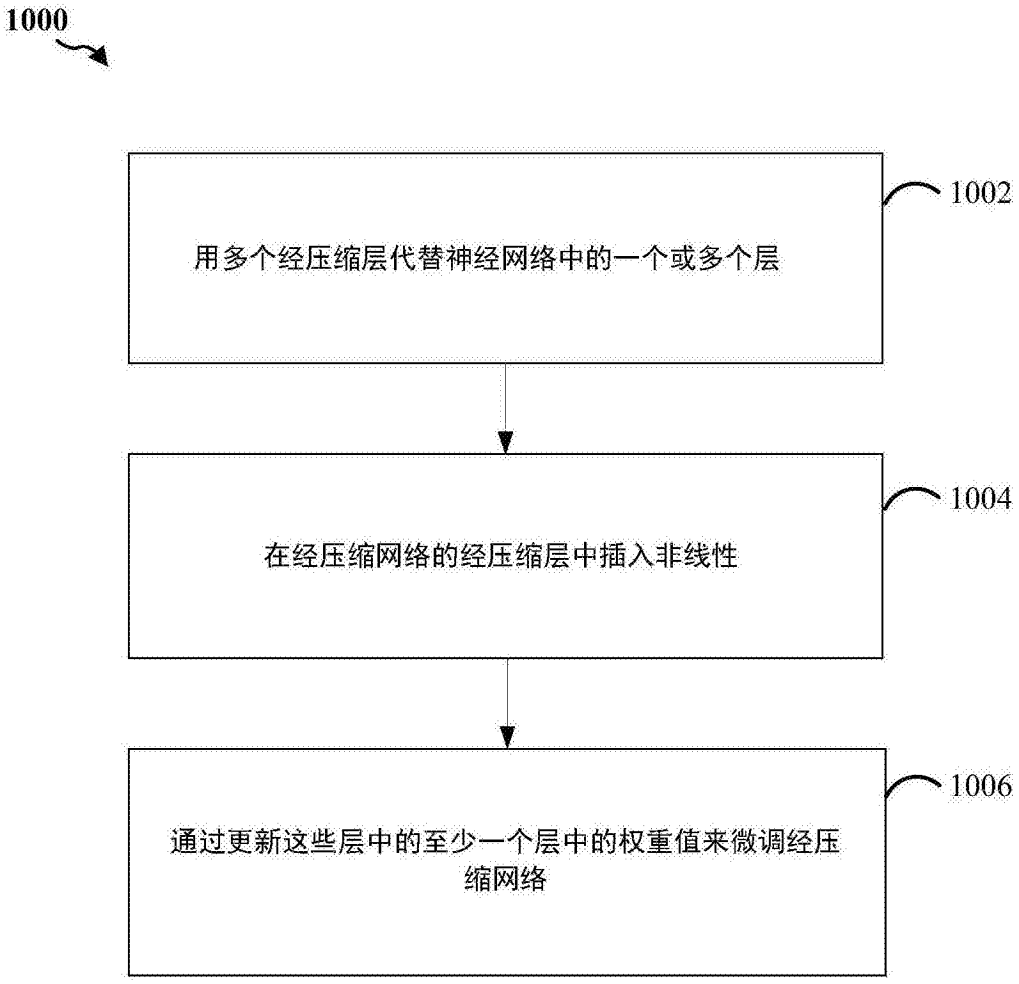

1000
用多个经压缩层代替神经网络中的一个或多个层
1002
在经压缩网络的经压缩层中插入非线性
1004
通过更新这些层中的至少一个层中的权重值来微调经压缩网络
1006

图10

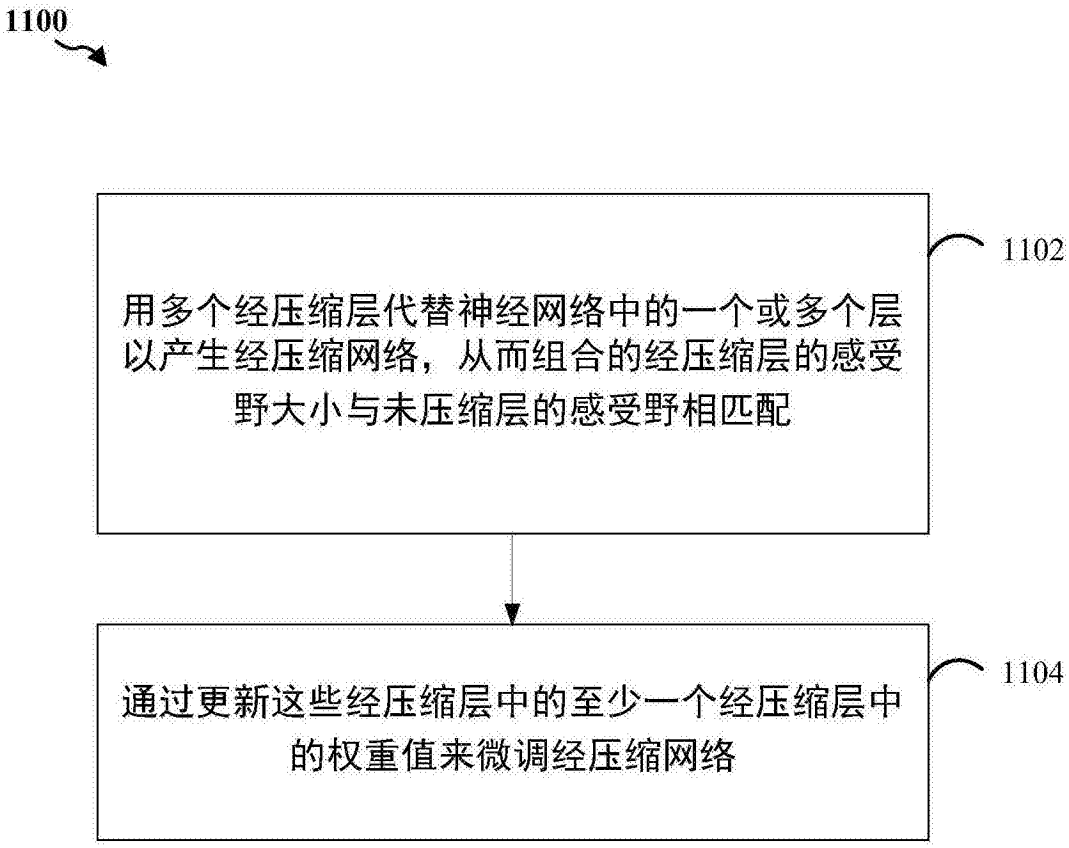

1100
用多个经压缩层代替神经网络中的一个或多个层以产生经压缩网络，从而组合的经压缩层的感受野大小与未压缩层的感受野相匹配
1102
通过更新这些经压缩层中的至少一个经压缩层中的权重值来微调经压缩网络
1104

图11

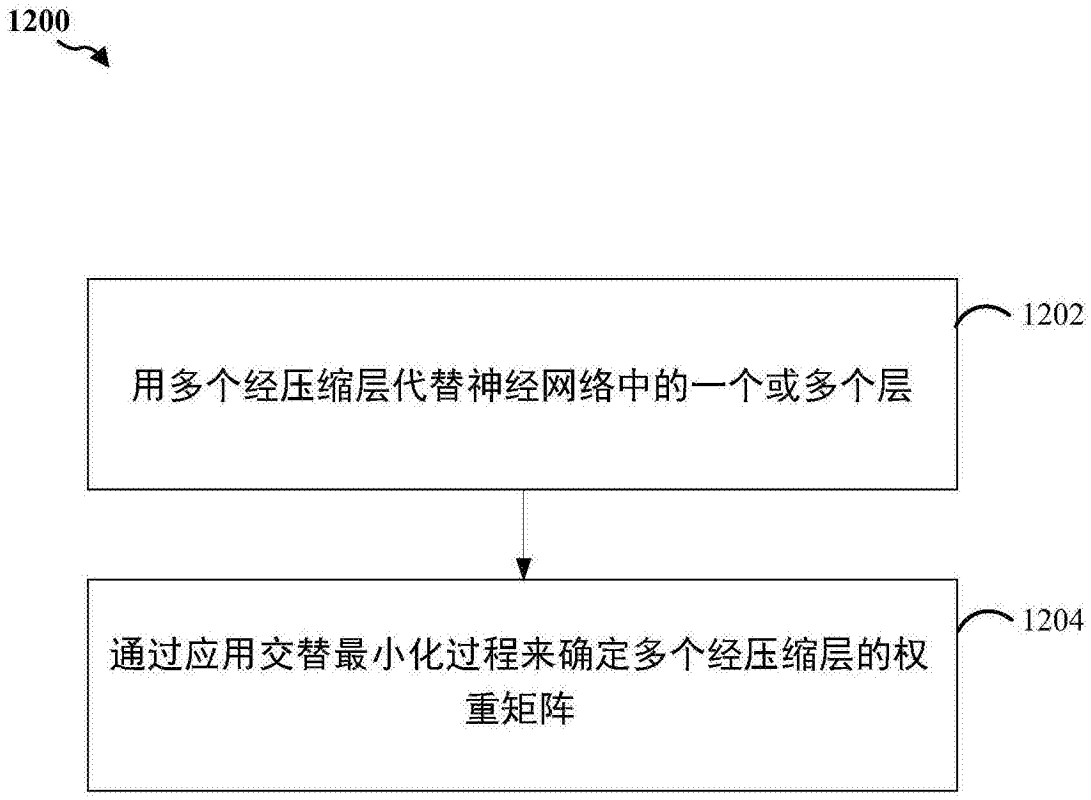

1200
用多个经压缩层代替神经网络中的一个或多个层
1202
通过应用交替最小化过程来确定多个经压缩层的权重矩阵
1204

图12

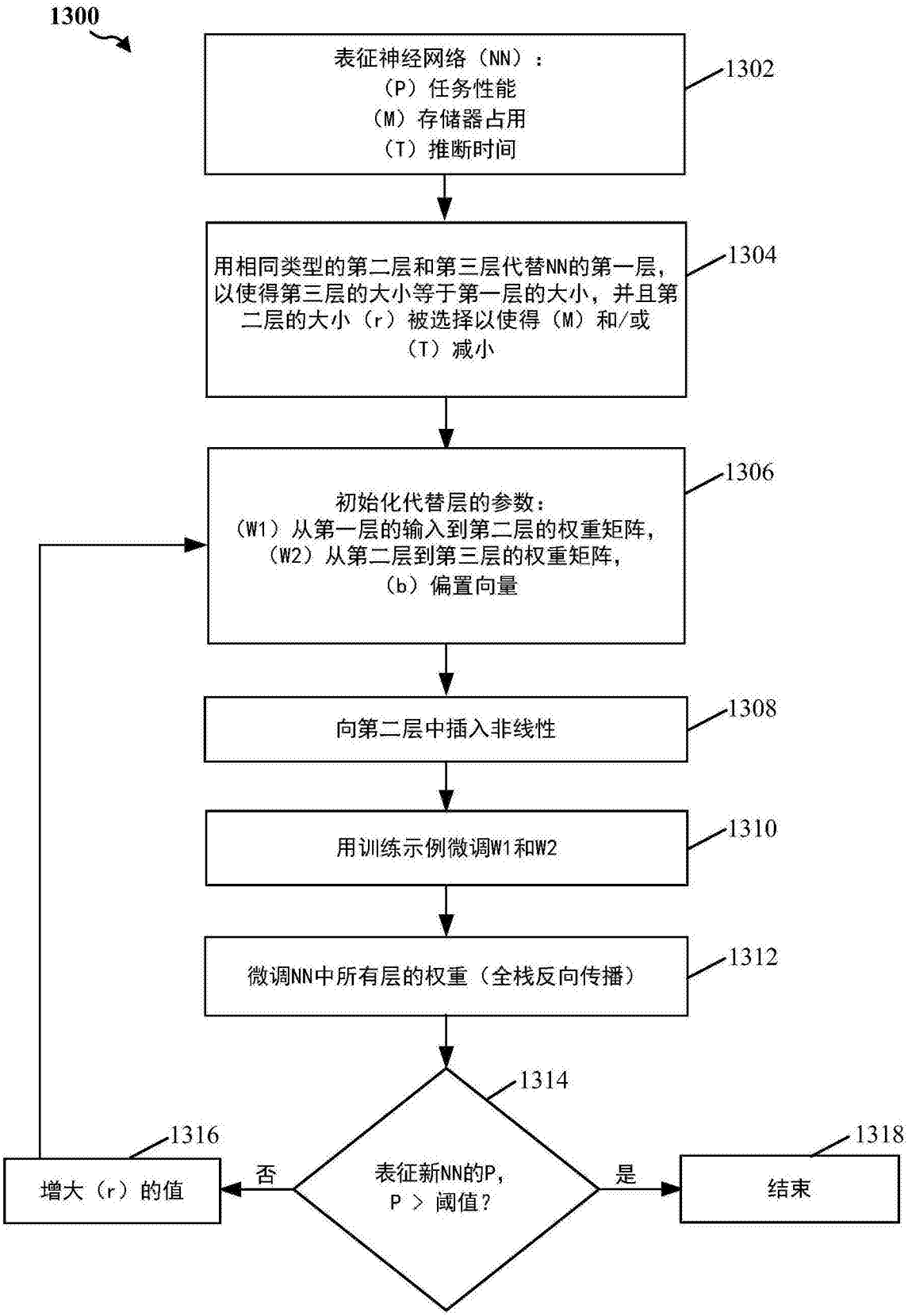

1300
表征神经网络（NN）：
（P）任务性能
（M）存储器占用
（T）推断时间
1302
用相同类型的第二层和第三层代替NN的第一层，以使得第三层的大小等于第一层的大小，并且第二层的大小（r）被选择以使得（M）和/或（T）减小
1304
初始化代替层的参数：
（W1）从第一层的输入到第二层的权重矩阵，
（W2）从第二层到第三层的权重矩阵，
（b）偏置向量
1306
向第二层中插入非线性
1308
用训练示例微调W1和W2
1310
微调NN中所有层的权重（全栈反向传播）
1312
表征新NN的P，
P > 阈值？
1314
否
是
增大（r）的值
1316
结束
1318

图13