(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 107533665 A
(43)申请公布日 2018.01.02

(21)申请号 201680024211.6

(22)申请日 2016.03.11

(30)优先权数据
62/154,097 2015.04.28 US
14/848,288 2015.09.08 US

(85)PCT国际申请进入国家阶段日
2017.10.26

(86)PCT国际申请的申请数据
PCT/US2016/022158 2016.03.11

(87)PCT国际申请的公布数据
WO2016/175925 EN 2016.11.03

(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 R·B·托瓦

(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 李小芳 袁逸

(51)Int.Cl.
G06N 3/04(2006.01)
G06N 3/08(2006.01)
G06N 7/00(2006.01)

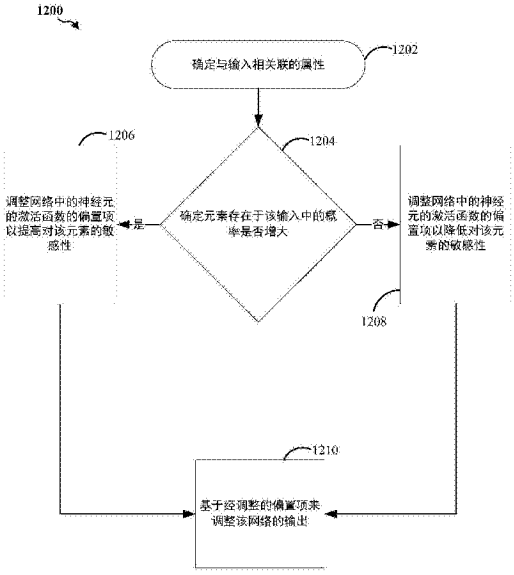
权利要求书2页 说明书16页 附图13页

(54)发明名称

经由偏置项在深度神经网络中纳入自顶向下信息

(57)摘要

一种对深度神经网络进行偏置的方法包括确定元素存在于去往该网络的输入中的概率是否增大。该方法还包括调整该网络中的神经元的激活函数的偏置以提高对该元素的敏感性。在一个配置中，该偏置是在不调整网络权重的情况下被调整的。该方法进一步包括基于该偏置来调整该网络的输出。



CN 107533665 A

1.一种对深度神经网络进行偏置的方法,包括:

确定元素存在于去往所述网络的输入中的概率是否增大;

调整所述网络中的神经元的激活函数的偏置以提高对所述元素的敏感性,所述偏置是在不调整所述网络的权重的情况下被调整的;以及

至少部分地基于所述偏置来调整所述网络的输出。

2.如权利要求1所述的方法,其特征在于,进一步包括将所述偏置的调整量确定为常数、突触权重的函数、或响应于目标类呈现的激活的函数。

3.如权利要求1所述的方法,其特征在于,调整所述偏置包括缩放所述偏置。

4.如权利要求3所述的方法,其特征在于,所述调整的范围至少部分地基于什么很可能存在于所述输入中和/或什么很可能不存在于所述输入中的先验知识。

5.如权利要求1所述的方法,其特征在于,所述调整是在所述网络的内部层级处执行的。

6.一种用于对深度神经网络进行偏置的装备,包括:

用于确定元素存在于去往所述网络的输入中的概率是否增大的装置;

用于调整所述网络中的神经元的激活函数的偏置以提高对所述元素的敏感性的装置,所述偏置是在不调整所述网络的权重的情况下被调整的;以及

用于至少部分地基于所述偏置来调整所述网络的输出的装置。

7.如权利要求6所述的装备,其特征在于,进一步包括用于将所述偏置的调整量确定为常数、突触权重的函数、或响应于目标类呈现的激活的函数的装置。

8.如权利要求6所述的装备,其特征在于,所述用于调整所述偏置的装置包括用于缩放所述偏置的装置。

9.如权利要求8所述的装备,其特征在于,所述调整的范围至少部分地基于什么很可能存在于所述输入中和/或什么很可能不存在于所述输入中的先验知识。

10.如权利要求6所述的装备,其特征在于,所述用于调整的装置是在所述网络的内部层级处被执行的。

11.一种用于对深度神经网络进行偏置的装置,包括:

存储器单元;以及

耦合至所述存储器单元的至少一个处理器,所述至少一个处理器被配置成:

确定元素存在于去往所述网络的输入中的概率是否增大;

调整所述网络中的神经元的激活函数的偏置以提高对所述元素的敏感性,所述偏置是在不调整所述网络的权重的情况下被调整的;以及

至少部分地基于所述偏置来调整所述网络的输出。

12.如权利要求11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将所述偏置的调整量确定为常数、突触权重的函数、或响应于目标类呈现的激活的函数。

13.如权利要求11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缩放所述偏置来调整所述偏置。

14.如权利要求13所述的装置,其特征在于,所述调整的范围至少部分地基于什么很可能存在于所述输入中和/或什么很可能不存在于所述输入中的先验知识。

15.如权利要求11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所

述网络的内部层级处调整所述偏置。

16.一种其上记录有程序代码的非瞬态计算机可读介质，所述程序代码由处理器执行并且包括：

用于确定元素存在于去往网络的输入中的概率是否增大的程序代码；

用于调整所述网络中的神经元的激活函数的偏置以提高对所述元素的敏感性的程序代码，所述偏置是在不调整所述网络的权重的情况下被调整的；以及

用于至少部分地基于所述偏置来调整所述网络的输出的程序代码。

17.如权利要求16所述的计算机可读介质，其特征在于，所述程序代码进一步包括用于将所述偏置的调整量确定为常数、突触权重的函数、或响应于目标类呈现的激活的函数的程序代码。

18.如权利要求16所述的计算机可读介质，其特征在于，所述用于调整所述偏置的程序代码包括用于通过缩放所述偏置来调整所述偏置的程序代码。

19.如权利要求18所述的计算机可读介质，其特征在于，所述调整的范围至少部分地基于什么很可能存在于所述输入中和/或什么很可能不存在于所述输入中的先验知识。

20.如权利要求16所述的计算机可读介质，其特征在于，所述用于调整所述偏置的程序代码包括用于在所述网络的内部层级处调整所述偏置的程序代码。

经由偏置项在深度神经网络中纳入自顶向下信息

[0001] 相关申请的交叉引用

[0002] 本申请要求于2015年4月28日提交且题为“Incorporating top-down information in deep neural networks via the bias term(经由偏置项在深度神经网络中纳入自顶向下信息)”的美国临时专利申请No.62/154,097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,尤其涉及用于基于元素存在于去往网络的输入中的概率是否增大来调整该网络中的神经元的激活函数的偏置项以提高对该元素的敏感性的系统和方法。

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,其中每一个神经元具有感受野并且共同地拼出一输入空间。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因此它们被越来越多地用于对象识别应用。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。支持向量机包括对数据进行归类的分离超平面(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裕量。换言之,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] 在本公开的一个方面,公开了一种对深度神经网络进行偏置的方法。该方法包括确定元素存在于去往该网络的输入中的概率是否增大。该方法还包括调整该网络中的神经元的激活函数的偏置以提高对该元素的敏感性。在一个配置中,该偏置是在不调整网络权重的情况下被调整的。该方法进一步包括至少部分地基于该偏置来调整该网络的输出。

[0013] 本公开的另一方面涉及一种装备,其包括用于确定元素存在于去往网络的输入中的概率是否增大的装置。该装备还包括用于调整该网络中的神经元的激活函数的偏置以提

高对该元素的敏感性的装置。在一个配置中，该偏置是在不调整网络权重的情况下被调整的。该装备进一步包括用于至少部分地基于该偏置来调整该网络的输出的装置。

[0014] 在本公开的另一方面，公开了一种用于对深度神经网络进行偏置的计算机程序产品。该计算机程序产品具有其上记录有非瞬态程序代码的非瞬态计算机可读介质。该程序代码由处理器执行并且包括用于确定元素存在于去往该网络的输入中的概率是否增大的程序代码。该程序代码还包括用于调整该网络中的神经元的激活函数的偏置以提高对该元素的敏感性的程序代码。在一个配置中，该偏置是在不调整网络权重的情况下被调整的。该程序代码进一步包括用于至少部分地基于该偏置来调整该网络的输出的程序代码。

[0015] 本公开的另一方面涉及一种用于对深度神经网络进行偏置的装置，该装置具有存储器单元和耦合至该存储器的一个或多个处理器。(诸)处理器被配置成确定元素存在于去往该网络的输入中的概率是否增大。(诸)处理器还被配置成调整该网络中的神经元的激活函数的偏置以提高对该元素的敏感性。在一个配置中，该偏置是在不调整网络权重的情况下被调整的。(诸)处理器被进一步配置成至少部分地基于该偏置来调整该网络的输出。

[0016] 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。

[0017] 附图简述

[0018] 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。

[0019] 图1解说了根据本公开的某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。

[0020] 图2解说了根据本公开的各方面的系统的示例实现。

[0021] 图3A是解说根据本公开的各方面的神经网络的示图。

[0022] 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。

[0023] 图4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。

[0024] 图5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。

[0025] 图6是解说神经分类器网络的图像、过滤器和神经元的示图。

[0026] 图7和8解说了根据本公开的诸方面的神经分类器网络的证据输入和激活输出的图表的示例。

[0027] 图9是解说根据本公开的诸方面的神经分类器网络的过滤器和神经元的示图。

[0028] 图10是解说根据本公开的诸方面的神经分类器网络的图像、过滤器和神经元的示图。

[0029] 图11和12是根据本公开的诸方面的调整神经分类器网络中的偏置的方法的流程图。

[0030] 详细描述

[0031] 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。

[0032] 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。

[0033] 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。

[0034] 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。

[0035] 在常规系统中，过滤器可被指定成修改或增强图像。另外，过滤器可被用来确定图像的一部分中是否存在特定元素。例如，过滤器可确定在图像的3x3像素部分中是否存在横线。因此，通过应用各种类型的过滤器，系统可确定图像中是否存在特定对象。相应地，过滤可被用来促成对图像进行分类。

[0036] 卷积可被指定用于图像的线性过滤。卷积输出是输入像素的加权和。权重矩阵可被称为卷积内核或过滤器。卷积可通过线性化图像和线性化过滤器的矩阵相乘来获得。

[0037] 在常规系统中，图像可基于图像的像素来分类。然而，在一些情形中，可能存在对象将存在于图像中或存在于图像中的概率增大的先验知识。本公开的诸方面涉及将网络偏置成趋向于基于对象将存在于图像中或存在于图像中的概率增大的先验知识来对该对象进行分类。

[0038] 图1解说了根据本公开的某些方面使用片上系统(SOC)100进行前述网络偏置的示例实现，SOC 100可包括通用处理器(CPU)或多核通用处理器(CPU)102。各变量(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，带权重的神经网络)、延迟、频率槽信息、以及任务信息可以被存储在与神经处理单元(NPU)108相关联的存储器块中或存储在专用存储器块118中。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。

[0039] SOC 100还可包括为具体功能定制的附加处理块(诸如图形处理单元(GPU)104、数字信号处理器(DSP)106、连通性块110(其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))以及例如可检测和识别姿势的多媒体处理器112。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)。SOC可基于ARM指令集。

[0040] SOC 100还可包括为具体功能定制的附加处理块(诸如GPU 104、DSP 106、连通性

块110(其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))以及例如可检测和识别姿势的多媒体处理器112。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)。

[0041] SOC 100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确定元素存在于去往网络的输入中的概率是否增大的代码。加载到通用处理器102中的指令还可包括用于调整该网络中的神经元的激活函数的偏置以提高对该元素的敏感性的代码。在一个配置中,该偏置是在不调整该网络的权重的情况下被调整的。加载到通用处理器102中的指令可进一步包括用于基于偏置来调整网络输出的代码。

[0042] 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可具有可执行本文所描述的方法的各种操作的多个局部处理单元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。

[0043] 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以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。此外,深度网络可以学习以表示和识别人类可能还没有考虑过的新类型的特征。

[0044] 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,则第一层可学习以识别输入流中的相对简单的特征(诸如边)。在另一示例中,如果向第一层呈递听觉数据,则第一层可学习以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以识别特征组合,诸如对于视觉数据识别简单形状或对于听觉数据识别声音组合。例如,更高层可学习以表示视觉数据中的复杂形状或听觉数据中的词语。再高层可学习以识别常见视觉对象或口语短语。

[0045] 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以识别轮子、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。

[0046] 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可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。

[0047] 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。

[0048] 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以识别车道标记,而与图像上部相关联的神经元可学习以识别交通信号灯、交通标志等。

[0049] 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(诸如限速标志的经剪裁图像),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,例如与经训练的网络300的输出322中所示的“标志”和“60”对应的那些神经元。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。

[0050] 为了调整权重,学习算法可为权重计算梯度向量。该梯度可指示在权重被略微调整情况下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)”。

[0051] 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。

[0052] 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。

[0053] 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,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。

[0054] 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。

[0055] 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。

[0056] 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。

[0057] 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普遍化。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。

[0058] 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过滤器,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,而是,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过滤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。

[0059] 例如,深度卷积网络的平行过滤器组可任选地基于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,平行过滤器组可被加载到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。

[0060] 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。

[0061] 图4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。

[0062] 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。

[0063] 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。

[0064] 图5是解说智能手机502上的AI应用的运行时操作500的框图。AI应用可包括预处理模块504，该预处理模块504可被配置(例如，使用JAVA编程语言被配置)成转换图像506的格式并随后对该图像进行剪裁和/或调整大小(508)。经预处理的图像可接着被传达给分类应用510，该分类应用510包含场景检测后端引擎512，该场景检测后端引擎512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512可被配置成进一步通过缩放(516)和剪裁(518)来预处理(514)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520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(522)并被传递通过分类应用510中的指数平滑块524。经平滑的结果可接着使得智能手机502的设置和/或显示改变。

[0065] 在一个配置中，机器学习模型(诸如神经网络)被配置成用于：确定元素存在于去往网络的输入中的概率是否增大；调整该网络中的神经元的激活函数的偏置以提高对该元素的敏感性；以及至少部分地基于该偏置来调整该网络的输出。该模型包括确定装置和/或调整装置。在一个方面，确定装置和/或调整装置可以是配置成执行所叙述功能的通用处理器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。在另一种配置中，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装置。

[0066] 根据本公开的某些方面，每个局部处理单元202可被配置成基于模型的一个或多个期望功能特征来确定模型的参数，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特征朝着期望的功能特征发展。

[0067] 经由偏置项在深度神经网络中纳入自顶向下信息

[0068] 如先前所讨论的，可能具有关于对象将存在于图像中或存在于图像中的概率增大的先验知识。例如，图像的时间/位置可以提供关于可能存在于该图像中的对象的信息。即，在一个示例中，如果图像是在橄榄球比赛中拍摄的，则该图像中存在橄榄球、草地、和/或头

盔的概率增大。作为另一示例，对象存在于图像中的概率可基于图像中其他对象的存在而增大。例如，滑雪板的图像包括雪的概率增大。

[0069] 尽管本公开的诸方面是关于确定图像中的对象来描述的，但本公开的诸方面并不限于确定图像中的对象。当然，本公开的诸方面还被构想用于确定任何元素是否存在或存在于去往网络的输入中的概率是否增大。例如，本公开的诸方面可被用于确定特定声音是否存在于音频输入中。

[0070] 在一个配置中，将网络偏置成趋向于基于对象将存在于图像中或存在于图像中的概率增大的先验知识来对该对象进行分类。可以指定该偏置以防止误报。即，本公开的诸方面对偏置进行缩放以放大对图像中所检测到的对象的响应，而不是基于对象存在的概率来提高分类器神经元的输出。

[0071] 图6解说了图像600和可被应用于图像600的过滤器602-608的示例。如图6所示，图像600是踢球比赛的图像。在此示例中，该图像包括绿草610、红球612、蓝队队员614、以及紫队队员616。过滤器包括对横线进行过滤的横向过滤器602、对纵线进行过滤的纵向过滤器604、对绿色对象进行过滤的绿色过滤器606、以及对红色/紫色对象进行过滤的红色/紫色过滤器608。图6的过滤器是示例性过滤器。由于本公开的诸方面是针对要被应用于输入的各种过滤器来构想的，因此本公开的诸方面并不限于图6的过滤器。

[0072] 在本示例中，在将过滤器602-608应用于图像600之后，网络的输出可为：

[0073] 1．0.24-球

[0074] 2．0.60-蓝队

[0075] 3．0.15-紫队

[0076] 4．0.01-树

[0077] 该输出指的是基于从输入导出的证据所确定的对象存在于输入中的概率。在此示例中，球存在于图像中有24％的概率，蓝队队员存在于图像中有60％的概率，紫队队员存在于图像中有15％的概率，并且树存在于图像中有1％的概率。

[0078] 如图6所示，每个过滤器602-608具有去往与特定对象(例如，类)相关联的分类器神经元的输入。在此示例中，出于解说目的，粗线指示来自过滤器的强输出，而细线指示来自过滤器的弱输出。随着关于对象存在的证据量增加，来自过滤器的输出强度增大。例如，基于过滤器确定有红色对象存在于图像中的证据，从红色/紫色过滤器608至红球神经元618的输出为强。

[0079] 然而，如图6所示，由于纵向过滤器604未找到紫队616的任何证据，因此从纵向过滤器604至紫队神经元620的输出为弱。如先前所讨论的，纵向过滤器604确定纵线是否存在于图像中。即，纵向过滤器604不针对与紫队616相关联的特征(诸如穿紫色衬衣的人)进行过滤。由此，由于紫队队员616与纵线不相关联，因此纵向过滤器604与紫队神经元620之间存在弱连接。

[0080] 根据本公开的诸方面，网络元素之间(诸如过滤器和神经元)的连接可被称为突触。此外，分类器神经元可被称为输出神经元和/或对象神经元。分类器神经元、输出神经元和对象神经元指的是基于来自过滤器的输入来从激活函数输出值的神经元。

[0081] 如先前所讨论的，图像600包括红球612和穿紫色衬衣的个体(例如，紫队队员616)。然而，在图像600中，与其他对象相比，红球612相对较小。而且，在图像600中，穿紫色

衬衣的个体不如其他对象(诸如穿蓝色衬衣的群体)那样多。因此,红球612和穿紫色衬衣的个体可能基于网络输出而被漏掉或假定不存在。

[0082] 然而,在本配置中,分类被指定以确定紫队队员616是否存在于图像中。在常规系统中,基于图像600是蓝队614与紫队616踢球的图像的先验知识,可基于图像600包含紫队队员616的概率来增大对紫队分类器神经元(例如,紫队神经元620)的响应。然而,紫队队员616有可能不存在于图像中。因此,基于图像600包含紫队队员616的概率来增大对紫队神经元620的响应(例如,激活值输出)可能导致误报。

[0083] 由此,除了防止误报以外,期望缓解与其他对象相比相对较小和/或不如其他对象那样多的对象的不正确或弱分类。根据本公开的诸方面,基于对象将存在于图像中或对象存在于图像中的概率增大的先验知识,可以调整激活函数的偏置以使得过滤器的输出基于该偏置而有所调整。在一个配置中,可基于对象存在于图像中的概率来调整至分类器神经元的突触的偏置。作为示例,可基于紫队队员616存在于图像中的概率来调整至紫队神经元620的突触622的偏置。

[0084] 在一些情形中,可能不期望基于对象将存在于图像中或对象存在于图像中的概率增大的先验知识来调整过滤器的权重以更改网络输出。具体地,各过滤器的权重已从众多训练轮中确定。因此,在训练之后调整权重可能更改训练结果并导致错误值。

[0085] 另外,直接改变激活值可导致网络对不存在的对象(例如,幻象)进行分类。因此,在一个配置中,偏置项被缩放以放大很可能指示对象存在的响应。即,在本配置中,对偏置进行缩放将激活函数的工作范围改变成对输入更敏感。式1示出了关于激活函数的方程。

[0086] $$激活=f(\sum_i w_i x_i+\gamma b_i) \quad (1)$$

[0087] 在式1中,w_i是权重,x_i是来自较低层(诸如过滤器)的激活值输出,且γb_i是偏置项。具体地,γ是偏置的调整量且b_i是偏置。根据式1,可以缩放通往特定分类器神经元的所有突触的偏置项。即,可基于偏置来增大或减小去往分类器神经元的输入的增益。

[0088] 图7解说了具有在x轴上的去往分类器神经元的输入($\sum_i w_i x_i+\gamma b_i$)以及在y轴上的从分类器神经元输出的激活函数值(式1)的坐标图700。该激活函数值可被称为激活值,且去往分类器神经元的输入可被称为证据输入。x轴上的证据输入是关于对象存在的证据量的值。在此示例中,输入值范围从-10到10,以使得值-10指示几乎没有关于对象存在于图像中的证据且10指示有大量关于对象存在的证据。此外,激活值是对象存在于图像中的概率,其基于关于该对象存在于该图像中的证据量(例如,x轴输入)。由此,如图7所示,激活值随去往分类器神经元的证据输入增加而增大。即,去往分类器神经元的强证据输入导致强激活值输出。

[0089] 另外,图7解说了图表700上所标绘的众多线。这些线指示调整输入偏置的结果。例如,第一线702指示输入和激活的基线(例如,无偏置调整)。在此示例中,如第一线702所示,证据输入0导致约0.5的激活值。另外,第二线704提供将偏置调整1.5的示例。如第二线704所示,证据输入0导致约0.9的激活值。

[0090] 相应地,如图7所示,尽管第一线702和第二线704接收到相同的证据输入值,但从分类器神经元输出的激活值基于经缩放偏置而有所调整。

[0091] 应当注意,该偏置可以正向调整或负向调整。例如,图7解说了正向调整和负向调整两者。第二线704标绘了将偏置调整1.5的坐标。第三线706标绘了将偏置调整-1.5的坐

标。

[0092] 如先前所讨论的，可基于项目存在于输入中的先验知识来正向调整偏置。例如，由于鸟与树相关联，因此可在给出鸟的图像时正向调整针对树的偏置。此外，可基于项目不存在于输入中的先验知识来负向调整偏置。例如，由于棒球与橄榄球比赛不相关联，因此可在给出橄榄球比赛的图像时负向调整针对棒球的偏置。

[0093] 应当注意，偏置被应用于分类器神经元的每个输入。即，偏置被应用于每件证据，诸如每个过滤器的输出。例如，基于图6的示例，偏置可被应用于被输入到紫队神经元620的每个突触622。如先前所讨论的，针对对象存在所确定的值可基于过滤器类型而变化。

[0094] 例如，基于图6的示例，可以指定横向过滤器602以确定横线是否存在于图像中。相应地，由于红球几乎没有纵线，因此从横向过滤器602至红球神经元618的证据输入值较低。即，横向过滤器602几乎找不到与红色蹴球相关联的横线的证据。由此，由于偏置被应用于来自每个过滤器的输入，因此对象存在的概率基于从每个过滤器找到的关于该对象的证据量而增大。

[0095] 如先前所讨论的，基于图6的示例，具有未经调整的偏置的网络的输出可为：

[0096] 1．0.24-球

[0097] 2．0.60-蓝队

[0098] 3．0.15-紫队

[0099] 4．0.01-树

[0100] 在本配置中，基于图6的示例，基于球将存在于图像中或存在于图像中的概率增大的先验知识而针对球缩放偏置。基于为球所应用的正向偏置，网络的输出可为：

[0101] 1．0.50-球

[0102] 2．0.35-蓝队

[0103] 3．0.05-紫队

[0104] 4．0.00-树

[0105] 如以上关于针对球所调整的正向偏置提供的输出所示，与未经调整的偏置输出相比，球的概率从24％改变为50％。

[0106] 在本配置中，基于图6的示例，基于树存在于图像中的概率增大的先验知识而针对树缩放偏置。基于此配置，网络的输出可为：

[0107] 1．0.10-球

[0108] 2．0.35-蓝队

[0109] 3．0.05-紫队

[0110] 4．0.02-树

[0111] 如以上关于针对树所调整的正向偏置提供的输出所示，与未经调整的偏置输出相比，树的概率从1％改变为2％。即，由于树不存在于图6的图像600中，因此针对树缩放偏置并不会使树存在的概率显著增大。

[0112] 图8解说了具有表示从过滤器输入到分类器神经元的证据值的x轴和表示从分类器神经元输出的激活函数值的y轴的图表800。在图8中，不同曲线指示调整输入偏置的结果。例如，第一线802指示输入和激活的未经调整的基线。在此示例中，如第一线802所示，当未针对证据输入调整偏置时，输入-1导致约0.24的激活。另外，第二线804提供针对证据输

入将偏置调整0.5的示例。如第二线804所示，输入-1导致约0.5的激活。由此，如先前所讨论的，在未经调整的网络输出中，对象(诸如球)的值为0.24。此外，如以上所描述的，当为对象调整了偏置时，该值为0.5。

[0113] 另外，如图8所示，对于具有低证据值(诸如-5)的第二对象，来自第一线802的未经调整激活值为0.01。此外，第二线804提供针对第二对象的证据输入将偏置调整0.5的示例。如第二线804所示，证据输入的值-5导致约0.02的激活值。由此，如上所述，在未经调整的网络输出中，第二对象的激活值为0.01。此外，如以上所描述的，当为第二对象调整了偏置时，该激活值为0.02。如先前所讨论的，由于几乎没有关于第二对象存在的证据，因此调整证据输入的偏置将不会使激活值显著变化。

[0114] 在一个配置中，偏置是作为通往对象的权重的函数来调整的。例如，如果要调整球的偏置，则将与突触权重成比例的调整项从该球的分类器神经元反向传播。

[0115] 图9解说了在顶层(层J)处的分类器神经元连接至中间层(层I)处的因对象而异的过滤器的网络的示例900。这些分类器被连接至较低层(层H)处的通用过滤器。在一个示例中，可针对球的证据来调整偏置。由此，在此示例中，可在顶层给出调整值，以使得该调整值(γ_{ij})与该网络中的突触权重成比例地从球神经元902向该网络反向传播。在此示例中，调整值可在已知对象存在于图像中或对象存在于图像中的概率增大时被应用于顶层。

[0116] 例如，如图9所示，从球过滤器906至球神经元902的突触904的权重较高。然而，从其他因对象而异的过滤器至球神经元902的其他突触908的权重较弱。因此，与从球神经元902反向传播至层I中的其他因对象而异的过滤器的调整值相比，反向传播至球过滤器906的调整值更强。即，调整值与突触权重成比例地从层I处的每个因对象而异的过滤器反向传播至层J处的分类器神经元。

[0117] 另外，调整值基于从层I处的因对象而异的过滤器至层H处的每个通用过滤器的突触的权重来从这些因对象而异的过滤器反向传播至层H处的通用过滤器。

[0118] 用于基于每个突触的权重来确定调整值的方程如下：

[0119] $$\gamma_{ij} = \gamma_0 w_{ij} \quad \forall j \in \text{球类} \qquad (2)$$

[0120] $$\gamma_{hi} = \gamma_{ij} w_{hi} \qquad (3)$$

[0121] 在式2和3中，基于图8的示例，w_{ij}是从层J至层I的突触的权重，w_{hi}是从层H至层I的突触的权重，γ_0是在输出神经元处给出的偏置调整量，γ_{ij}是应用于从层J至层I的突触的调整值，且γ_{hi}是应用于从层H至层I的突触的调整值。

[0122] 在另一配置中，作为调整特定对象(例如，类)的偏置的替代，可针对特定特征(诸如红色对象、和/或具有圆形边缘的对象)来调整偏置。在此示例中，可能不存在关于图像中的对象的先验知识。然而，在此示例中，网络可以搜索特定对象，诸如紫色衬衫。因此，可以在网络中的任何层调整偏置。例如，基于图9，可以针对层I处的紫色图像过滤器910调整偏置且调整值可与从层H至层I的每个突触的权重成比例地反向传播至层H处的过滤器。用于将调整值反向传播至连接到层I处的过滤器的每个突触的方程如下：

[0123] $$\gamma_{ij} = 0 \qquad (4)$$

[0124] $$\gamma_{hi} = \gamma_0 w_{hi} \qquad (5)$$

[0125] 在式4和5中，基于图9的示例，w_{hi}是从层H至层I的突触的权重，γ_0是在输出神经元

处给出的偏置调整量，γ_{ij}是应用于从层J至层I的突触的调整值，且γ_{hi}是应用于从层H至层I的突触的调整值。在此配置中，$\gamma_{ij}=0$，这是因为调整从层I反向传播至层H，而不是被应用于层J并从层J反向传播。

[0126] 在另一配置中，可基于所测得的该网络对示例性图像的响应来调整偏置。例如，可将图像呈现给网络并响应于该图像来测量该网络的响应。此外，可基于该响应来调整偏置。可在该网络的内部层级处执行该调整。

[0127] 图10解说了基于所测得的对呈现给网络1000的图像1002的响应来生成偏置的示例。如图10所示，网络1000包括分类器神经元的顶层(层J)、因对象而异的过滤器的中间层(层I)、以及通用过滤器的底层(层H)。此外，如图10所示，图像1002被呈现给网络1000。在此示例中，图像1002是具有树叶背景的紫球。如图10所示，图像1002的紫球并不是作为对象神经元中的对象存在的。因此，为了确定图像1002中的对象的调整值，图像1002被呈现给网络1000以测量网络1000的响应。

[0128] 在本示例中，当图像1002被呈现给网络1000时，在诸神经元、突触、以及层处测量该网络的激活。例如，如图10所示，这些激活分布在各种过滤器、突触、以及神经元处。具体地，在此示例中，紫色过滤器1004、绿色过滤器1006、红球过滤器1008、紫色正方形过滤器1010、以及树过滤器1012是响应于图像1002而被激活的过滤器。此外，这些激活分布在诸分类器神经元处，以使得树神经元1014、紫队神经元1016和球神经元1018被激活。应当注意，在图10中，带有粗线的突触表示响应于图像1002而被激活的突触。在图10的示例中，圈相对于过滤器/神经元的大小指示激活水平，以使得较大圈表示比较小圈更强的激活。

[0129] 在确定特定对象的激活之后，可作为该激活的函数来调整偏置。例如，可将新图像呈现给网络并观察贯穿该网络的激活模式。该偏置随后与突触所连接的神经元的激活成比例地被分发给每个突触。在此示例中，偏置是自底向上调整的，以使得该偏置中的一些分布在每层处的突触之间。在此配置中，可基于下式来自底向上调整偏置：

[0130] $$\gamma_{ij}=\frac{\gamma_0}{N_{ij}\text{突触}}x_i \qquad (6)$$

[0131] $$\gamma_{hi}=\frac{\gamma_0}{N_{hi}\text{突触}}x_h \qquad (7)$$

[0132] 在式6和7中，基于图10的示例，γ_0是在输出神经元处给出的偏置调整量，γ_{ij}是应用于从层J至层I的突触的调整值，且γ_{hi}是应用于从层H至层I的突触的调整值，x_i是从层I的特定突触输出的值，且x_h是从层H的特定突触输出的值。

[0133] 在另一配置中，基于下式来从输出反向传播调整值：

[0134] $\gamma_{ij}=(\gamma_0 w_{ij})x_j$ (8)

[0135] $\gamma_{hi}=\gamma_{ij}w_{hi}$ (9)

[0136] 在式8和9中，基于图10的示例，γ_0是在输出神经元处给出的偏置调整量，γ_{ij}是应用于从层J至层I的突触的调整值，且γ_{hi}是用于从层H至层I的突触的调整值，x_j是层J处的激活模式，w_{hi}是从层H至层I的突触的权重，且w_{ij}是从层I至层J的突触的权重。

[0137] 基于本公开的诸方面，给出了用于调整偏置的多个配置。在一个配置中，可作为常数来调整偏置。可在从知识图类型源确定自顶向下信号时将作为常数来调整偏置。例如，可

在已知鸟的图像包含树的图像的概率增大时作为常数来调整偏置。式1可被用于作为常数来调整偏置。

[0138] 在另一配置中，作为突触权重的函数来调整偏置。偏置可作为突触权重的函数被调整以使得给定对象的重要权重被偏置。附加地或替换地，偏置可作为突触权重的函数被调整以使得调整值通过网络反向传播。用于作为突触权重的函数来调整偏置的方程为：

[0139] 激活＝f(∑$_i$w$_i$x$_i$+γ(w$_i$)b$_i$) (10)

[0140] 在式10中，w$_i$是权重，()是偏置调整(例如，偏置变化)，x$_i$是从较低层输出的值，且b$_i$是偏置。

[0141] 在另一配置中，作为响应于目标类呈现的激活的函数来调整偏置。此配置可在从呈现给网络的示例导出自顶向下信号时使用。例如，如图10所示，将图像1002呈现给网络1000并基于网络中的激活分布来确定偏置。用于作为响应于目标类呈现的激活的函数来调整偏置的方程可基于下式：

[0142] 激活＝f(∑$_i$w$_i$x$_i$+γ(x$_i$)b$_i$) (11)

[0143] 在式11中，w$_i$是权重，()是偏置调整(例如，偏置变化)，x$_i$是从较低层输出的值，且b$_i$是偏置。

[0144] 此外，可以相加地或相乘地应用偏置调整。偏置的应用可取决于激活函数。

[0145] 可基于下式来相加地应用偏置调整：

[0146] 激活＝f(∑$_i$w$_i$x$_i$+(γ+b$_i$)) (12)

[0147] 在式12中，w$_i$是权重，γ是偏置调整(例如，偏置变化)，x$_i$是从较低层输出的值，且b$_i$是偏置。

[0148] 在一个配置中，基于式1来相乘地应用偏置调整。由于偏置是从原始值缩放得来的，因此相乘地应用偏置可能是期望的。

[0149] 图11解说了调整机器学习网络(诸如神经分类器网络)中的激活函数的偏置的方法1100。在框1102，网络确定元素存在于去往该网络的输入中的概率是否增大。在框1104，该网络调整该网络中的神经元的激活函数的偏置项以提高对该元素的敏感性。在一个配置中，该偏置是在不调整网络权重的情况下被调整的。此外，在框1106，该网络基于该偏置来调整该网络的输出。

[0150] 图12解说了调整机器学习网络(诸如神经分类器网络)中的激活函数的偏置的方法1200。在框1202，网络确定与输入(诸如图像)相关联的属性。作为示例，该属性可包括图像的时间、图像的位置、和/或存在于该图像中的特定对象。基于所确定的属性，在框1024，该网络确定元素存在于该输入中的概率是否增大。

[0151] 如果该元素存在于去往该网络的输入中的概率增大，则在框1206，该网络调整该网络中的神经元的激活函数的偏置项以提高对该元素的敏感性。此外，在框1210，该网络基于经调整的偏置项来调整网络输出。

[0152] 如果该元素存在于去往该网络的输入中的概率并未增大，则在框1208，该网络调整该网络中的神经元的激活函数的偏置项以降低对该元素的敏感性。此外，在框1210，该网络基于经调整的偏置项来调整网络输出。

[0153] 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(诸)软件组件和/或(诸)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。

[0154] 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。此外，“确定”可包括解析、选择、选取、确立及类似动作。

[0155] 如本文中所使用的，引述一列项目中的“至少一者”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中的至少一个”旨在涵盖：a、b、c、a-b、a-c、b-c、以及a-b-c。

[0156] 结合本公开所描述的各种解说性逻辑块、模块、以及电路可用设计成执行本文所描述的功能的通用处理器、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、或任何其他此类配置。

[0157] 结合本公开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实施。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。在替换方案中，存储介质可以被整合到处理器。

[0158] 本文所公开的方法包括用于达成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。

[0159] 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。

[0160] 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。

[0161] 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。

[0162] 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。

[0163] 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，本公开的各方面产生对处理器、计算机、机器或实现此类方面的其它系统的机能的改进。

[0164] 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。

[0165] 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。

[0166] 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。

[0167] 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。

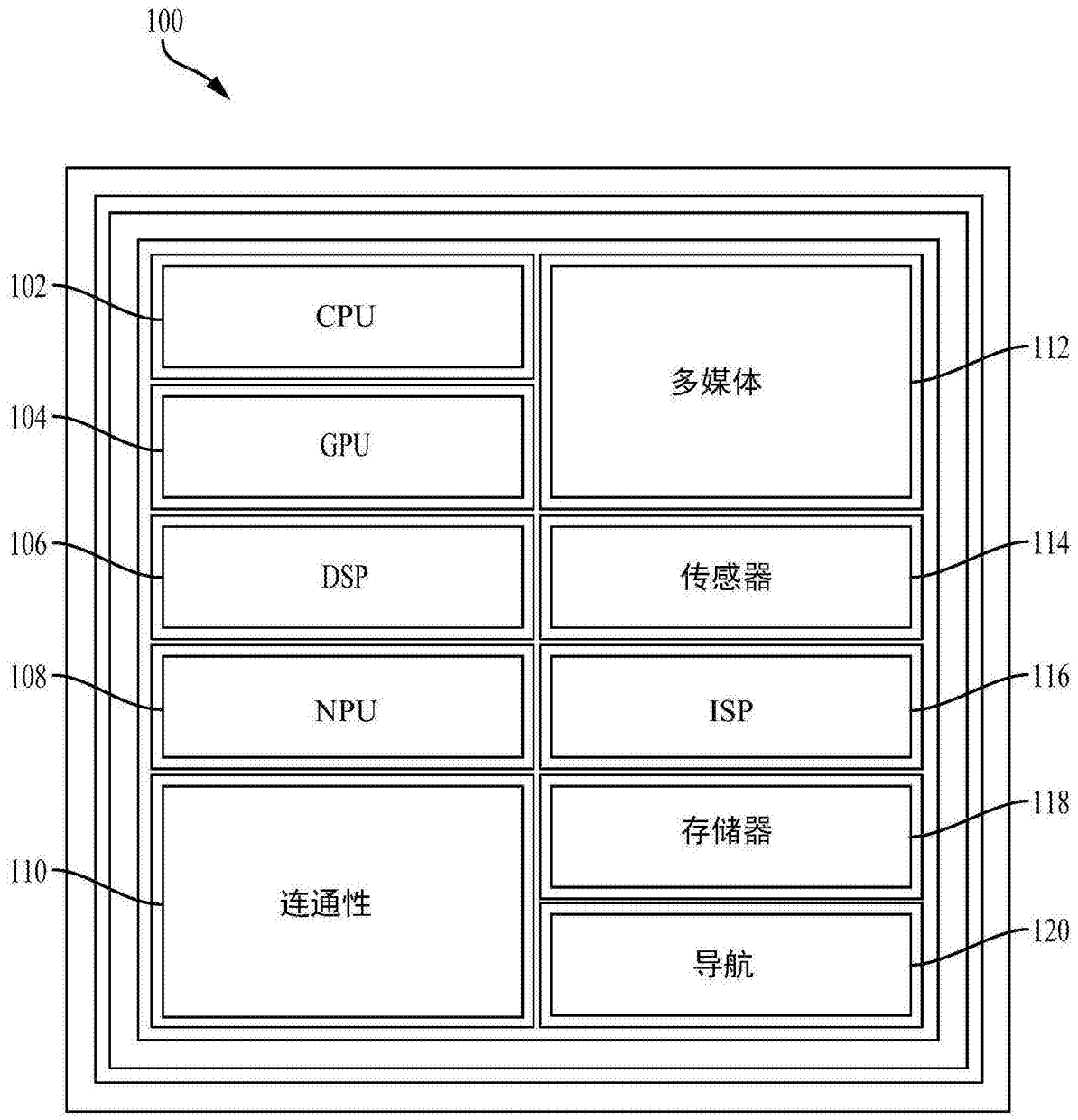

100
102
CPU
104
GPU
106
DSP
108
NPU
110
连通性
112
多媒体
114
传感器
116
ISP
118
存储器
120
导航

图1

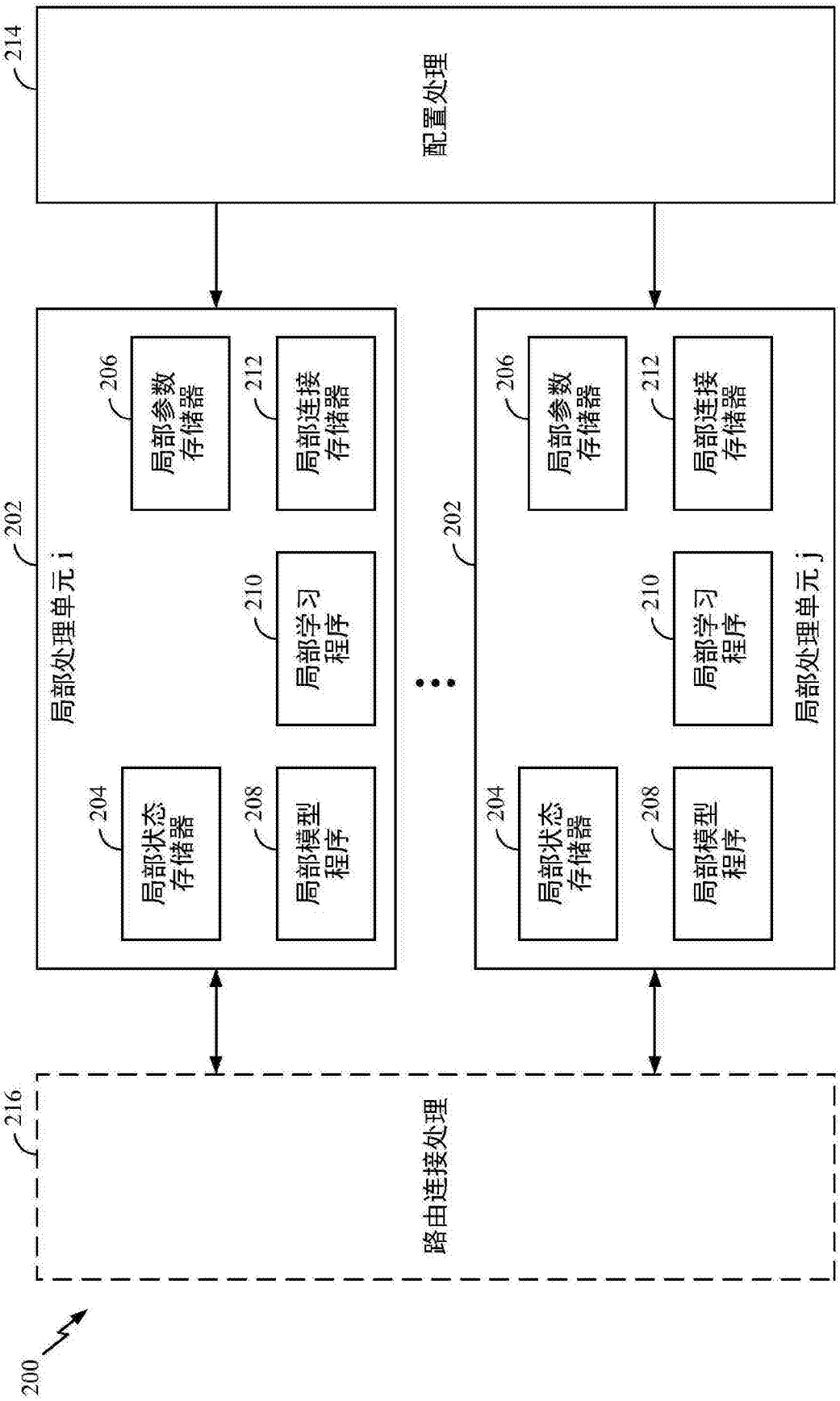

214
配置处理
202
局部处理单元 i
206
局部参数存储器
212
局部连接存储器
210
局部学习程序
204
局部状态存储器
208
局部模型程序
202
局部处理单元 j
206
局部参数存储器
212
局部连接存储器
210
局部学习程序
204
局部状态存储器
208
局部模型程序
216
路由连接处理
200

图2

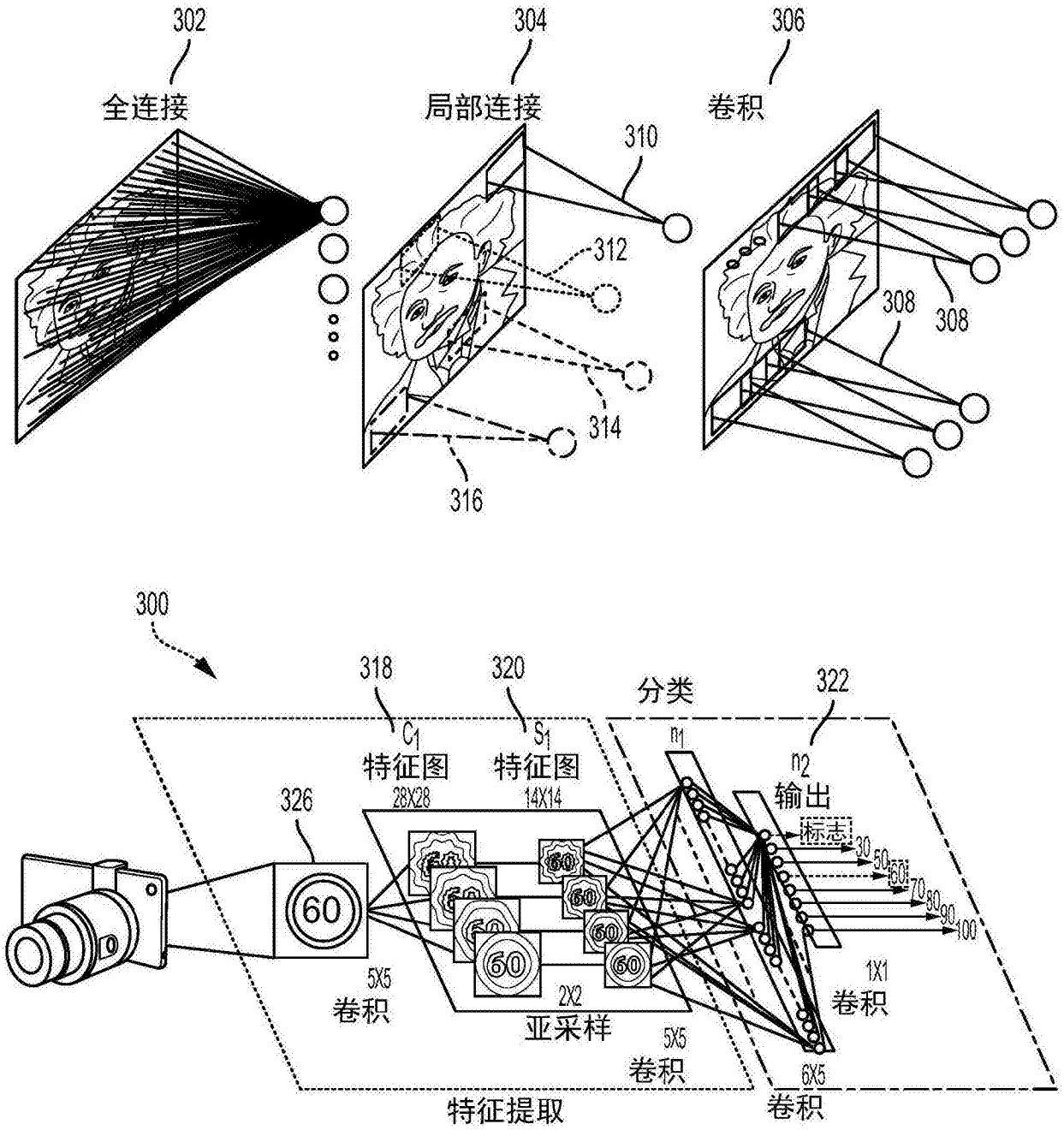

302
全连接
304
局部连接
306
卷积
310
312
314
316
308
308
300
318
320
分类
322
C1
特征图
28X28
S1
特征图
14X14
n1
n2
输出
标志
326
60
5X5
卷积
2X2
亚采样
5X5
卷积
1X1
卷积
6X5
卷积
特征提取

图3A

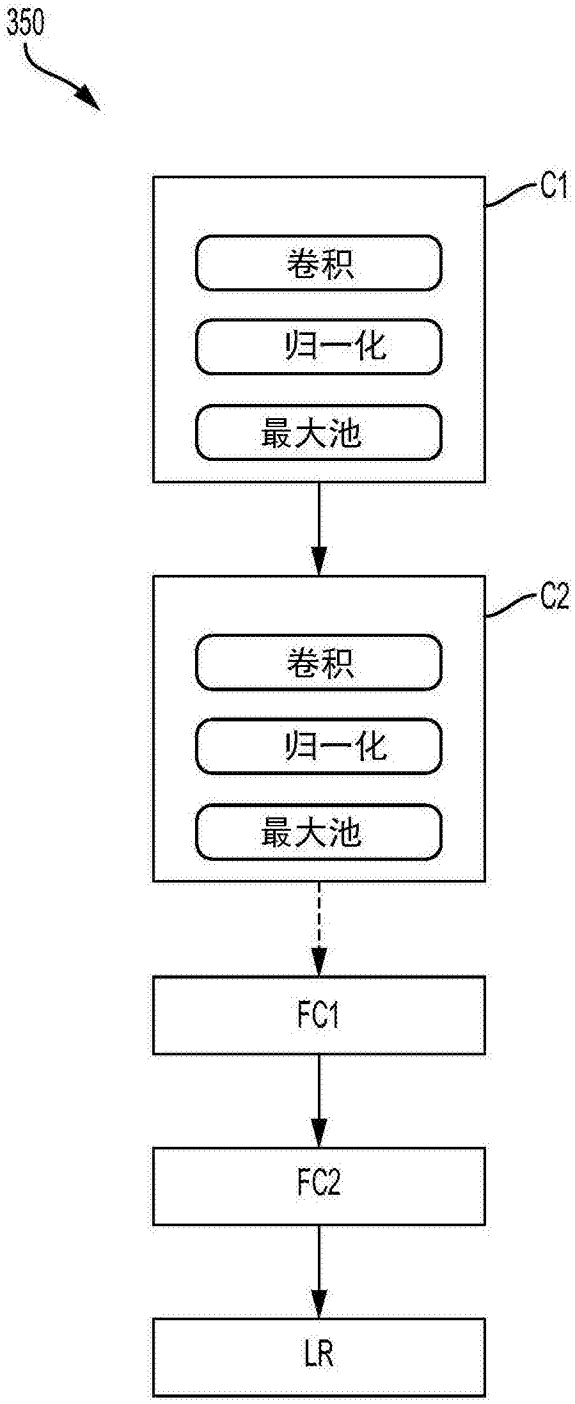


图3B

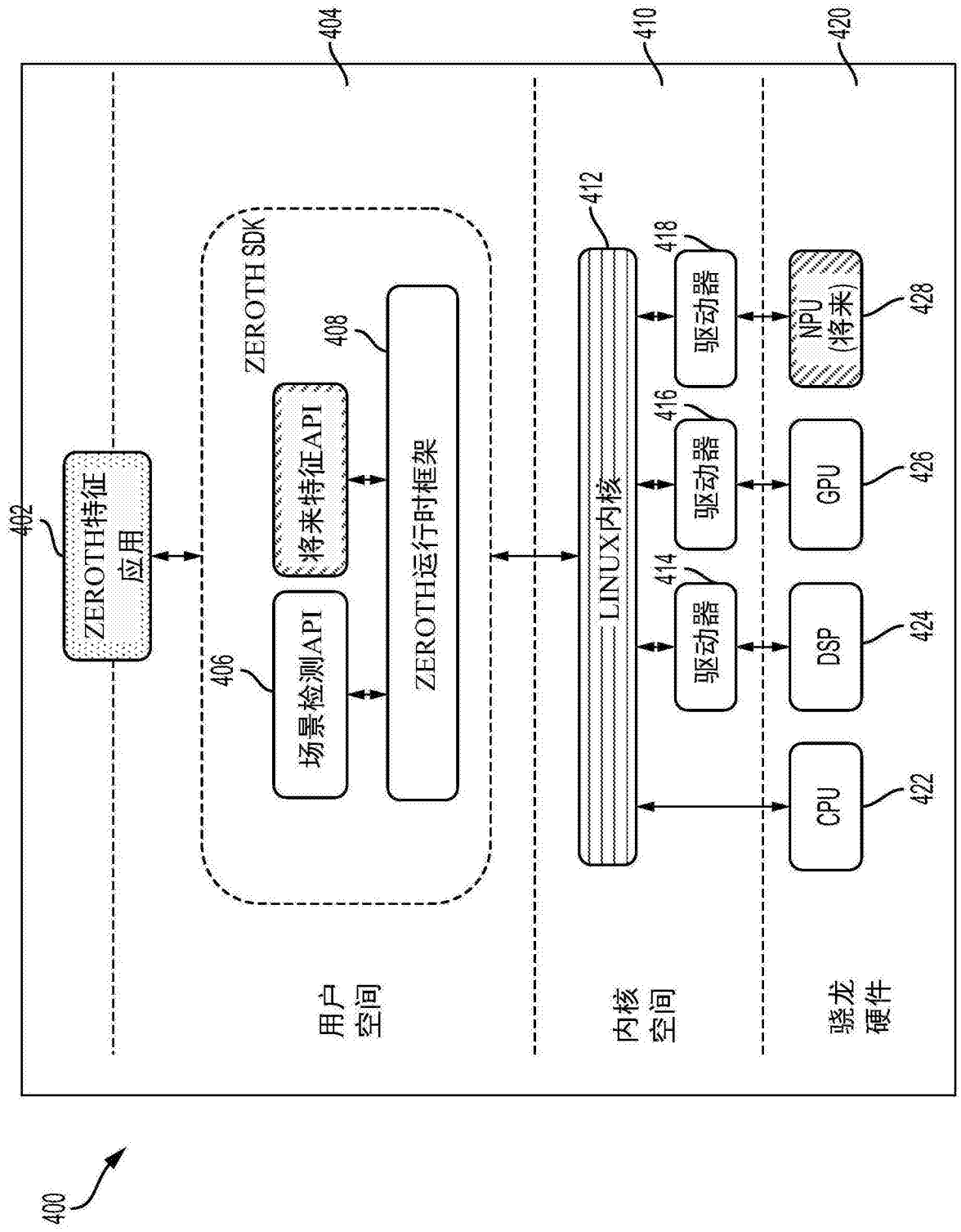

400
404
410
420
402
ZEROTH特征
应用
ZEROTH SDK
408
将来特征API
406
场景检测API
ZEROTH运行时框架
412
LINUX内核
418
驱动器
416
驱动器
414
驱动器
NPU
(将来)
428
GPU
426
DSP
424
CPU
422
用户
空间
内核
空间
骁龙
硬件

图4

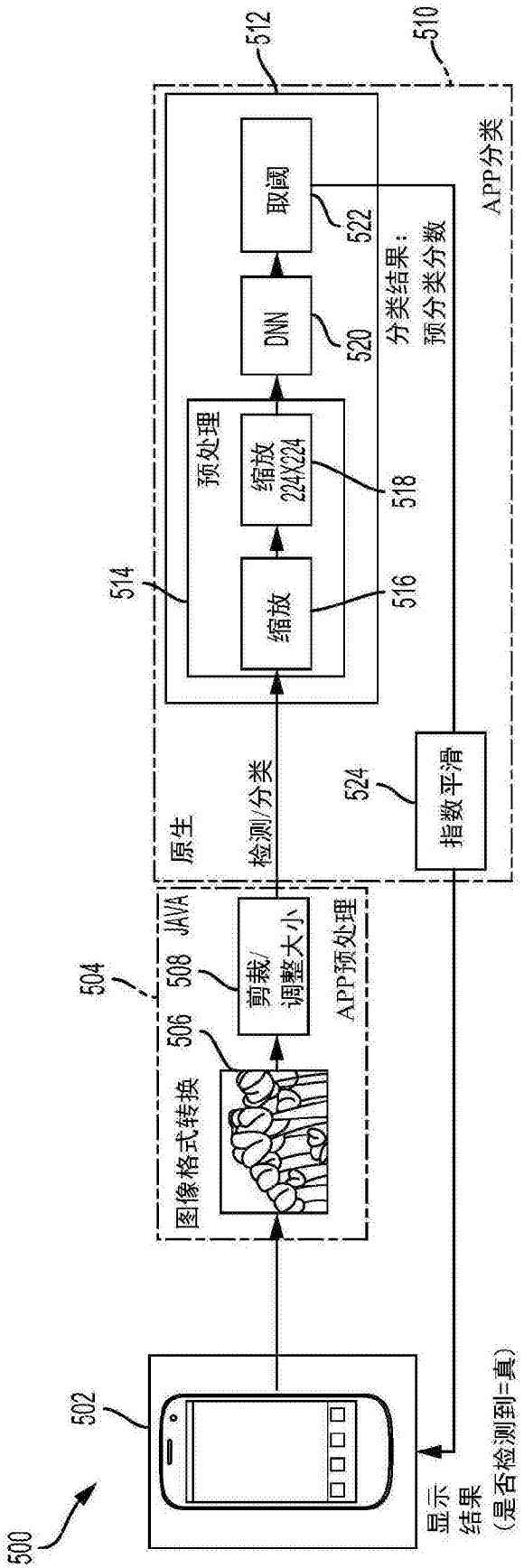

500
502
显示结果
(是否检测到=真)
504
图像格式转换
506
JAVA
508
剪裁/调整大小
APP预处理
原生
检测/分类
514
预处理
缩放
516
缩放224X224
518
DNN
520
取阈
522
分类结果：
预分类分数
APP分类
512
510
524
指数平滑

图5

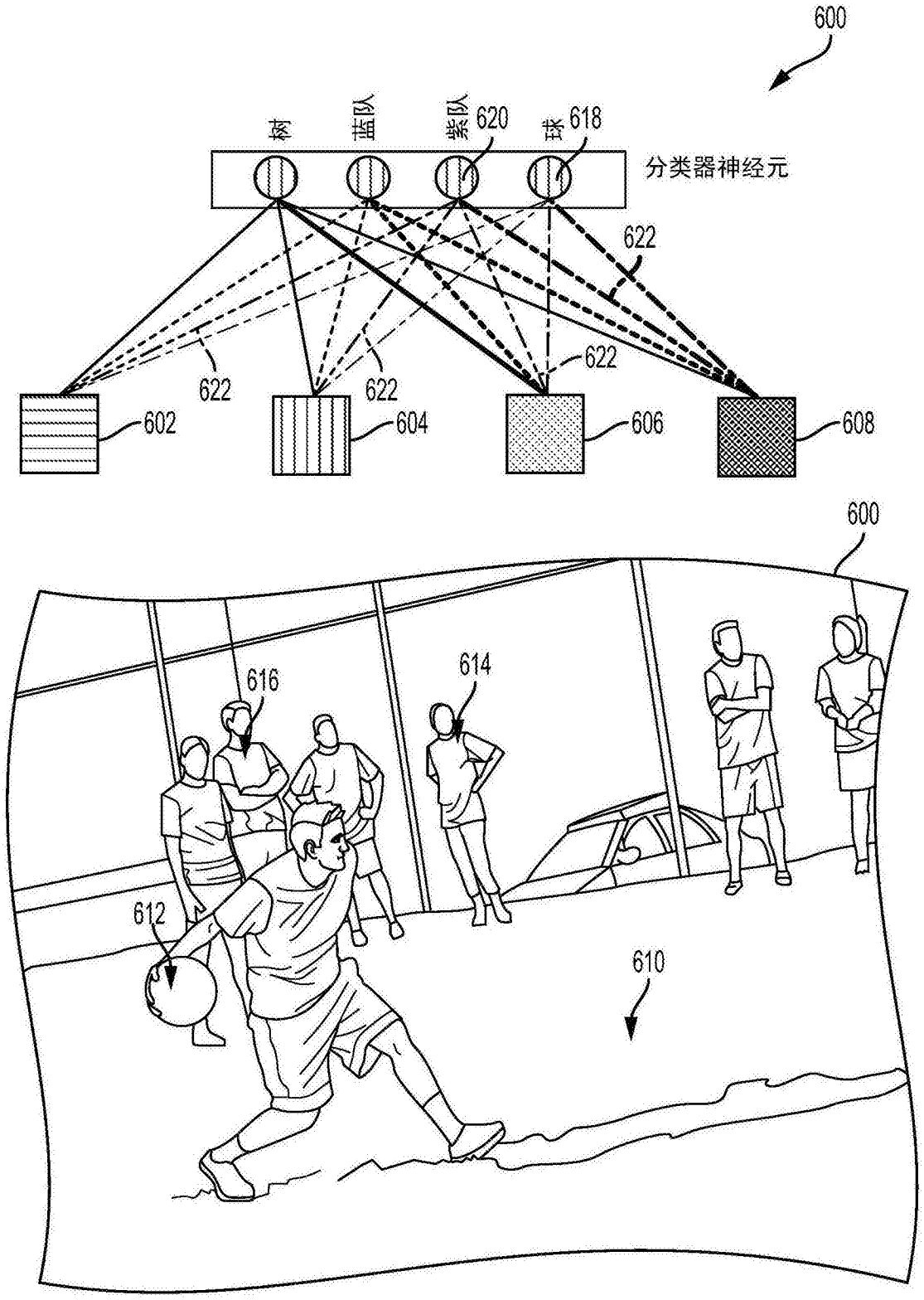

600
客
蓝队
紫队
620
618
分类器神经元
622
622
622
602
604
606
608
600
616
614
612
610

图6

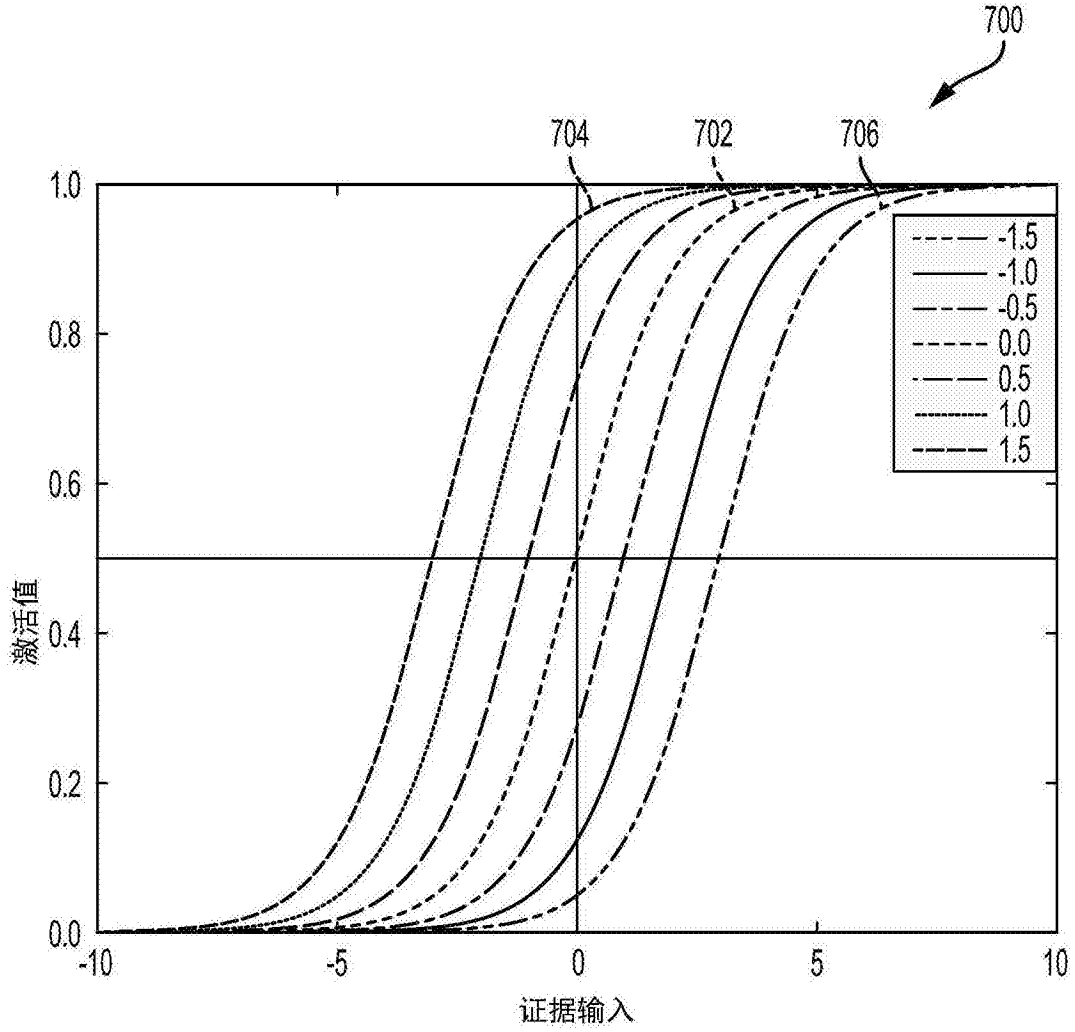

700
704
702
706
1.0
0.8
0.6
0.4
0.2
0.0
激活值
-10
-5
0
5
10
证据输入
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

图7

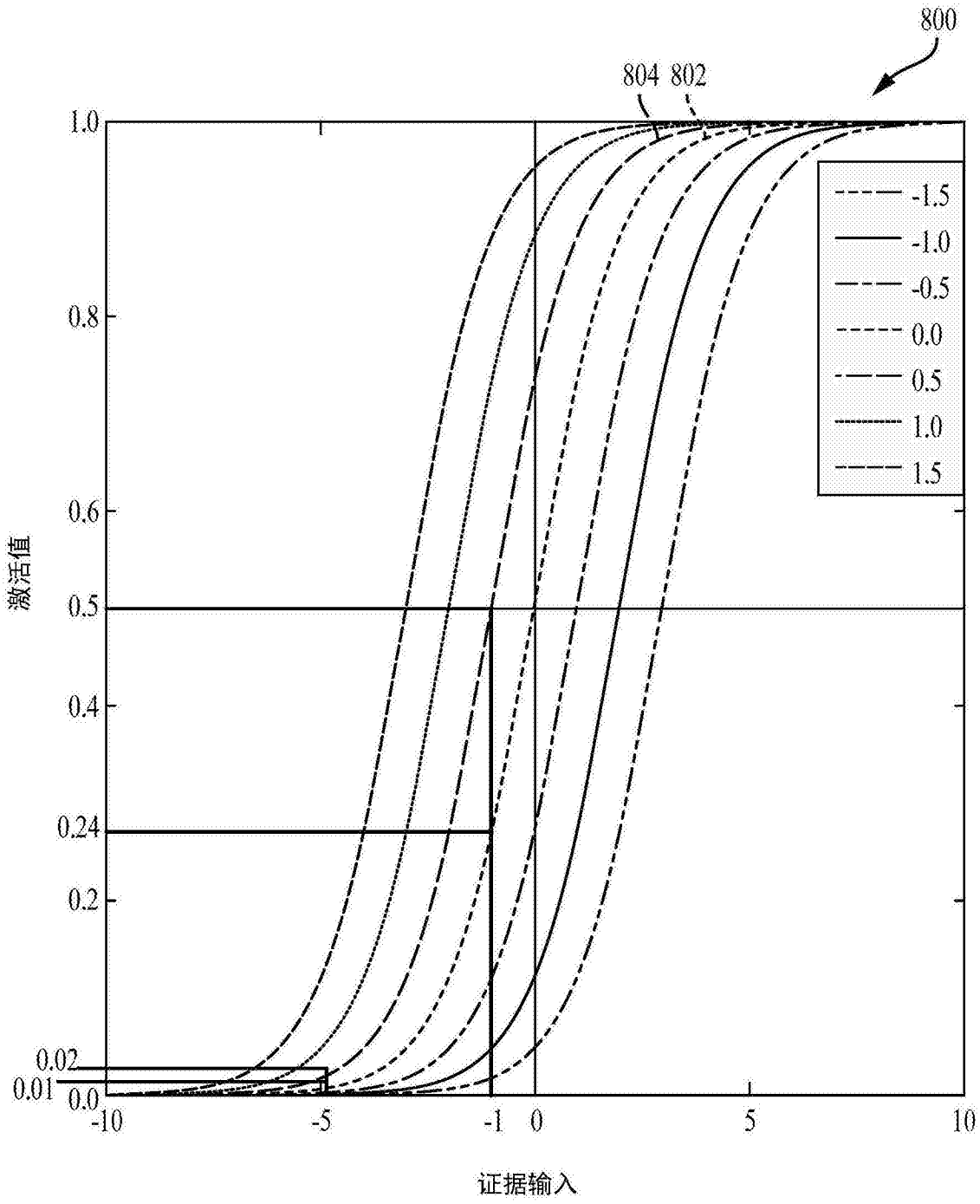

800
804 802
1.0
0.8
0.6
0.5
0.4
0.24
0.2
0.02
0.01
0.0
-10
-5
-1
0
5
10
激活值
证据输入
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

图8

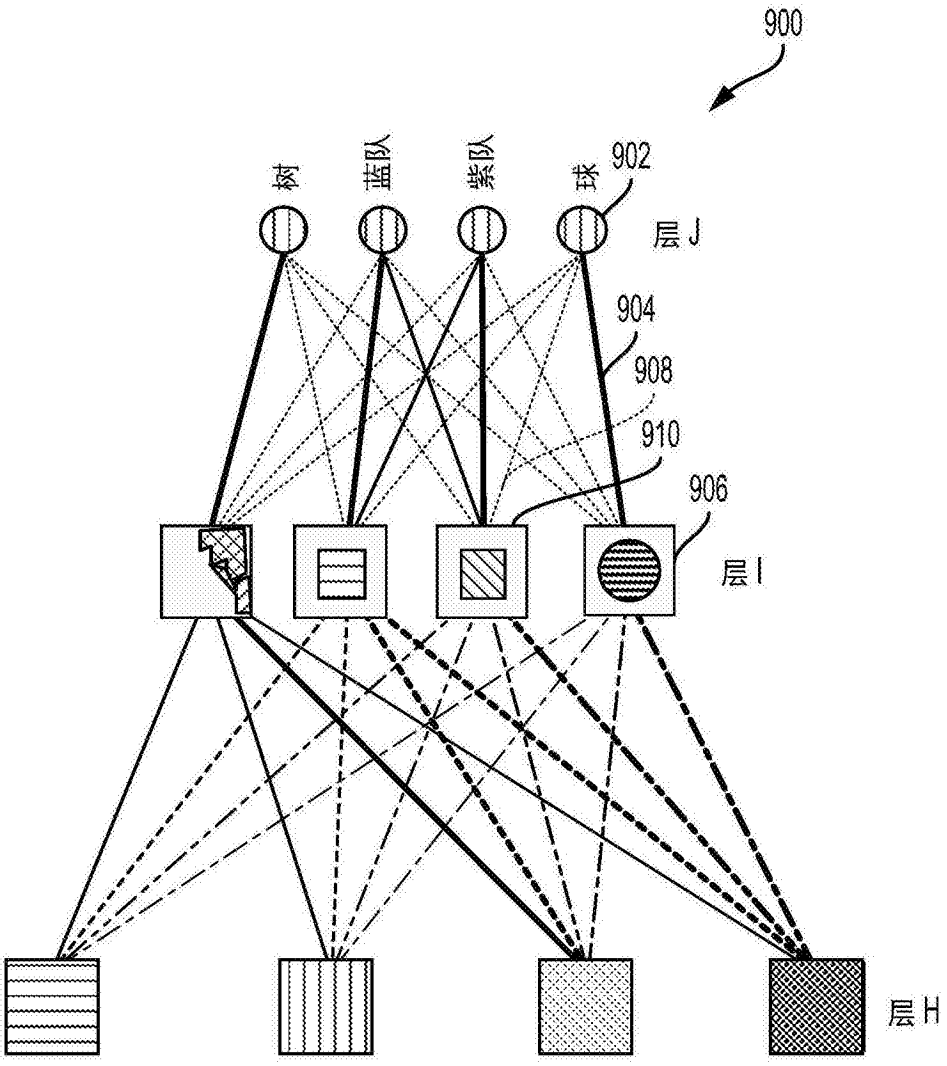


图9

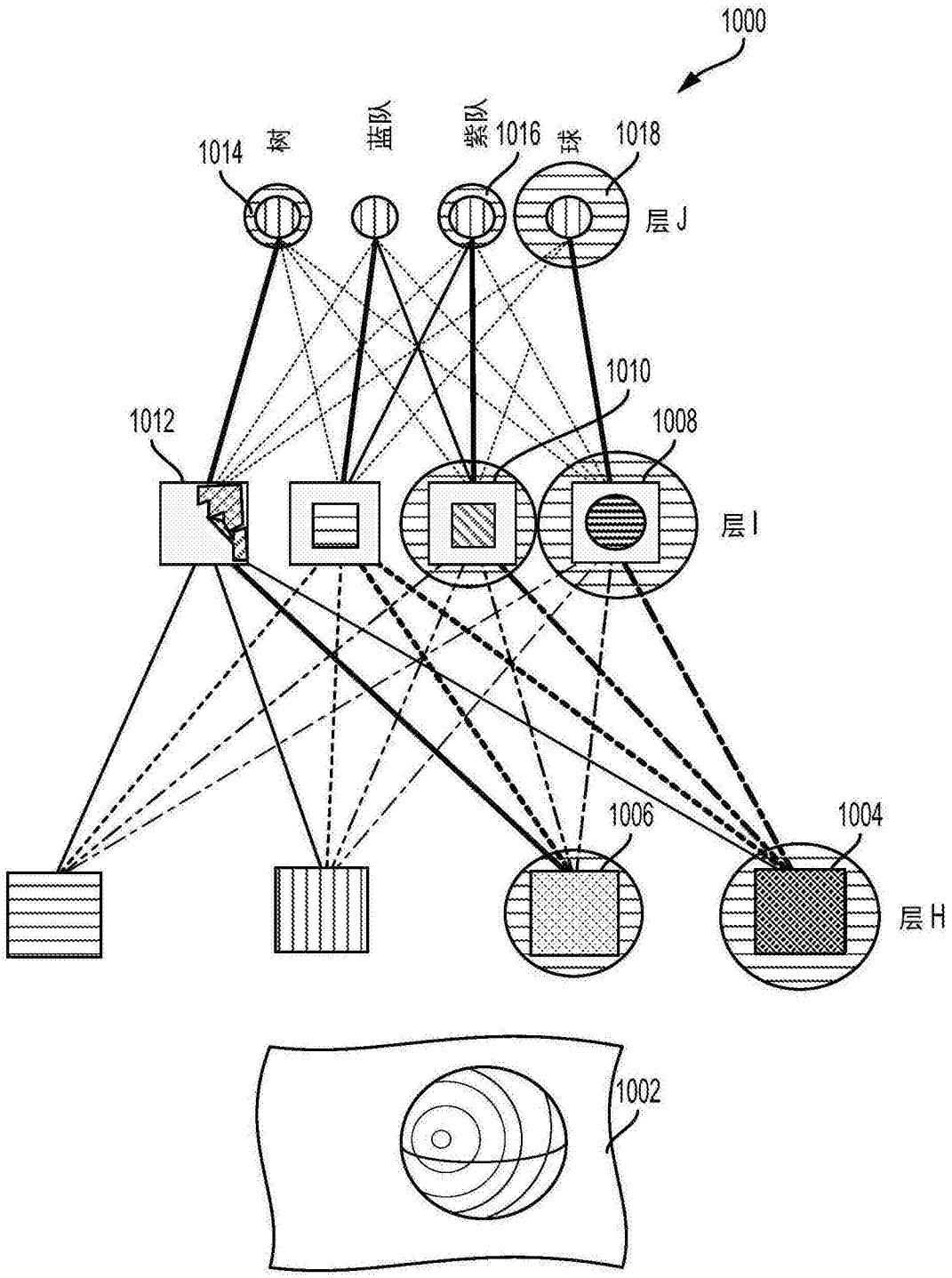


图10

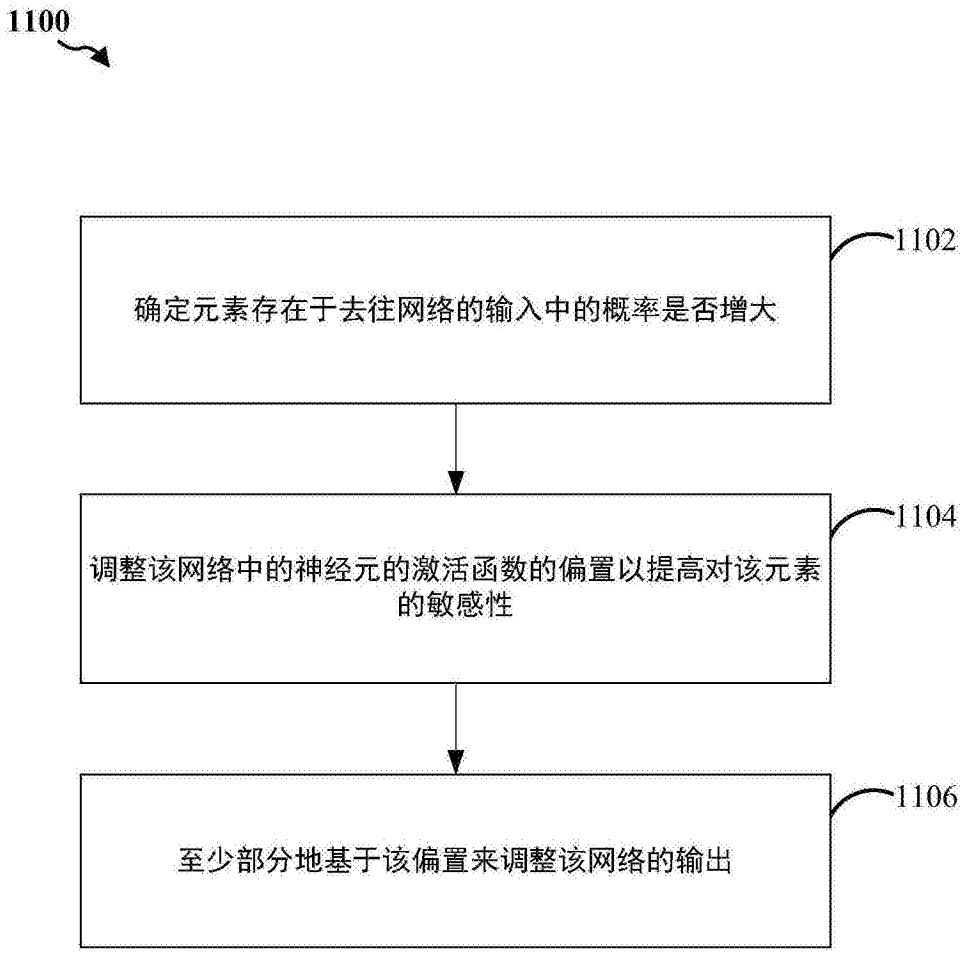

1100
确定元素存在于去往网络的输入中的概率是否增大
1102
调整该网络中的神经元的激活函数的偏置以提高对该元素的敏感性
1104
至少部分地基于该偏置来调整该网络的输出
1106

图11

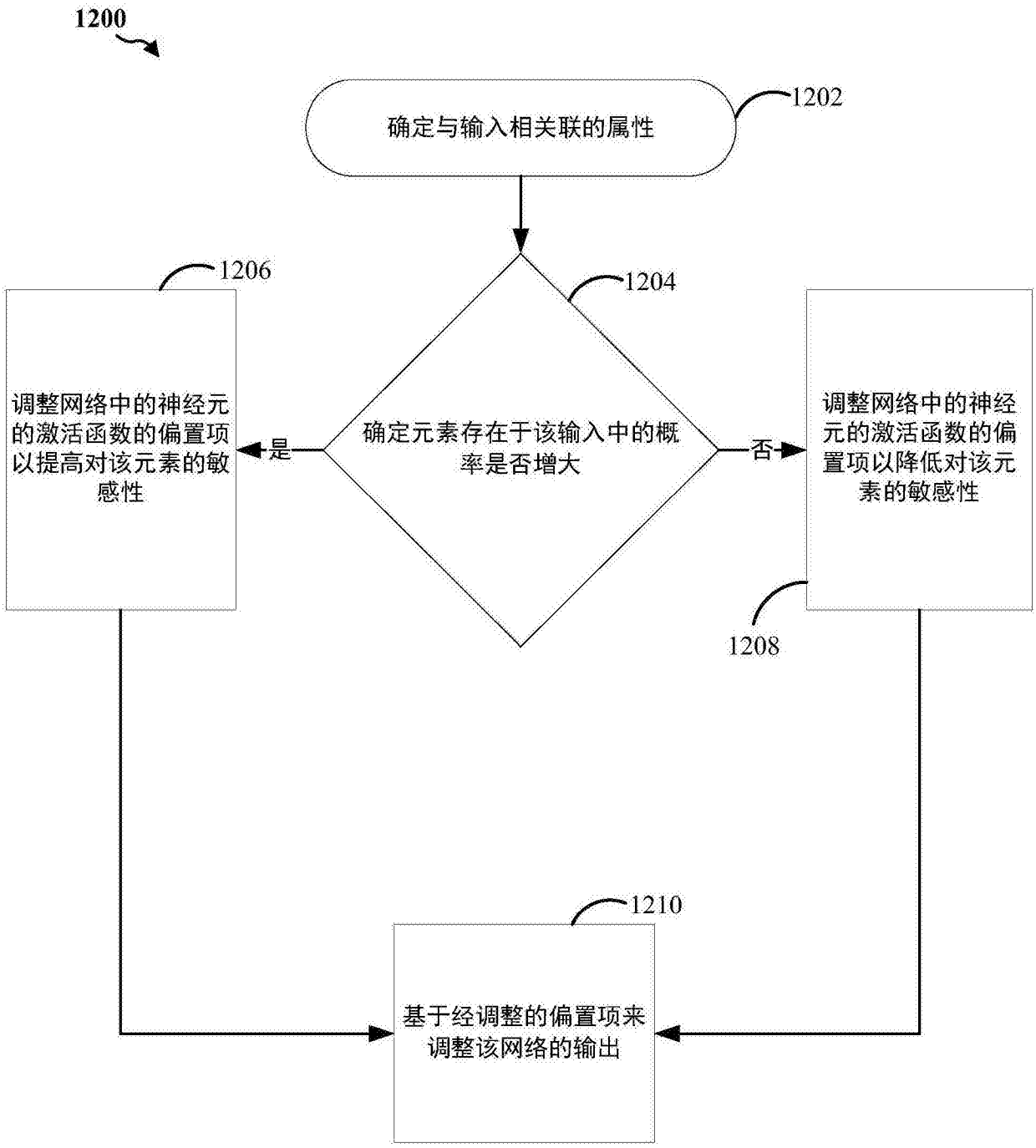

1200
确定与输入相关联的属性
1202
确定元素存在于该输入中的概率是否增大
1204
是
否
调整网络中的神经元的激活函数的偏置项以提高对该元素的敏感性
1206
调整网络中的神经元的激活函数的偏置项以降低对该元素的敏感性
1208
基于经调整的偏置项来调整该网络的输出
1210

图12