



(12)发明专利申请

(10)申请公布号 CN 107924486 A

(43)申请公布日 2018.04.17

(21)申请号 201680050371.8

(22)申请日 2016.08.04

(30)优先权数据

62/213,591 2015.09.02 US

15/077,873 2016.03.22 US

(85)PCT国际申请进入国家阶段日

2018.02.28

(86)PCT国际申请的申请数据

PCT/US2016/045636 2016.08.04

(87)PCT国际申请的公布数据

W02017/039946 EN 2017.03.09

(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 S·马宗达 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 周敏 陈炜

(51)Int.Cl.

G06N 3/08(2006.01)

G06N 99/00(2006.01)

G06K 9/62(2006.01)

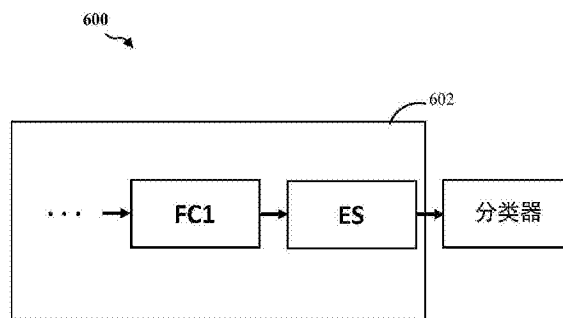
权利要求书2页 说明书11页 附图8页

(54)发明名称

用于分类的强制稀疏

(57)摘要

一种用于对输入进行分类的装置包括分类器和特征提取器。特征提取器被配置成基于输入生成特征向量。特征向量还被配置成将该特征向量的数个元素设为零以产生稀疏特征向量。稀疏特征向量具有与由特征提取器生成的特征向量相同的维度。然而，稀疏特征向量包括比由特征提取器生成的特征向量更少的非零元素。特征向量被进一步配置成将稀疏特征向量转发给该分类器以对输入进行分类。



1. 一种用于对输入进行分类的装置,包括:
分类器;以及
特征提取器,所述特征提取器被配置成:
从所述输入生成特征向量;
将所述特征向量的数个元素设为零以产生稀疏特征向量,所述稀疏特征向量具有与所生成的特征向量相同的维度并且包括比所生成的特征向量更少的非零元素;以及
将所述稀疏特征向量转发给所述分类器以对所述输入进行分类。
2. 如权利要求1所述的装置,其特征在于,所述特征提取器被进一步配置成至少部分基于性能度量来确定所述数个元素的数量。
3. 如权利要求2所述的装置,其特征在于,所述特征提取器被进一步配置成在设备上计算所述性能度量。
4. 如权利要求2所述的装置,其特征在于,所述特征提取器被进一步配置成在设备外计算所述性能度量。
5. 如权利要求1所述的装置,其特征在于,所述特征提取器被训练来确定所述特征向量中要保留的元素的数量。
6. 如权利要求5所述的装置,其特征在于,所述特征提取器是使用惩罚所述特征向量的低稀疏性的成本函数来训练的。
7. 如权利要求6所述的装置,其特征在于,所述成本函数包括L1范数或L0范数。
8. 如权利要求1所述的装置,其特征在于,所述特征提取器被进一步配置成二进制化所述稀疏特征向量的各元素。
9. 如权利要求1所述的装置,其特征在于,所述特征提取器被进一步配置成量化所述稀疏特征向量的各元素。
10. 一种用于对输入进行分类的方法,包括:
从所述输入生成特征向量;
将所述特征向量的数个元素设为零以产生稀疏特征向量,所述稀疏特征向量具有与所生成的特征向量相同的维度并且包括比所生成的特征向量更少的非零元素;以及
将所述稀疏特征向量转发给分类器以对所述输入进行分类。
11. 如权利要求10所述的方法,其特征在于,进一步包括至少部分基于性能度量来确定所述数个元素的数量。
12. 如权利要求11所述的方法,其特征在于,进一步包括在设备上计算所述性能度量。
13. 如权利要求11所述的方法,其特征在于,进一步包括在设备外计算所述性能度量。
14. 如权利要求10所述的方法,其特征在于,进一步包括确定要所述特征向量中要保留的元素的数量。
15. 如权利要求14所述的方法,其特征在于,所述确定包括使用惩罚所述特征向量的低稀疏性的成本函数来训练特征提取器。
16. 如权利要求15所述的方法,其特征在于,所述成本函数包括L1范数或L0范数。
17. 如权利要求10所述的方法,其特征在于,进一步包括二进制化所述稀疏特征向量的各元素。
18. 如权利要求10所述的方法,其特征在于,进一步包括量化所述稀疏特征向量的各元

素。

19. 一种用于对输入进行分类的设备,包括:

用于从所述输入生成特征向量的装置;

用于将所述特征向量的数个元素设为零以产生稀疏特征向量的装置,所述稀疏特征向量具有与所生成的特征向量相同的维度并且包括比所生成的特征向量更少的非零元素;以及

用于将所述稀疏特征向量转发给分类器以对所述输入进行分类的装置。

20. 一种其上编码有用于对输入进行分类的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:

用于从所述输入生成特征向量的程序代码;

用于将所述特征向量的数个元素设为零以产生稀疏特征向量的程序代码,所述稀疏特征向量具有与所生成的特征向量相同的维度并且包括比所生成的特征向量更少的非零元素;以及

用于将所述稀疏特征向量转发给分类器以对所述输入进行分类的程序代码。

用于分类的强制稀疏

[0001] 相关申请的交叉引用

[0002] 本申请要求于2015年9月2日提交的题为“ENFORCED SPARSITY FOR CLASSIFICATION(用于分类的强制稀疏)”的美国临时专利申请No.62/213,591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机器学习,且更具体地涉及改进特征提取和分类的系统和方法。

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。具体而言,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,依此类推。深度神经网络可被训练以识别特征阶层并且因此它们已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达到了优异的结果,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] 在本公开的一方面,公开了一种用于对输入进行分类的装置。该装置包括分类器和特征提取器。特征提取器被配置成从输入生成特征向量。特征向量还被配置成将该特征向量的数个元素设为零以产生稀疏特征向量。稀疏特征向量具有与由特征提取器生成的特征向量相同的维度。稀疏特征向量包括比由特征提取器生成的特征向量更少的非零元素。特征向量被进一步配置成将稀疏特征向量转发给分类器以对输入进行分类。

[0013] 在本公开的另一方面,公开了一种用于对输入进行分类的方法。该方法包括从输入生成特征向量。该方法还包括将该特征向量的数个元素设为零以产生稀疏特征向量。稀疏特征向量具有与所生成的特征向量相同的维度。稀疏特征向量还包括比所生成的特征向

量更少的非零元素。该方法进一步包括成将稀疏特征向量转发给分类器以对输入进行分类。

[0014] 在本公开的又一方面,公开了一种用于对输入进行分类的设备。该设备包括用于从输入生成特征向量的装置。该设备还包括用于将该特征向量的数个元素设为零以产生稀疏特征向量的装置。稀疏特征向量具有与所生成的特征向量相同的维度。稀疏特征向量还包括比所生成的特征向量更少的非零元素。该设备进一步包括用于成将稀疏特征向量转发给分类器以对输入进行分类的装置。

[0015] 根据本公开的再一方面,给出了一种非瞬态计算机可读介质。该非瞬态计算机可读介质其上编码有用于分类的程序代码。该程序代码由处理器执行并且包括用于从输入生成特征向量的程序代码。该程序代码还包括用于将该特征向量的数个元素设为零以产生稀疏特征向量的程序代码。稀疏特征向量具有与所生成的特征向量相同的维度。稀疏特征向量还包括比所生成的特征向量更少的非零元素。该程序代码进一步包括用于将稀疏特征向量转发给分类器以对输入进行分类的程序代码。

[0016] 本公开的附加特征和优点将在下文描述。本领域技术人员应当领会,本公开可容易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是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意图。

[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是解说根据本公开的各方面的包括特征提取器的示例性机器学习模型的框图。

[0026] 图7解说了根据本公开的各方面的用于特征提取和输入分类的方法。

[0027] 详细描述

[0028] 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。

[0029] 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。

[0030] 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。

[0031] 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。

[0032] 用于分类的强制稀疏

[0033] 在分类任务中,经由特征提取器输出的特征向量通常可能是密集的(例如,包含许多非零元素)。具有此类密集的特征向量可不利地影响存储器要求和分类器等待时间。此外,具有大量的具有较小的非零特征值的特征向量元素可表示有噪声的特征向量,这进而可降低分类准确性。

[0034] 本公开的各方面涉及改进的特征提取和分类准确性。在本公开中,强制稀疏(ES)过程被采用使得给定特征向量的仅最高K个或百分之K的特征值或元素被保留。其它值可被设为零(0),从而产生具有比给定特征向量更少的非零值的稀疏特征向量。然而,特征向量的维度(例如,特征向量中元素的数量)可被维持。通过增加特征向量的稀疏性,例如,可使用更少存储器来存储用于重新训练的特征。此外,更高的稀疏性也可改进分类性能(例如,分类速度和准确性)。

[0035] 图1解说了根据本公开的某些方面的使用片上系统(SOC) 100进行前述强制稀疏和特征提取的示例实现,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。

[0036] 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)。

[0037] 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接收来自特征提取器的特征向量的代码。加载到通用处理器102中的指令还可包括用于保留特征向量的一百分比的元素以产生稀疏特征向量的代码。此外,加载到通用处

理器102中的指令还可包括用于将稀疏特征向量转发给分类器的代码。

[0038] 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。

[0039] 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类似的特征,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。

[0040] 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。

[0041] 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,机动车辆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。

[0042] 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。

[0043] 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。

[0044] 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。

[0045] 深度卷积网络(DCN)可以用受监督式学习来训练。在训练期间,可向DCN呈递图像(诸如限速标志的经裁剪图像326),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。

[0046] 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)”。

[0047] 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。

[0048] 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。

[0049] 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中。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。

[0050] 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。

[0051] 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。

[0052] 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。

[0053] 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。

[0054] 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,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。

[0055] 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。

[0056] 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。

[0057] 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的运行时的操作期间执行支持计算。

[0058] 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。

[0059] 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) 来访问。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。

[0060] 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 (例如, 使用JAVA编程语言) 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)。经预处理的图像可随后被传达给分类应用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可以被 (例如, 使用C编程语言) 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可被配置成通过缩放 (516) 和剪裁 (518) 来进一步预处理 (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 $\times$ 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的各种处理块借助深度神经网络来进一步处理图像像素。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。经平滑的结果可随后使得智能手机502的设置和/或显示改变。

[0061] 在一种配置中, 机器学习模型被配置成用于从输入生成特征向量。该模型还被配置成用于将该特征向量的数个元素设为零以产生稀疏特征向量。机器学习模型进一步被配置成用于将该稀疏特征向量转发给分类器。机器学习模型包括生成装置、设置装置、和/或转发装置。在一个方面, 生成装置、设置装置、和/或转发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。在另一配置中, 前述装置可以是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。

[0062] 根据本公开的某些方面, 每个局部处理单元202可被配置成基于神经网络的一个或多个期望功能性特征来确定神经网络的参数, 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。

[0063] 图6是解说根据本公开的各方面的包括特征提取器602的示例性机器学习模型600的框图。参见图6, 示出了特征向量602的全连接层FC1和强制稀疏单元。在一些方面, 全连接层可例如包括深度卷积网络 (DCN) 或其它网络的层 (例如, 最后层)。

[0064] 全连接层FC1可向强制稀疏 (ES) 单元提供特征向量。在此示例性配置中, 输入 (例如, 图像像素、话音等) 可被传递通过DCN的多个层以提取特定特征并经由全连接层FC1输出特征向量。

[0065] 典型地, DCN可采用矫正器线性单元 (ReLU) 或参数化矫正器线性单元 (PReLU) 来矫正该数据。ReLU通过将负特征向量值设为0并保持正值来矫正该数据。另一方面, PReLU保持正特征向量值并线性地缩放负值。然而, 相比于所提议的强制稀疏方法, 两者均在训练和推断中增加的存储器消耗和提高的等待时间下产生特征向量。

[0066] 根据本发明的各方面, 强制稀疏可经由ES单元来应用。即, 包括在特征向量中的数据可被提供给强制稀疏 (ES) 单元来稀疏数据或使得数据稀疏。使用ES单元, 最高K%的数据

元素可被保持,其中K是整数。即,具有最高值的K个元素或K%的元素可被保留。特征向量中剩余的元素可被设为零。如此,与所提供的特征向量具有相同维度或元素数量的稀疏特征向量可被产生,其仅针对K个元素或K%的元素包括非零值。

[0067] 在一些方面,K的值可被离线计算或确定。例如,K的值可基于跨0%和100%之间的各个K值的参数扫描来确定。

[0068] 相反,K的值也可在线确定。例如,K的值可通过保留来自用户的一组训练和验证样本并执行跨K的扫描来确定。

[0069] 尽管ES单元在图6中被示为在DCN的外部,但在一些方面,ES单元可被纳入DCN或其它网络内。在一个示例中,稀疏化可通过将最小绝对误差(L1)成本函数作为DCN或其它特征提取器的训练规程的一部分来应用而被纳入DCN中。在一些方面,成本函数可被配置成惩罚特征向量中所包括的非零元素的数量。此外,使误差的L1范数最小化可迫使非零特征值的数量为较小数量。如此,DCN可学习稀疏因子,以及一些情形中的特征向量的最有利的或期望的稀疏因子。

[0070] 在DCN中,通过使得权重作为预测标记和实际标记之间的误差的函数来更新,训练得到进展。该误差是惩罚项,并且一个目的是将误差减小到零。根据本公开的各方面,可以添加第二惩罚项。第二惩罚项可包括期望稀疏的层的激活的范数。由于目的是最小化特征向量中的非零元素的数目,因此该第二惩罚项可在一些方面包括该层中非零项的数量的计数。非零项的计数可以是L0范数。

[0071] 然而,该量将不是可区分的并且可能导致训练网络(例如,使用反向传播训练网络,其使用成本函数的梯度来驱动权重更新)中的困难。因此,在一些方面,L1范数(这些项的绝对值的总和、而不是各项的计数)可被用作第二惩罚项。通过强制较小的绝对值总和,这些项的数量可被间接地迫使成走向零(或可被向下取阈到零的非常小的数量)。

[0072] 在这个范例中,非零项的“最优”数量(例如,K)可作为成本函数最小化的一部分来确定。在该情形中,非零项的数量可被认为最优,因为其最小化误差且最大化稀疏性。

[0073] 在一个示例中,包括元素 $F[-1 \ 0 \ 2 \ -3 \ 5 \ 7 \ 9 \ 4 \ -1 \ 2]$ 的特征向量可被提供给ES单元。ES单元可以例如被配置成保持最高20%的特征向量值或元素。因此,ES单元可确定特征向量的十个(10)元素中的最高两个(2)元素或特征值。因此,ES单元可输出具有非零特征值或元素的稀疏性的经修改的特征向量 $F'[0 \ 0 \ 0 \ 0 \ 0 \ 7 \ 9 \ 0 \ 0 \ 0]$ 。因此,稀疏特征可经由ES单元被输出并提供给分类器。

[0074] 在一些方面,ES单元可保留最多K%的特征向量。替换地,ES单元可保留特征向量元素的最高K%的绝对值、正值、或负值。此外,选择最高K%是用于选择特征向量的幸存特征值或非零元素的实例化。

[0075] 稀疏特征向量可以是特别有益的,因为其可降低用于存储用于重新训练的特征的存储器消耗。增加的稀疏性也可有助于实现更快的分类器训练和推断,因为执行了更少的计算。另外,稀疏特征向量可以改善分类器准确性。此外,稀疏特征向量还帮助执行涉及计算特征空间中两个或更多个特征向量之间的距离的任务,例如以形成“类似”特征的群集,或构建如“最近邻居分类器”的更简单的分类器。

[0076] 在一些方面,经修改的特征向量的元素值可被二进制化或量化。例如,在以上示例中,F的二进制化版本可以是 $F'[0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0]$ 。稀疏特征向量可进而被呈现给分

类器。

[0077] 在第二个示例中,元素值可以被量化。在此示例中,全部“幸存量”或K个最高值可用1来编码,而其它值全部用0编码。例如,如果期望稀疏性为80%且向量大小为10,则8个最低量可被设为0而两个幸存量(例如,最高元素值)可被设为1。

[0078] 图7解说了用于特征提取和输入分类的方法700。在框702,该过程从输入生成特征向量。输入可用是图像、语音、话音、或其它输入数据。在框704,该过程将该特征向量的数个元素设为零以产生稀疏特征向量。稀疏特征向量具有与所生成的特征向量相同的维度并包括比所生成的特征向量更少的非零元素。

[0079] 在一些方面,例如,元素的数量可以基于性能度量(诸如,分类器等待时间、分类器准确性、分类器速度、和/或存储器使用)来确定。性能度量可在设备上(例如,在执行分类任务的设备上)或在设备外被确定。

[0080] 进一步,在框706,该过程将稀疏特征向量转发给分类器。在一些方面,该过程可进一步量化该稀疏特征向量的各元素以进一步减少存储器消耗。

[0081] 在一些方面,该过程可进一步训练特征提取器来确定特征向量中要保留的元素的数量。训练可包括惩罚特征向量的低稀疏性的成本函数的应用。在一些方面,成本函数可包括最低绝对误差(L1-范数)或L0范数正则化。

[0082] 在一些方面,方法700可由SOC 100(图1)或系统200(图2)来执行。即,举例而言但不作为限定,方法700的每个元素可由SOC 100或系统200、或者一个或多个处理器(例如,CPU 102和局部处理单元202)和/或其中所包括的其他组件来执行。

[0083] 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。

[0084] 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。

[0085] 如本文中所使用的,引述一系列项目中的“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中的至少一个”旨在涵盖:a、b、c、a-b、a-c、b-c、以及a-b-c。

[0086] 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中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。

[0087] 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。替换地, 存储介质可以被整合到处理器。

[0088] 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。

[0089] 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。

[0090] 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。

[0091] 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。

[0092] 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中所描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。

[0093] 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。可随后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。

[0094] 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。

[0095] 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。

[0096] 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,本文中所描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中所描述的方法和技术的任何其他合适的技术。

[0097] 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。

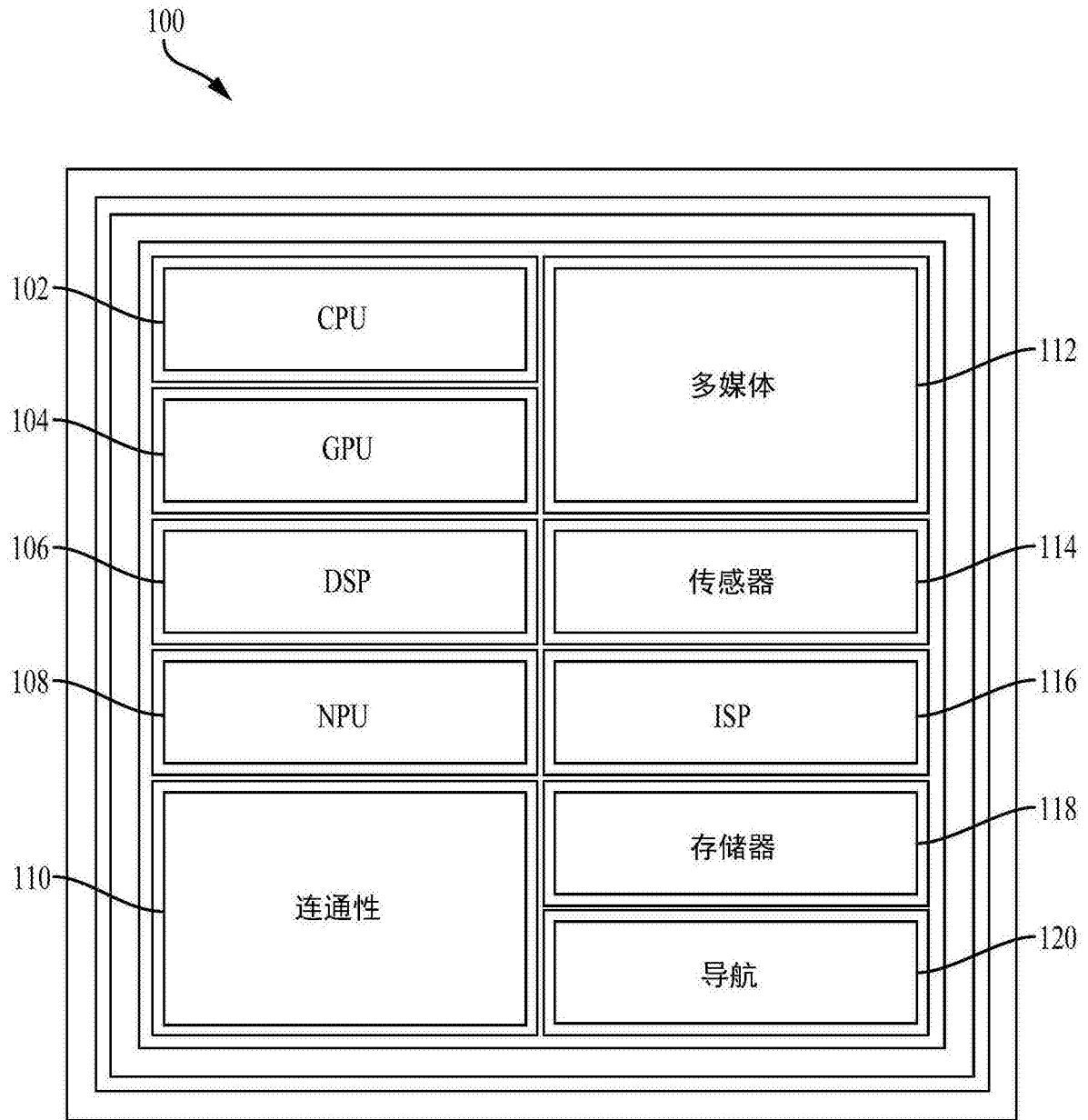


图1

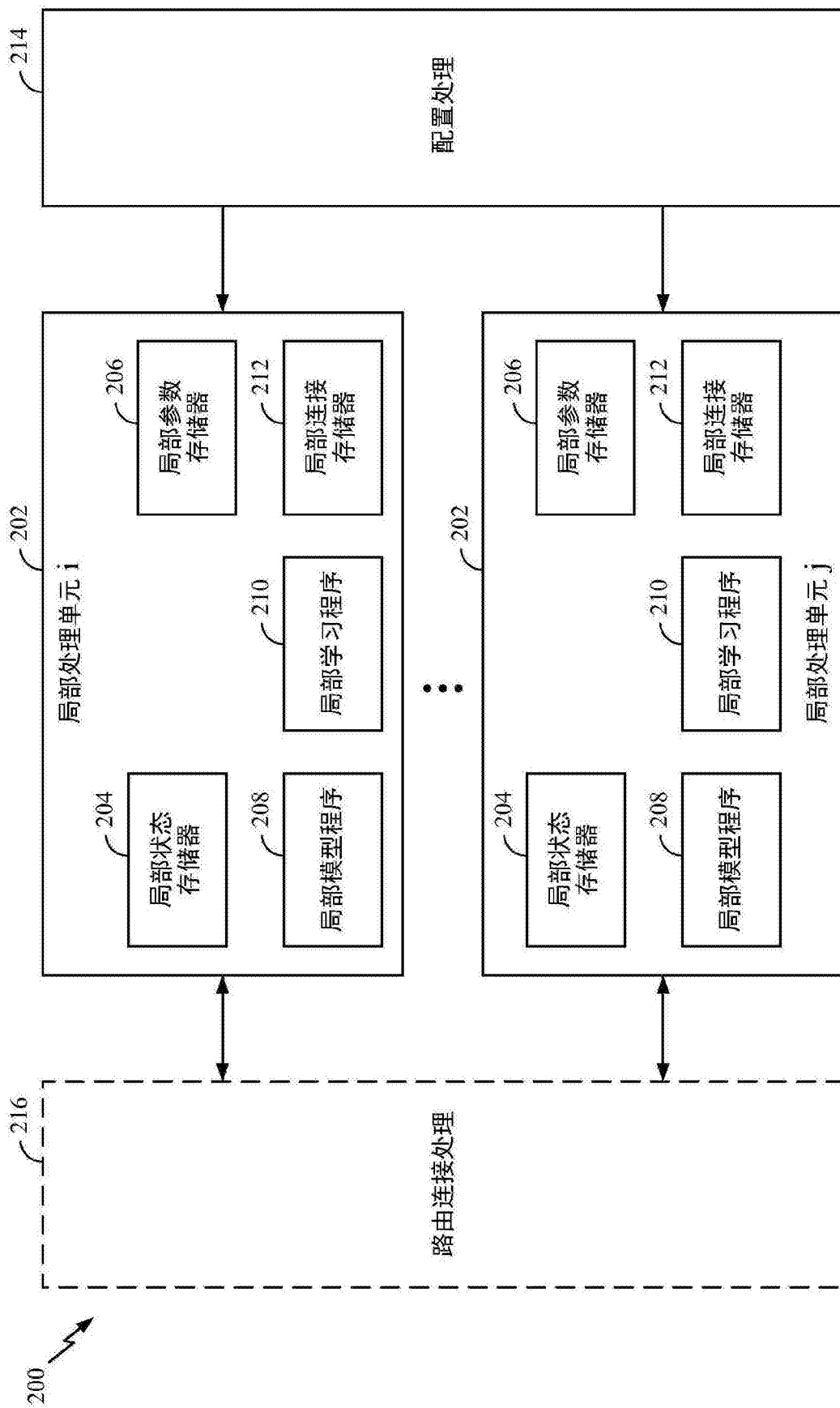


图2

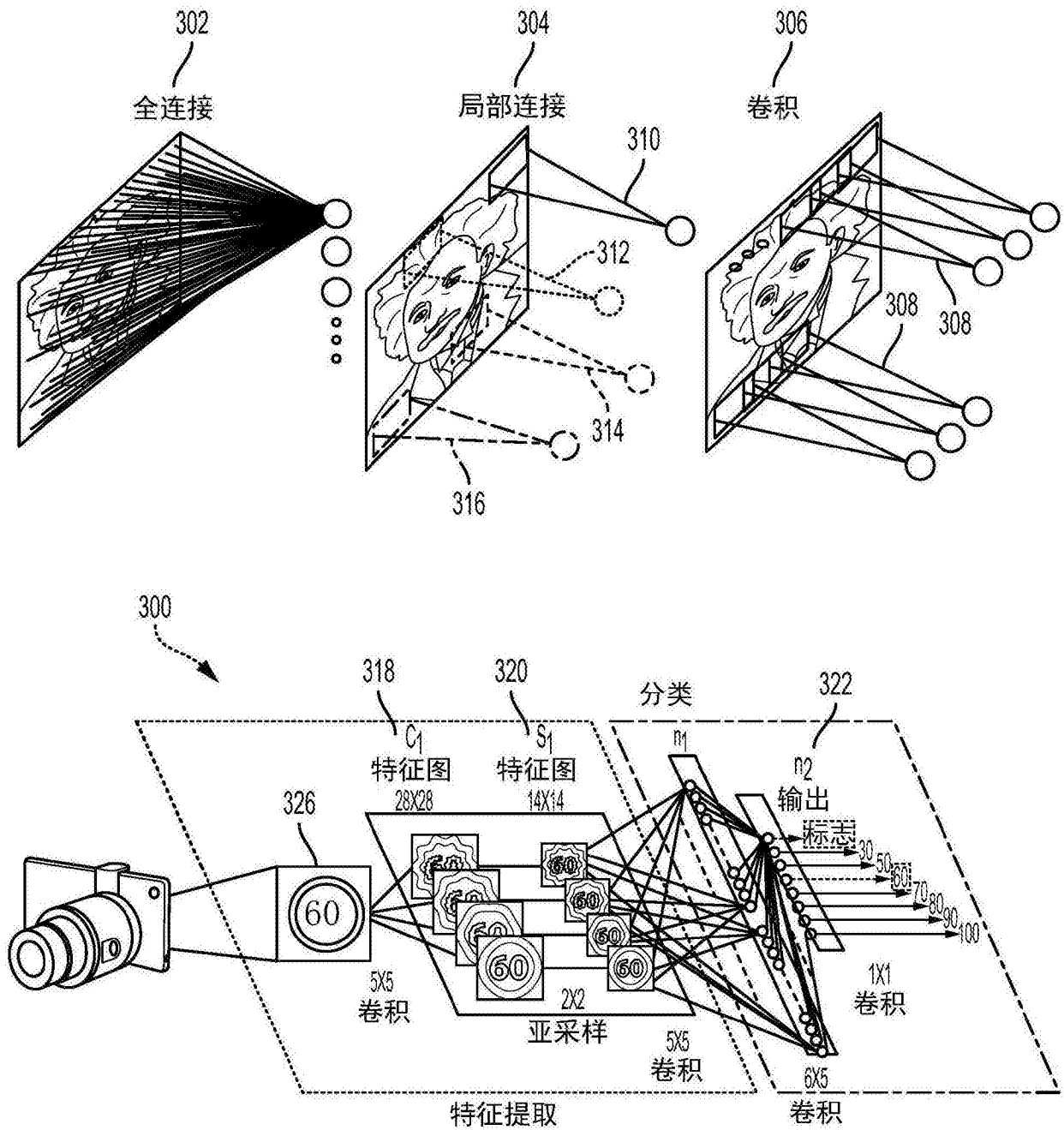


图3A

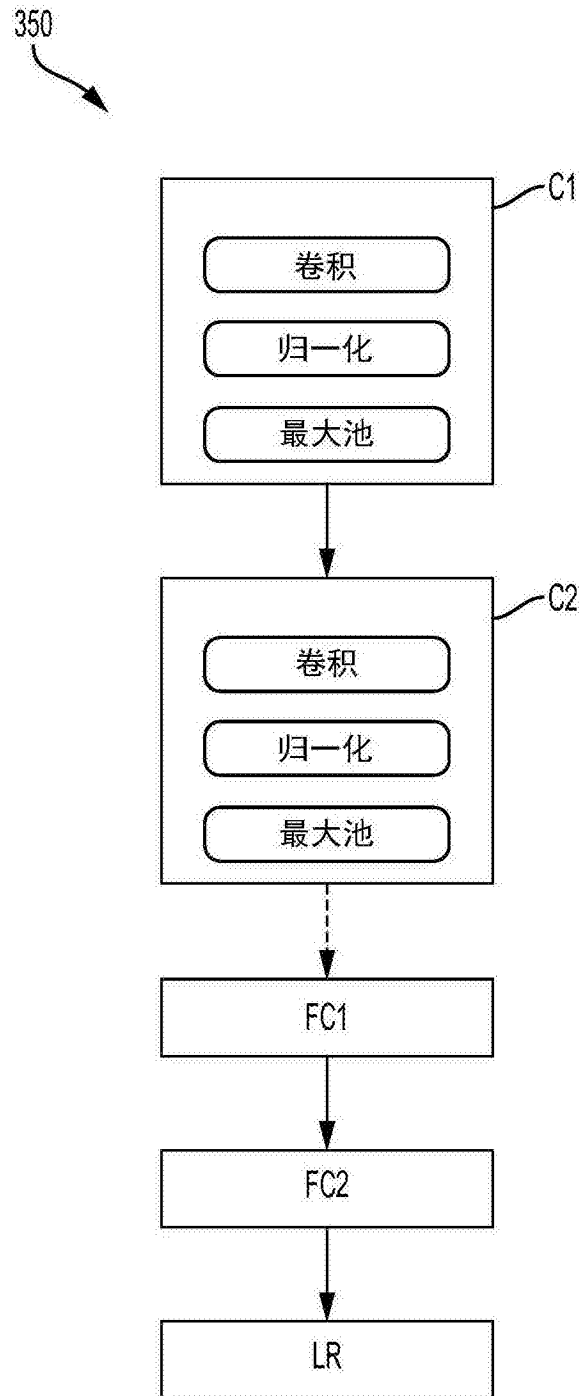


图3B

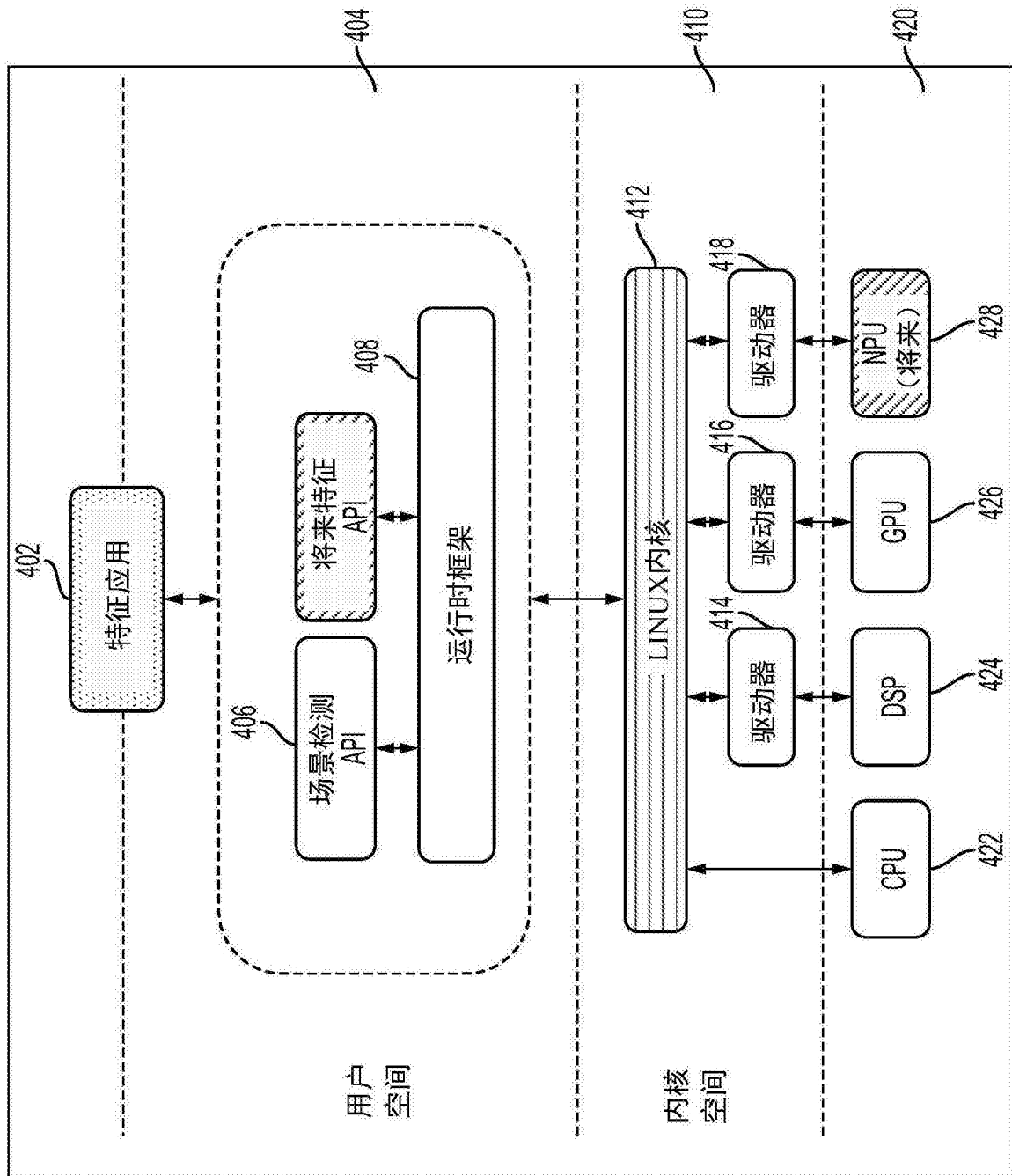


图4

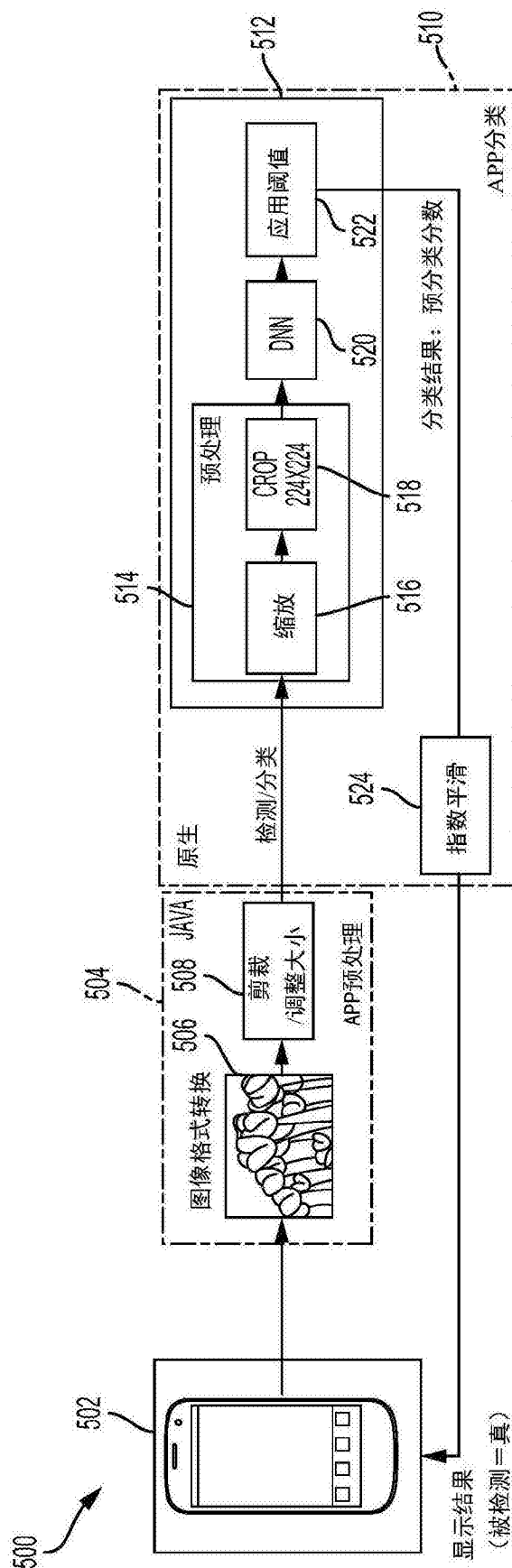


图5

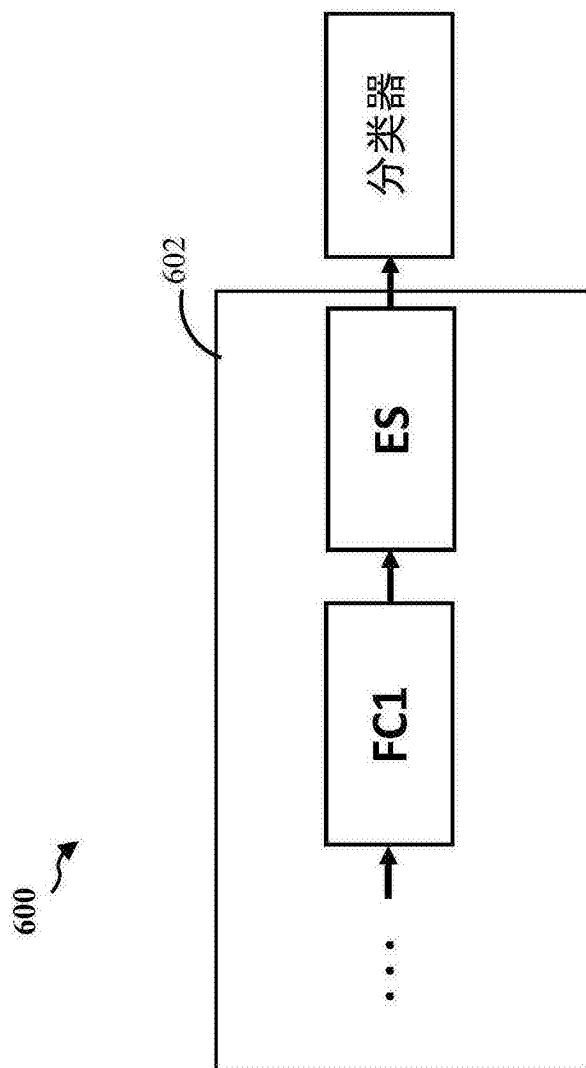


图6

700

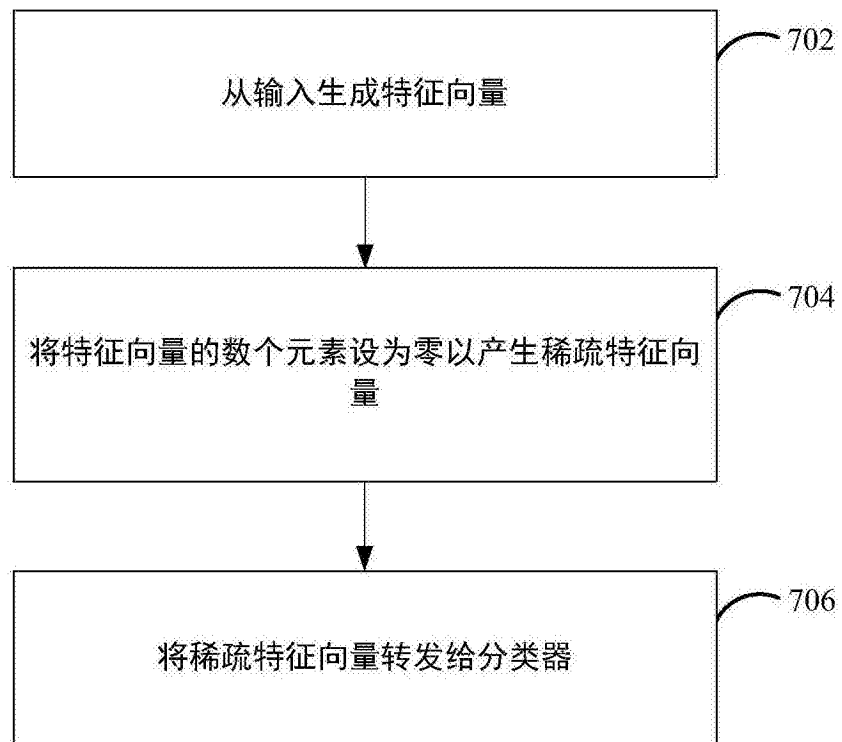


图7