(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105447498 B
(45)授权公告日 2021.01.26

(21)申请号 201510608615.7
(22)申请日 2015.09.22
(65)同一申请的已公布的文献号
申请公布号 CN 105447498 A
(43)申请公布日 2016.03.30
(30)优先权数据
62/053,692 2014.09.22 US
14/663,233 2015.03.19 US
(73)专利权人 三星电子株式会社
地址 韩国京畿道水原市
(72)发明人 冀正平 伊利亚·奥夫相尼科夫
王一兵 石立龙
(74)专利代理机构 北京铭硕知识产权代理有限公司 11286
代理人 刘灿强 尹淑梅
(51)Int.Cl.
G06K 9/62(2006.01)
G06N 3/08(2006.01)
(56)对比文件
CN 102411708 A,2012.04.11
CN 101299233 A,2008.11.05
US 2014079315 A1,2014.03.20
US 2013339281 A1,2013.12.19
审查员 司马成

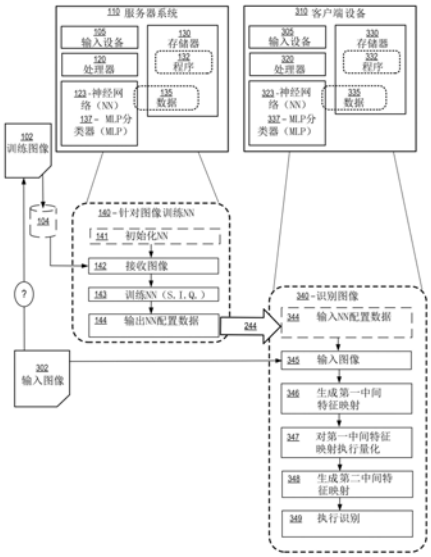
权利要求书2页 说明书14页 附图13页

(54)发明名称

配置有神经网络的客户端设备、系统和服务器系统

(57)摘要

提供了配置有神经网络的客户端设备、系统和服务器系统。一种配置有神经网络的客户端设备，所述客户端设备包括：处理器、存储器、用户接口、通信接口、电源和输入设备，其中，存储器包括从服务器系统接收的训练神经网络，所述服务器系统已经训练并配置了用于客户端设备的神经网络。公开了训练神经网络的服务器系统和方法。



CN 105447498 B

1.一种配置有训练神经网络的客户端设备,所述客户端设备包括:

处理器、存储器、用户接口、通信接口、电源和输入设备;

所述存储器包括从服务器系统接收的训练神经网络,所述服务器系统已经训练并配置了将用作训练神经网络的基于服务器系统的神经网络,

其中,基于服务器系统的神经网络被配置为生成特征映射,其中,所述特征映射包括从输入图像得到的多个权重值,

其中,基于服务器系统的神经网络还被配置为:对所述特征映射进行统一量化操作或监督迭代量化操作,以在不改变所述特征映射的维度情况下将所述多个权重值中的每个权重值的位数从第一预定数量减小为比第一预定数量小的第二预定数量。

2.根据权利要求1所述的客户端设备,其中,所述输入设备被配置为捕获图像并且将图像输入数据存储在存储器中。

3.根据权利要求1所述的客户端设备,所述客户端设备还包括被配置为映射图像输入数据的多层感知器分类器。

4.根据权利要求1所述的客户端设备,其中,训练神经网络包括卷积神经网络。

5.根据权利要求1所述的客户端设备,其中,监督迭代量化操作执行图像输入数据的反向传播。

6.根据权利要求1所述的客户端设备,其中,训练神经网络被配置为执行目标识别。

7.根据权利要求1所述的客户端设备,所述客户端设备包括智能电话、平板电脑和便携式电子设备。

8.一种利用客户端设备提供目标识别的系统,所述系统包括:

服务器系统,被配置为训练神经网络以执行目标识别,并且将训练神经网络输出到所述客户端设备,

其中,所述神经网络被配置为生成特征映射,其中,所述特征映射包括从输入图像得到的多个权重值,

其中,所述神经网络还被配置为:对所述特征映射进行统一量化操作或监督迭代量化操作,以在不改变所述特征映射的维度情况下将所述多个权重值中的每个权重值的位数从第一预定数量减小为比第一预定数量小的第二预定数量。

9.根据权利要求8所述的系统,所述系统还包括:

客户端设备,包括处理器、存储器、用户接口、通信接口、电源和输入设备;并且

所述存储器包括从所述服务器系统接收的训练神经网络。

10.根据权利要求8所述的系统,其中,所述服务器系统包括训练图像的数据库。

11.根据权利要求8所述的系统,其中,所述神经网络包括卷积神经网络。

12.根据权利要求11所述的系统,其中,所述卷积神经网络包括至少两个层。

13.根据权利要求11所述的系统,其中,所述卷积神经网络包括卷积层和子采样层。

14.根据权利要求8所述的系统,其中,所述服务器系统包括多层感知器分类器。

15.根据权利要求8所述的系统,其中,针对所述神经网络的学习技术包括反向传播、限制玻尔兹曼机、自动编码器解码技术中的一者。

16.一种服务器系统,所述服务器系统包括:

输入设备,被配置为接收训练图像;

神经网络，包括至少两个层对，每个层对包括卷积层和子采样层；

多层感知器分类器；

其中，所述神经网络被配置为执行卷积层中的中间权重的量化，

所述神经网络还被配置为响应于应用到卷积层的输入而在所述子采样层中生成中间特征映射；所述神经网络被配置为执行所述多层感知器分类器中的权重的量化，并且

所述神经网络被配置为响应于应用到量化的权重多层感知器的所述特征映射而在所述多层感知器分类器中生成分类输出，

其中，卷积层中的中间权重的量化和所述多层感知器分类器中的权重的量化中的至少一个量化为：用于在不改变与所述至少一个量化对应的特征映射的维度情况下将所述特征映射的每个权重的位数从第一预定数量减小为比第一预定数量小的第二预定数量的统一量化操作或监督迭代量化操作。

配置有神经网络的客户端设备、系统和服务器系统

[0001] 本申请要求于2014年9月22日提交的第62/053,692号美国临时专利申请和2015年3月19日提交的第14/663,233号美国专利申请的优先权，上述专利申请的公开内容通过引用包含于此。

技术领域

[0002] 本发明涉及具有减小的神经网络权重精度的目标识别，更具体地，涉及配置有神经网络的客户端设备、系统和服务器系统。

背景技术

[0003] 机器(即，计算机)越来越多地用于提供机器视觉或目标识别。目标识别提供给用户各种有益的工具。

[0004] 在某些情况下，目标识别依赖于包括神经网络的算法。即，设备可以通过使用神经网络来识别目标在输入图像之内。通常，神经网络已经被训练为通过训练图像的预先使用来识别目标。如果对目标使用更多的训练图像，则该目标识别处理会变得更有辨识能力。

[0005] 通常，神经网络包括互相连接的“神经元”的系统。神经网络计算来自输入的值，并且由于神经网络的自适应性(adaptive nature)而能够进行机器学习以及模式识别。

[0006] 用于图像识别的神经网络需要用于学习和用于识别的数据处理，该数据处理是存储器加强的和处理密集的，因此在计算上是昂贵的。的确，在计算处理期间，权重的值需要用于存储和用于处理的存储空间。

[0007] 如可预期的，增大训练数据集的大小提高神经网络的性能。遗憾的是在诸如智能电话等的移动设备中，存储器和处理能力相对有限。因此，移动设备的越来越普遍的使用通常尚未受益于图像识别技术。

[0008] 需要的是，提供在具有有限资源的计算设备上提高神经网络的性能的方法和设备。

发明内容

[0009] 本说明书给出了服务器系统、客户端设备以及方法的实例，其中，所述服务器系统用于图像识别训练，所述图像识别训练可以生成用于神经网络的配置信息；所述客户端设备基于下载的神经网络配置信息，利用神经网络来执行图像识别；所述方法的使用可以帮助克服现有技术的问题和限制。

[0010] 在一个实施例中，客户端设备被配置为具有神经网络。所述客户端设备包括：处理器、存储器、用户接口、通信接口、电源和输入设备，其中，存储器包括从服务器系统接收的训练神经网络，所述服务器系统已经训练并配置了用于客户端设备的神经网络。

[0011] 输入设备被配置为捕获图像并且将图像输入数据存储在存储器中。客户端设备还包括被配置为映射图像输入数据的多层感知器(MLP)分类器。神经网络可以包括卷积神经网络。神经网络被配置为生成特征映射；特征映射包括从输入图像得到的多个权重值；神经

网络可以被配置为对所述特征映射进行量化操作。量化操作可以包括统一量化、自适应量化、均匀量化和监督迭代量化中的一者。量化操作可以执行图像输入数据的反向传播(BP)。神经网络被配置为执行目标识别。客户端设备可以包括智能电话、平板电脑和便携式电子设备。

[0012] 在另一个实施例中，提供一种利用客户端设备提供目标识别的系统。所述系统包括：服务器系统，被配置为训练神经网络以执行目标识别，并且将神经网络输出到客户端设备。

[0013] 系统还可以包括：客户端设备，包括处理器、存储器、用户接口、通信接口、电源和输入设备；并且存储器包括从所述服务器系统接收的训练神经网络。服务器系统包括训练图像的数据库。神经网络包括卷积神经网络；卷积神经网络包括至少两个层；卷积神经网络包括卷积层和子采样层；服务器系统包括多层感知器(MLP)分类器。服务器系统还被配置为采用用于训练的学习技术；学习技术可以包括反向传播(BP)、限制玻尔兹曼机、自动编码器解码技术中的一者。

[0014] 在另一个实施例中，提供一种服务器系统，所述服务器系统包括：输入设备，被配置为接收训练图像；神经网络，包括至少两个层对，每个层对包括卷积层和子采样层；多层感知器(MLP)分类器；其中，神经网络被配置为执行卷积层中的中间权重的量化，神经网络还被配置为响应于应用到卷积层的输入而在子采样层中生成中间特征映射；神经网络被配置为执行MLP分类器中的权重的量化，神经网络被配置为响应于被应用到量化的权重MLP的所述特征映射而在MLP分类器中生成分类输出。

[0015] 在另一个实施例中，包括用于训练被配置为使用客户端设备进行目标识别的神经网络的方法，所述方法可以包括以下步骤：利用神经网络接收输入图像；通过神经网络来执行训练通过，训练通过包括量化操作；使用来自量化操作的权重值来配置神经网络；配置由客户端设备使用的神经网络；以及将神经网络存储在服务器系统中和将神经网络输出到客户端设备中的至少一者。

[0016] 在又一个实施例中，包括存储在非临时机器可读介质上的计算机程序产品。计算机程序产品可以包括用于通过执行如下方法来训练被配置为使用客户端设备进行目标识别的神经网络的机器可执行指令，其中，所述方法包括以下步骤：利用神经网络接收输入图像；通过神经网络执行训练通过，训练通过包括量化操作；使用来自量化操作的权重值来配置神经网络；对由客户端设备使用的神经网络进行配置；以及将神经网络存储在服务器系统中和将神经网络输出到客户端设备中的至少一者。

[0017] 比现有技术的优势包括：具有可接受的或者甚至未降低的性能为根据实施例制作的客户端设备节约了成本。的确，用于存储并且还对识别目标进行处理的存储器需求可以是较低的，这使得区域需求较低、功耗较低，因此成本较低。

[0018] 这些优势对移动设备而言会是重要的。的确，这种设备可以能实时执行片上图像识别，这在诸如情境感知的移动应用中会是有用的。

附图说明

[0019] 本发明的特征和优点通过结合附图进行的以下描述而明显，其中：

[0020] 图1是根据实施例的样本服务器系统和样本客户端设备的框图、根据实施例的流

程图以及其他相关方面的组合图。

[0021] 图2是图1的服务器或客户端的任一者的神经网络的样本框图的组合图并且还示出了根据实施例的可能的操作。

[0022] 图3是示出根据实施例的用于解释图2的组件的操作的卷积层和子采样层的样本对的图。

[0023] 图4是示出根据实施例在图1的训练流程图中的神经网络训练处理的流程图。

[0024] 图5是用于服务器系统的神经网络的样本框图，其还示出了在何处可以根据实施例来应用图4的流程图的量化操作。

[0025] 图6是可以在一维自适应量化操作中使用的样本方程表达式。

[0026] 图7是用于示出通过应用图6的方程表达式得到的自适应量化操作的一维示例的样本图。

[0027] 图8是用于示出可以在图4的流程图中使用的自适应量化操作的二维示例的样本图。

[0028] 图9示出用于执行自适应量化操作的实施例的样本方程表达式。

[0029] 图10是可以在一维均匀量化操作中用于确定的量化值的样本方程。

[0030] 图11是用于示出通过应用图10的方程表达式的版本得到的均匀量化操作的一维示例的样本图。

[0031] 图12是用于客户端设备的神经网络的样本框图，其还示出了在何处可以根据实施例来应用量化操作。

[0032] 图13是用于执行模拟的针对手写识别的一组样本训练图像。

[0033] 图14是用于执行模拟的样本卷积神经网络中的样本数据结构的概念图。

[0034] 图15是示出将实施例的模拟性能与现有技术对比的模拟结果的表。

[0035] 图16是用于将实施例的模拟性能与现有技术对比的条形图。

[0036] 图17是用于将实施例的模拟性能与现有技术对比的另一个条形图。

[0037] 图18示出用于将实施例的模拟性能与现有技术对比的多组条形图。

具体实施方式

[0038] 这里公开的是提供在具有有限资源的客户端上高效操作神经网络的方法和设备。通常，该方法和设备提供使用计算密集的学习处理在具有大量计算资源的设备(诸如服务器)上建立卷积神经网络(CNN)。一旦建立，该神经网络可以移植到具有相对有限计算资源的客户端设备(诸如智能电话)。

[0039] 神经网络对各种计算上复杂的任务是有用的。例如，神经网络可以用于目标识别。目标识别可以提供面部识别、环境监视、控制产品和制造、协助医疗诊断以及各种其他类似的处理。

[0040] 神经网络的类型包括：具有单向逻辑的仅单层或两层的神经网络、具有复杂的多路输入的神经网络、具有多方向反馈回路的神经网络以及具有多个层的神经网络。通常，这些系统在它们的编程中使用算法来确定它们的函数的控制和组织。大部分系统使用“权重”(可以表示为值)来改变吞吐量的参数以及改变至神经元的连接。神经网络可以是自主的，并且从通过使用训练数据的集完成的先前训练学习。

[0041] 为了给这里的教导提供一些内容，现在介绍一些方面。

[0042] 如这里讨论的，术语“服务器”通常指的是具有鲁棒性计算资源的计算资源。示例性资源包括用于执行这里描述的任务的那些重要的资源，并且可以包括大量的存储器、处理能力和数据存储等。在示例性实施例中，服务器包括传统的服务器(诸如刀片式服务器)、大型计算机的主机、个人计算机的网络、或者仅个人计算机(PC)。

[0043] 如这里讨论的，术语“客户端”通常指的是具有计算资源的缩减集的计算资源。示例性资源包括用于执行这里描述的任务的那些重要的资源，并且可以包括最小量的存储器、处理能力和数据存储等。在一些实施例中，客户端包括成像能力以提供被输入到神经网络中的输入图像的收集。

[0044] 如这里讨论的，术语“神经网络”通常指具有自适应性的统计学的学习算法，因此在机器学习中是有用的。神经网络可以包括多个人工节点，所述多个人工节点作为“神经元”、“处理元素”、“单元”或通过其他类似术语而公知，这些节点连接到一起形成模仿生物神经网络的网络。通常，神经网络包括自适应权重(即，通过学习算法调谐的数值参数)的集，并且能够近似于它们的输入的非线性函数。自适应权重从概念上讲是在训练和预测期间激活的神经元之间的连接强度。通常，神经网络根据非线性、分布式、并行的以及本地处理和适应的原理而操作。

[0045] 如这里描述的，术语“多层感知器(MLP)”通常指的是将输入数据的集映射到适当的输出的集上的前馈人工神经网络。因此，MLP分类器可以从神经网络的输出来执行识别。MLP可以包括有向图中的节点的层的序列，各层连接到下一层。除了输入节点之外，各个节点是可以具有非线性激活函数的神经元(或处理元素)。

[0046] 通常，“卷积”是对两个函数f和g的数学操作，生成可以视为原始函数之一的修改版本的第三函数，给出两个函数之间的重叠区域作为原始函数之一被转化的量的函数。

[0047] 通常，术语“卷积神经网络”是前馈型人工神经网络，其中，按照个体神经元响应于视野中的重叠区域的方式来平铺个体神经元。卷积网络是多层感知器(MLP)的变量，并且被设计为使用最小量的预处理。当用于图像识别时，卷积神经网络使用小神经元集合的多个层，其中，所述小神经元集合查看被称为“接受域”的输入图像的小部分。然后，平铺这些集合的结果使得它们重叠以获得原始图像的更好表现。针对每个这样的层重复此操作。有利地，卷积神经网络可以使用在卷积层中共享的权重。因此，针对各层中的每个像素使用同一个滤波器(权重组)。这样既减小了需要的存储器大小又改善了性能。

[0048] 如这里讨论的，术语“图像”指数字化图像数据的二维阵列，然而，这仅是例示并且不受限制。如这里讨论的，提供给服务器的图像可以通过诸如照相机(未示出)的另一个设备来收集，可以用中间工具(诸如软件)来准备以用于训练，并被配置为模仿由客户端(例如)提供的图像的形式。能够以数字化形式(诸如从智能电话中的照相机)提供由客户端收集的图像。在示例性实施例中，训练图像的诸如维度和像素数的方面通常等于生成图像的维度和像素数。此外，仅出于将训练图像与客户端的操作期间收集的其他图像区别开来的目的，将用于训练服务器的图像称为“训练图像”。通过客户端收集的图像被称为“产品图像”以及其他类似有区别的术语。

[0049] 如这里讨论的，“目标”可以出现在或者包含在图像内。例如，汽车(目标)可以出现在风景(目标的集合)的图片(图像)中。

[0050] 如这里讨论的，术语“程序”、“软件”、“应用”、“插件”和其他类似的术语指非暂时性机器可读介质上存储的机器可执行指令。机器可执行指令提供通过计算资源的控制和任何认为适合的相关联组件来执行方法。

[0051] 在示例性实施例中，方法和设备对于根据图像数据执行目标识别来说是有用的。示例性实施例应被认为仅说明性的，并且不限于这里的教导。因此，公开的方法和设备可以同样良好地用于其他应用。

[0052] 图1描述了服务器系统110(也称为服务器110)和客户端设备310(也称为客户端310)的方面。缩简的流程图与服务器110和客户端310中的每一者相关联。第一流程图140描述了如通过服务器110执行的训练神经网络140的方面。第二流程图340描述了如通过客户端310执行的识别图像的方面。

[0053] 图1示出示例性服务器系统110的方面。流程图140介绍针对服务器系统110的示例性操作。训练图像102可以可选择地存储在训练数据库104中。简单地说，可以使用训练图像102来训练服务器系统110，其中，服务器系统110可以从该训练图像102中生成学习后的权重的配置。还要在此更详细地讨论训练。

[0054] 服务器系统110包括输入设备105，其中，该输入设备105被配置为诸如从训练数据库104接收训练图像102。输入设备105可以包括诸如扫描器的数字转换器或者其他类似的设备。可以以类似的其他方式来实现输入设备105。

[0055] 服务器系统110还包括多层感知器(MLP)分类器137。神经网络(NN)123和MLP分类器137可以使用诸如“反向传播(BP)”的监督学习技术来训练神经网络(NN)123。在一些实施例中，MLP分类器137是标准线性感知器的变型，甚至可以区分不是线性可分离的数据。作为神经网络，可以认为MLP分类器137是神经网络(NN)123的部分。因此，在图1中成组地一起示出神经网络(NN)123和MLP分类器137。

[0056] 服务器系统110还包括可以存储程序132的存储器130。另外，服务器系统110包括处理器120和神经网络(NN)123。数据135可以存储在存储器130中并且可以存储在神经网络(NN)123中。

[0057] 另外，图1示出样本客户端设备310的方面。与客户端设备相关联的是介绍针对客户端设备310的示例性操作的流程图340。简单地说，客户端设备310可以根据输出操作244进行接收。输出操作244提供由服务器系统110生成的学习后的权重的配置。因此，一旦适当地配置，客户端设备310可以能够识别输入图像302中的训练图像102。

[0058] 客户端设备310包括被配置为接收输入图像302的输入设备305，以用于执行图像识别。客户端设备310还包括可以存储程序332的存储器330、处理器320和神经网络(NN)323。数据335可以存储在存储器330中，也可以存储在神经网络(NN)323中。

[0059] 客户端设备310还可以包括多层感知器(MLP)分类器337。MLP分类器337可以从神经网络(NN)323的输出执行识别。如上所述，MLP可以包括有向图中的节点的层的序列，各层连接到下一层。除了输入节点之外，各个节点是可以具有非线性激活函数的神经元(或处理元素)。神经网络(NN)323和MLP分类器337可以使用被称为反向传播(BP)的监督学习技术来训练网络。作为神经网络，MLP分类器337可以认为是神经网络(NN)323的部分，或者作为单独的模块。在这里给出的实施例中，给出作为单独模块的MLP分类器337，从而能够更好地描述神经网络(NN)323和神经网络(NN)123之间的共性。

[0060] 在服务器系统110和客户端设备310的框图中示出一些相似性。虽然有这些相似性，但是一些功能和要求可以非常不同。例如，服务器系统110不一定是便携的。更具体而言，服务器系统110可以具有大容量存储器130和用于生成配置的大量处理120。可以以各种形式给出客户端设备310，一些客户端设备310可以是便携的或不是便携的。客户端设备310可以是便携式个人电子设备且其还包括触摸屏。客户端设备310的示例包括智能电话(诸如来自加州的库比蒂诺的苹果公司的iPhone，或者来自加州的山景城的谷歌公司的实施安卓操作系统的设备)、平板电脑或其他类似设备。

[0061] 神经网络(NN)123和323包括人工智能(AI)技术，并且可以分别用于学习和识别。本说明书主要从作为卷积神经网络的神经网络(NN)123和323的方面展开描述，但是可以使用其他类型的神经网络。

[0062] 在神经网络(NN)123和323的一些实施例中，使用单层。单层可以包含许多神经元。每个神经元可以认为是处理单元。每个神经元可以执行诸如非线性变换的变换。在层内的神经元中存在加权连接。加权连接可以在神经网络的存储器中存储为权重值。学习算法可以被配置为学习权重值。

[0063] 在一些实施例中，可以使用深度学习。深度学习包括由类似于前述层的多个层组成的分层结构(hierarchical architecture)。此外，加权连接位于层内的神经元中并且还与所述多个层交叉。因此，深度学习需要存储器130的大容量。深度学习的示例包括反向传播(BP)、限制玻尔兹曼机、自动编码器解码等。现在描述针对神经网络(NN)的深度学习结构的示例。

[0064] 图2是描述神经网络(NN)423的示例性方面的框图。神经网络(NN)423是上述神经网络(NN)123、神经网络(NN)323或两者的示例。

[0065] 在图2的示例中，神经网络(NN)423至少包括第一卷积(C1)层454和第一子采样(S1)层458的第一层对421。神经网络(NN)423还包括第二卷积(C2)层464和第二子采样(S2)层468的第二层对422。虽然并不要求，但是对421可以类似于对422。在一些实施例中，对421和对422具有不同数量的特征映射。

[0066] 可以通过神经网络(NN)423(更具体而言，通过第一卷积层454)来接收图像402。因此，神经网络(NN)423可以因接收图像402而生成结果474。结果474是在已经完成了任意次迭代(iteration)之后，在第二子采样层468中生成的输出。在训练模式中，图像402是训练图像，并且结果474可以是所生成的用于识别其他输入图像中的图像402的学习后的权重。在目标识别模式中，图像402是输入图像，MLP分类器337可以使用结果474来表示输入图像402之内被识别的特定目标。现在，针对图3更详细地描述层对421和422。

[0067] 图3是示出卷积层554和子采样层558的样本层对521的图。对521的层554和子采样层558是对421，422或421和422两者的层的实施例的示例。根据接收图像502(针对该示例，可以包含权重核511和512)来解释这些实施例。

[0068] 通过存储特征映射531和532各自的权重，卷积层554可以包含特征映射531和532。实际上，可以存在比两个特征映射多得多的特征映射，但是为了简单起见，本图中仅示出了两个。在训练操作时，层554中的这些特征映射的每个特征映射可以是通过使权重核511、512与本地权重核卷积而生成的卷积操作结果。在目标识别操作时，层554中的这些特征映射的每个特征映射可以与输入图像卷积，以对生成特征映射的目标产生可能的识别。

[0069] 此外，子采样层558可以包含与卷积层554相同数量的特征映射。在这种情况下，子采样层558包含第一特征映射551和第二特征映射552。层558中的特征映射(551，552)中的每个可以是卷积层554的子采样特征映射的结果。子采样以结构化的方式减小特征映射的维度。子采样可以是对特征映射之内或横跨不同特征映射的神经元的最大子采样或平均子采样。

[0070] 特征映射531、532、551和552是真实的特征映射。然而，在本公开之内的一些实例中，因为它们还可以被另一个层处理，以生成特征映射的更新版本，所以可以将它们称为“中间”特征映射。

[0071] 现在返回到图1，针对服务器系统110和客户端设备310描述示例性的服务器至设备解决方案的方面。这些解决方案涉及对服务器系统110执行训练/学习处理，然后将用于神经网络(NN)123的学习后的配置(即，网络权重)输出到客户端设备310。该输出在图1中被示出为输出操作244。输出操作244将用于神经网络(NN)123的配置信息提供给客户端设备310。配置信息可以存在于存储器中，例如作为存储器130、数据存储(未示出)或另一个合适的存储库中的附加数据135，直到输出该配置信息为止。配置信息可以输出到诸如图1中示出的客户端设备310的各种客户端设备。客户端设备310可以是接收或输入神经网络配置的移动设备。

[0072] 将领会到，在一些实施例中，被诸如车载服务器系统110的客户端310使用的存储和计算的资源的至少一部分可以是远程的。训练神经网络的服务器系统110可以位于“云”(即，从客户端设备310中远程地实现服务器系统110的至少一部分设备)中。在一些实施例中，可以通过公共或非专用服务来提供服务器系统110。在一些实施例中，可以将客户端设备310配置为使得利用客户端设备310执行的目标识别基本上等同于通过服务器系统110执行的目标识别。即，在一些实施例中，以充足的训练信息来将客户端设备310配置为可以执行对更多训练图像的目标识别而基本上不降低性能。可以实时测量比较性能，这对于移动设备来说会是有益的。

[0073] 在一些实施例中，可以降低对客户端设备310的计算需求。例如，在一些实施例中，较小的位分辨率可能足以用于存储和处理网络权重。可以不需要原始的双精度。这种节约可以使神经网络323能够以低精度计算操作。在一些实施例中，可以使用模拟存储器。

[0074] 现在，使用图1的流程图140来描述示例性训练处理。这些训练处理可以用于服务器系统110或其他系统。将仅作为示例来根据服务器系统110的组件描述流程图140的处理。

[0075] 根据可选择的操作141，可以在使用训练图像102之前对神经网络123的部分或全部进行初始化。在使用层对(诸如图3的层对)的实施例中，可以在应用训练图像之前将第一卷积层初始化。例如，该初始化可以具有随机权重或另一个非图像的特定配置。

[0076] 根据另一个操作142，例如，可以经由输入设备105接收训练图像。在图1的示例中，接收到的训练图像是可能来自训练图像数据库104中的图像102。为了训练不同的图像，可以将操作140重复许多次。为了训练在训练图像数据库104中的所有图像，可以将操作140重复许多次。

[0077] 根据另一个操作143，可以按照将简短描述的方式来训练神经网络(NN)123。根据另一个操作144，可以输出在操作143之后生成的神经网络(NN)123的配置。可以按照与之前描述的输出操作244基本上类似或相同的方式来执行操作144。

[0078] 图4更详细地描述了操作143的训练的示例性实施例。给出的实施例更详细地描述了具有减小的位分辨率的学习权重。这些实施例中的一些实施例被称为“监督迭代量化(S.I.Q.)”。

[0079] 在图4中，示出训练643的示例性方法。更具体地，图4描述用于训练神经网络来生成具有减小的位分辨率的权重的示例性方面。训练643的方法的结果是量化权重(quantized weights)，所述量化权重被输出到客户端设备310以用于识别。用于输出和特征映射的位分辨率的减小在训练中不是强制性的。

[0080] 在该示例中，训练643的方法包括训练通过操作610，其中，根据该训练通过操作610借助用于神经网络(NN)123的层对来执行训练通过。训练通过操作610包括量化操作620，其中，根据该量化操作610在一个或更多个接合点(juncture)处执行量化。因此，量化操作620以将要被简短地说明的方式来修改训练通过操作610。

[0081] 根据另一个可选择的操作630，进行关于是否已经达到最大迭代数量的询问。如果回答为“是”，则结束训练643的方法的运行。如果回答为“否”，则根据初始化操作640，可以将神经网络(NN)123的部分或全部再次进行初始化。一旦重新初始化，神经网络(NN)123可以应用通过量化操作620量化的权重的版本。

[0082] 在初始化操作640之后，运行可以返回到训练通过操作610等。可以执行若干次训练643的示例性方法的迭代。

[0083] 现在，更详细地描述训练通过操作610和嵌入在训练通过操作610中的量化操作620。说明书具体涉及使用具有层对(诸如图2中给出的层对)的神经网络(NN)的服务器系统110的实施例。

[0084] 图5是描述针对服务器系统110的实施例的示例性神经网络(NN)723的量化操作620的方面的框图。在该示例中，神经网络(NN)723类似于上述的神经网络(NN)123。

[0085] 参照图5，在示例性量化操作720中，神经网络(NN)723接收训练图像702，并且生成结果774。结果774代表生成的学习后的权重，该学习后的权重稍后可以用于识别训练图像702是否包含某个目标，在训练处理中限定了该目标的标号。神经网络(NN)723至少包括第一卷积(C1)层754和第一子采样(S1)层758的第一层对721。神经网络(NN)723还包括第二卷积(C2)层764和第二子采样(S2)层768的第二层对722。可以如上所述地制作层对721和722。

[0086] 在图5中，箭头示出训练通过操作610和量化操作620的效果。在实施例中，首先利用随机权重或另一个非图像的特定配置来将神经网络(NN)723初始化。将已经被输入设备接收的训练图像702应用于第一卷积层754。如上所述，第一卷积层754已经在训练图像702应用到它之前被初始化了。

[0087] 神经网络(NN)723可以被配置为响应于被应用到第一卷积层754的训练图像702而在第一子采样层758中生成第一中间特征映射。可以基于第一中间原始权重来生成该第一中间特征映射。

[0088] 作为该训练通过步骤的部分，第一中间特征映射可以应用于第二卷积层764。神经网络(NN)723可以被配置为响应于被应用到第二卷积层764的第一中间特征映射而在第二子采样层768中生成第二中间特征映射。该第二中间特征映射可以具有第二中间原始权重。

[0089] 可以使用MLP分类器137将得出的特征映射768用于作出关于什么目标最有可能出现在图像中的预测。基于预测和地面实况的误差，可以使用反向传播(BP)来调整(即，学习/

训练)权重。

[0090] 可以针对训练数据库中的训练图像的全部或部分重复训练处理。在那之后,使用基于k-means(稍后将详细描述)的自适应量化方法来量化学习后的权重。在仍然保留足够的原始数据的同时,量化减少权重的位数。因此,在这种具体情况下,量化将原始权重转换为低分辨率权重。

[0091] 在一些实施例中,将训练权重的处理和其后的量化视为一个训练时代(training epoch)。

[0092] 如果训练时代小于阈值(例如,在操作630处的回答为“否”),则可以重复训练通过操作610。在这种情况下,可以使用量化权重将神经网络(NN)723初始化。基于该初始化,神经网络(NN)723可以再次接收训练图像的全部或部分,使用反向传播(BP)来学习权重,然后量化学习后的权重。

[0093] 如果训练时代等于或大于阈值(即,在操作630处的回答为“是”),则可以输出量化权重。

[0094] 监督迭代量化(S.I.Q)的处理(如关于图5的示例性过程概述的)涉及训练图像的反向传播(BP)、自适应量化和迭代。

[0095] 在示例性实施例中,可以在由操作626的箭头示出的位置处可选择地对输入进行量化;可以在由操作621的箭头示出的位置处可选择地将第一中间特征映射量化;然后,可以在由操作627的箭头示出的位置处可选择地将第二中间特征映射量化。此外,第一卷积层可以被配置为利用在第一训练通过处生成的第二中间特征映射的版本而变得初始化。神经网络还可以被配置为响应于将接收到的训练图像再次应用于第一卷积层而在第一子采样层中生成更新的第一中间特征映射。处理器还可以被配置为执行更新的第一中间特征映射的量化,以生成被配置为应用于第二卷积层的更新的且量化的第一中间特征映射。神经网络可以被配置为响应于将更新的且量化的第一中间特征映射应用于第二卷积层而在第二子采样层中生成更新的第二中间特征映射。

[0096] 现在,更详细地描述量化。量化是在仍然试图保留矢量或矩阵的一些属性的同时,用于减少矢量或矩阵的位数的方法。在这种具体情况下,矢量或矩阵可以是连接的权重、输入图像或特征映射。减少的位数可以减小计算的复杂性。在量化方法中,分析原始权重值的数据分布,并且生成代替原始权重值而使用的量化权重值。在这些量化权重值的组彼此相似之处,发生计算上的节约。

[0097] 在示例性实施例中,自适应量化是由原始权重值计算量化权重值的量化方法。现在描述示例。

[0098] 在一维中,可以寻求使方程表达式(诸如图6的方程表达式)最小化的量化权重值。图6没有精确地示出方程,而是示出了其值将要被最小化的方程表达式。

[0099] 图7中示出量化的一维示例,其中,横轴示出原始权重值(w)在0和1之间的分布。这些原始的权重值(w)在它们自身当中形成群集,因此量化权重值(Q)计算为0.262、0.428、0.603和0.866。因为权重值(w)分布以25-30个原始权重值(其值不同)开始,并且被量化权重值(Q)(其值在组中相似)代替,所以实现了节约。

[0100] 图8中示出量化的二维示例。特征映射831A可以具有由二维中的点表示的值。可以设置用于量化的数量K,在这种情况下,对于特征映射831A中的点的三个主要群集,数量K可

以为3(K＝3)。可以通过操作820生成量化的特征映射831B,量化的特征映射831B的值可以由三个星号表示,每个星号位于各个群集的中心附近。能够以许多方式发现星号的值。例如,星号可以在一些初始位置处开始,然后迭代(将每个原始值分配给最接近的量化值,然后将“量化值”移动到其群集的中心等)。

[0101] 图9示出将要被最小化的样本方程表达式9A、9B和9C。除了表达式9A还包括下标l(其中,针对两个层对l＝1,2)以外,表达式9A与用于自适应量化的图6的表达式类似。表达式9B可以用于流程图643,其中,F_l表示结合的卷积层Cl和子采样层Sl(l＝1,2)的组合操作。表达式9C可以等同于表达式9B。可以经由交替搜索:固定(w),求解(Q);固定(Q),求解(w)来求解该问题。

[0102] 从上述可以看出,因为由原始的权重值来计算新的量化权重值,所以自适应量化在计算上会是昂贵的。当计算需要迭代搜索时,自适应量化甚至更加昂贵。如此,自适应量化在具有有限资源的客户端设备中不表现为片上实施。然而,对于具有大量资源的服务器系统来说,不存在同样的问题。

[0103] 另一类型的量化方法称为“均匀量化”。均匀量化通过典型的简单规则来选择量化值,然后确定许多原始值如何匹配每个量化值。因此,均匀量化在计算上不如自适应量化密集,但是结果不精确。现在描述一维示例。

[0104] 图10示出可以用于确定量化值Q的方程。根据需要,可以设置变量delta(Δ)。

[0105] 图11是用于示出图10的方程的应用的样本图。在该示例中,delta(Δ)的值设置为4,这将得到4个量化值。因此,这些量化值确定为0.125、0.375、0.625和0.875,而与横轴上的权重值的分布无关。

[0106] 在图11中,横轴中示出了原始权重值(w)在0和1之间的分布。(实际上,该分布与图7的分布相同)。然而,因为量化方法不同,所以在图7中结果是不同的。

[0107] 再次返回到图1,根据实施例的客户端设备310可以接收(例如,根据输出操作244来接收)发送的用于神经网络(NN)123的学习后的配置(即网络权重)。将基于训练图像102和数据库104中的许多其他训练图像来准备该配置。

[0108] 现在,使用流程图340(也在图1中)描述根据示例性实施例的用于识别图像的处理。这些处理可以用于客户端设备310或者客户端设备的其他配置。将仅作为示例来根据客户端设备340的组件描述流程图340的处理。

[0109] 根据可选择的输入操作344,输入神经网络(NN)123的配置。可以基本上按照之前描述的操作244来执行输入操作344。能够以许多方式来执行输入,例如,通过在网络上下载,从存储器设备加载等。神经网络(NN)123的配置可以输入到,例如,神经网络(NN)323和MLP分类器337。在其他情况下,神经网络(NN)123的配置可以在诸如制造的适当时间作为附加数据335存储在存储器330中。

[0110] 根据输入操作345,可以接收并输入输入图像。能够以许多方式(例如,经由被配置为接收输入图像302的输入设备305)来执行该步骤。

[0111] 根据第一生成操作346,可以生成第一中间特征映射。如稍后将更详细说明的,可以以许多方式(例如,通过将输入图像应用于第一卷积层)来执行该步骤。可选择地,可以在将接收的输入图像应用于第一卷积层之前执行接收到的输入图像的量化。

[0112] 根据另一个量化操作347,可以执行第一中间特征映射的量化。能够以许多方式进

行量化。优选地，另一个量化操作347是参照图10和图11描述的均匀量化的类型，并且对于客户端设备310来说不像由服务器系统110执行的自适应量化那样繁重。另一个量化操作347可以生成量化的第一中间特征映射。

[0113] 根据第二生成操作348，可以生成第二中间特征映射。如稍后将更详细说明的，可以以许多方式(例如，通过将量化的第一中间特征映射应用于第二卷积层)来执行该步骤。

[0114] 根据识别操作349，可以执行输入图像中的目标的识别。该识别将当然首先根据数学，并且可以通过第二中间特征映射执行。可选择地，可以在执行识别操作之前执行第二中间特征映射的量化。

[0115] 现在，针对使用具有层对(诸如图2的层对)的神经网络(NN)的客户端设备实施例来更详细地描述流程图340中提供的示例性处理的操作。

[0116] 图12是示出针对客户端设备的示例性神经网络(NN)1223和针对客户端设备的示例性MLP分类器1237的框图。在该示例中，神经网络(NN)1223与上述神经网络(NN)323基本上相似。如上所述，可以利用基于训练图像而准备的神经网络的配置，将神经网络(NN)1223和MLP分类器1237初始化。

[0117] 如图12中所示，神经网络(NN)1223接收输入图像1202，并且生成识别结果1249。识别结果1249表示训练图像是否在输入图像1202中。神经网络(NN)1223至少包括第一卷积(C1)层1254和第一子采样(S1)层1258的第一层对1221。神经网络(NN)1223还包括第二卷积(C2)层1264和第二子采样(S2)层1268的第二层对1222。层对1221和1222可以与以上关于其他层对的描述基本相同。

[0118] 在图12中，箭头示出流程图340的操作的效果。在示例性实施例中，将由输入设备接收的输入图像1202应用于第一卷积层1254。

[0119] 神经网络(NN)1223可以被配置为响应于被应用到第一卷积层1254的输入图像1202而在第一子采样层1258中生成第一中间特征映射。该第一中间特征映射可以具有第一中间原始权重。

[0120] 客户端的处理器(图12中未示出)可以被配置为执行第一中间特征映射的量化。该量化示出为操作1246，并且可以是统一量化或其他类型。量化可以生成量化的第一中间特征映射。因此，在此具体情况下，量化将第一中间原始权重转换为量化的第一中间特征映射的第一中间低分辨率权重。这可以包括从最后的输入图像1202和存储的学习后的配置权重两者获取的方面。

[0121] 量化的第一中间特征映射可以应用于第二卷积层1264。NN 1223可以被配置为响应于被应用到第二卷积层1264的量化的第一中间特征映射而在第二子采样层1268中生成第二中间特征映射。该第二中间特征映射可以具有第二中间原始权重。

[0122] 神经网络(NN)1223可以被配置为执行MLP分类器1237中的权重的量化，并且神经网络(NN)1223可以被配置为响应于应用到量化的权重MLP的第二特征映射而在MLP分类器1237中生成分类输出。

[0123] MLP分类器1237可以被配置为执行处于输入图像中的训练图像的识别。该识别可以通过第二中间特征映射来执行。执行该识别可以生成识别结果1249。

[0124] 此外，可以在由操作1256和1257的箭头示出的附加位置发生量化。具体而言，如操作1256的箭头所示，输入图像1202自身可以在被应用于第一卷积层1254之前被量化。然而，

如操作1257的箭头所示，第二中间特征映射可以在第二子采样层1268生成第二中间特征映射之后并且在执行识别操作之前被量化。此外，可以通过客户端设备的处理器来执行量化。

[0125] 现在描述实际模拟的示例。

[0126] 图13是用于手写识别的一组四十(40)个样本训练图像(MNIST手写数字)。为了MLP分类器，这些训练图像的最左列可以接收分类“0”，下一列接收分类“1”等等，共计10个类别。

[0127] 图14是按照以上用于执行模拟的样本卷积神经网络NN中的数据结构1401的图。图14未示出上述量化操作。

[0128] 数据结构1401包括已经通过以5×5核来卷积图像1402而生成的六(6)个特征映射(每个24×24像素)的集C11454。数据结构1401还包括已经通过以/2(除以2)比例来对集C1的特征映射进行子采样而生成的六(6)个特征映射(每个12×12像素)的集C21458。另外，数据结构1401包括已经通过与5×5核进行卷积而生成的十二(12)个特征映射(每个24×24像素)的集C21464。数据结构1401还包括已经通过以/2(除以2)比例来对集C1的特征映射进行子采样而生成的十二(12)个特征映射(每个12×12像素)的集C21468。此外，数据结构1401包括生成识别结果1449的MLP分类器的类别1437。

[0129] 可以离线执行训练，使用反向传播(BP)来学习权重。使用如上所述的监督迭代量化来将用于低位分辨率的学习后的权重量化。然后，考虑输入或特征映射的量化，执行识别的测试。根据分类器误差——越小越好，来评价性能。

[0130] 连接到输出神经元的分类器权重的维度远远大于卷积核权重，并且通常对于较高的分辨率需要更多的位。在这种情况下，将分类器的位分辨率设置为六(6)位的固定值。然后，参照性能评价卷积核的不同的位分辨率。

[0131] 图15是示出将实施例的模拟性能与现有技术对比的模拟结果的表。列是输入分辨率位的数量，行是权重分辨率位的数量。每行具有针对不同量化方法的三个子行，针对均匀量化的顶部子行，针对k-means自适应量化的中间子行以及底部子行。

[0132] 对于非常低的位分辨率，误差当然是非常高的。然而，将观察到，利用仅四位的分辨率，通过实施例实现了与利用第一子行和第二子行中的原始同样(低)的仅1.15％的误差。实际上，箭头指出了顶部子行中的原始低误差表值如何遇到用于仅四位分辨率的底部子行中的没有更逊色的表值。

[0133] 图16是用于将实施例的模拟性能与现有技术对比的条形图。条形图1610、1620、1643分别针对a)均匀量化，b)k-means自适应量化和c)实施例示出对低位分辨率(达到4位)的平均测试误差。条1643具有最小的误差。

[0134] 图17是用于将实施例的模拟性能与现有技术对比的另一个条形图。条形图1710、1720、1743分别针对a)均匀量化，b)k-means自适应量化和c)实施例示出对所有位分辨率的平均测试误差。条形图1743具有最小的误差。

[0135] 图18示出用于将实施例的模拟性能与现有技术对比的多组条形图。沿着横轴示出的每组三个条形图是针对同一误差。在每组中，a)最左条是针对均匀量化，b)中间条是针k-means自适应量化，c)最右条是每个实施例。纵轴示出要求满足误差水平(即，针对不能更高的误差水平)的总位数。对于示出的所有测试误差，最右条需要要求满足误差水平的最少总位数。

[0136] 上述设备和/或系统执行函数、处理和/或方法。这些函数、处理和/或方法可以通过包括逻辑电路的一个或更多个设备来实现。这种设备可以选择性地称为计算机等。其可以是单独的设备或计算机，诸如通用计算机，或具有一个或多个附加功能的设备的部分。

[0137] 应由系统设计者、制造商、用户或其他类似的相关方来对性能的标准做出判断。如这里使用的术语“大量的”通常涉及所得系统性能的充足。

[0138] 逻辑电路可以包括可能出于通用或专用的目的而编程的处理器，诸如微控制器、微处理器、数字信号处理器(DSP)等。示例可以包括处理器120和320。

[0139] 逻辑电路也可以包括诸如存储器的非临时性计算机可读存储介质。这种介质可以具有不同类型，包括但不限于易失性存储器、非易失性存储器(NVM)、只读存储器(ROM)；随机存取存储器(RAM)；磁盘存储介质；光学存储介质；智能卡、闪存装置等。示例可以包括存储器130和330。

[0140] 这些单独或与其他结合的存储介质可以具有存储在其上的数据。用于存储在存储介质中的数据的示例包括数据135和335。

[0141] 此外，这些存储介质可以存储处理器能够读取和执行的程序。更具体而言，程序可以包括以处理器在读取时能够执行的编码格式的指令。示例包括程序132和332。

[0142] 运行程序通过物理量的物理操作而执行，并且可能引起将要被执行的函数、处理、动作和/或方法，和/或使其他设备或组件或块执行这样的函数、处理、行动和/或方法的处理器。通常，仅为了方便起见，优选地实施和描述作为各种相互连接的有区别的软件模块或特征的程序。这些，连同数据一起被单独地并且也被共同地称为软件。在某些情况下，软件与硬件结合，以混合称为“固件”。

[0143] 此外，这里描述方法和算法。这些方法和算法不必与任何具体的逻辑装置或其他设备固有地联系。相反，这些方法和算法可以被程序有利地实施以供诸如通用计算机、专用计算机、微处理器等的计算机器使用。

[0144] 这种详细的描述包括位于至少一个计算机可读介质中的程序操作的流程图、显示图像、算法和符号表示。由于使用单组流程图来描述程序和方法两者而实现节约。所以，在流程图根据框描述方法的同时，它们也同时地描述程序。

[0145] 在上述方法中，可以按照动作的肯定步骤，或者使书面记载了能够发生的事情发生来执行每个操作。可以由整个系统或装置，或者由它的仅一个或多个组件来进行这样的动作或者使事情发生。此外，操作的顺序不被约束为所示出的顺序，并且根据不同的实施例可以是不同的顺序。此外，在某些实施例中，可以添加新的操作，或者可以修改或删除个别操作。添加的操作可以是例如来自于在主要描述不同的系统、设备、装置或方法的同时所提到的内容。

[0146] 本领域技术人员将能够根据被视为整体的本说明书实现本发明。包括详情以提供透彻的理解。在其他情况下，为了防止不必要地模糊本发明，没有描述公知的方面。加之，本说明书中对任何现有技术所做出的任何参考不是并且不应当被认为承认或以任何形式暗示该现有技术在任何国家形成公知常识的部分。

[0147] 本说明书包括一个或更多个示例，但是不限制发明可以如何实现。发明的实施例或示例的确可以根据所描述的内容来实现，或者也可以有差别地并且还与其他当前或未来的技术相结合地实现。其他实施例包括这里描述的特征的组合和子组合，包括例如，等同于

如下情况的实施例，即：以与描述的实施例不同的顺序提供或应用特征；从一个实施例中提取个别特征并且将该特征插入到另一个实施例中；从实施例中去除一个或多个特征；或者在提供这种组合和子组合中包括的特征的同时，既从一个实施例中去除特征又添加从另一个实施例中提取的特征。

[0148] 在本文件中，短语“被构造为”和/或“被配置为”表示构造和/或配置的一个或更多个实际状态，其根本上与前述的这些短语的元件或特征的物理特性相关联，如此，达到了远超过仅描述的预期用途。如本领域技术人员在看到本公开之后将明了的，可以以超出本文件中示出的任何示例的任意数量的方式来实现任何这些元件或特征。

[0149] 权利要求限定了被视为新颖的和非显而易见的元素、特征和步骤或操作的某些组合和子组合。针对其他这种组合和子组合的附加权利要求可以存在于本文件或相关文件中。

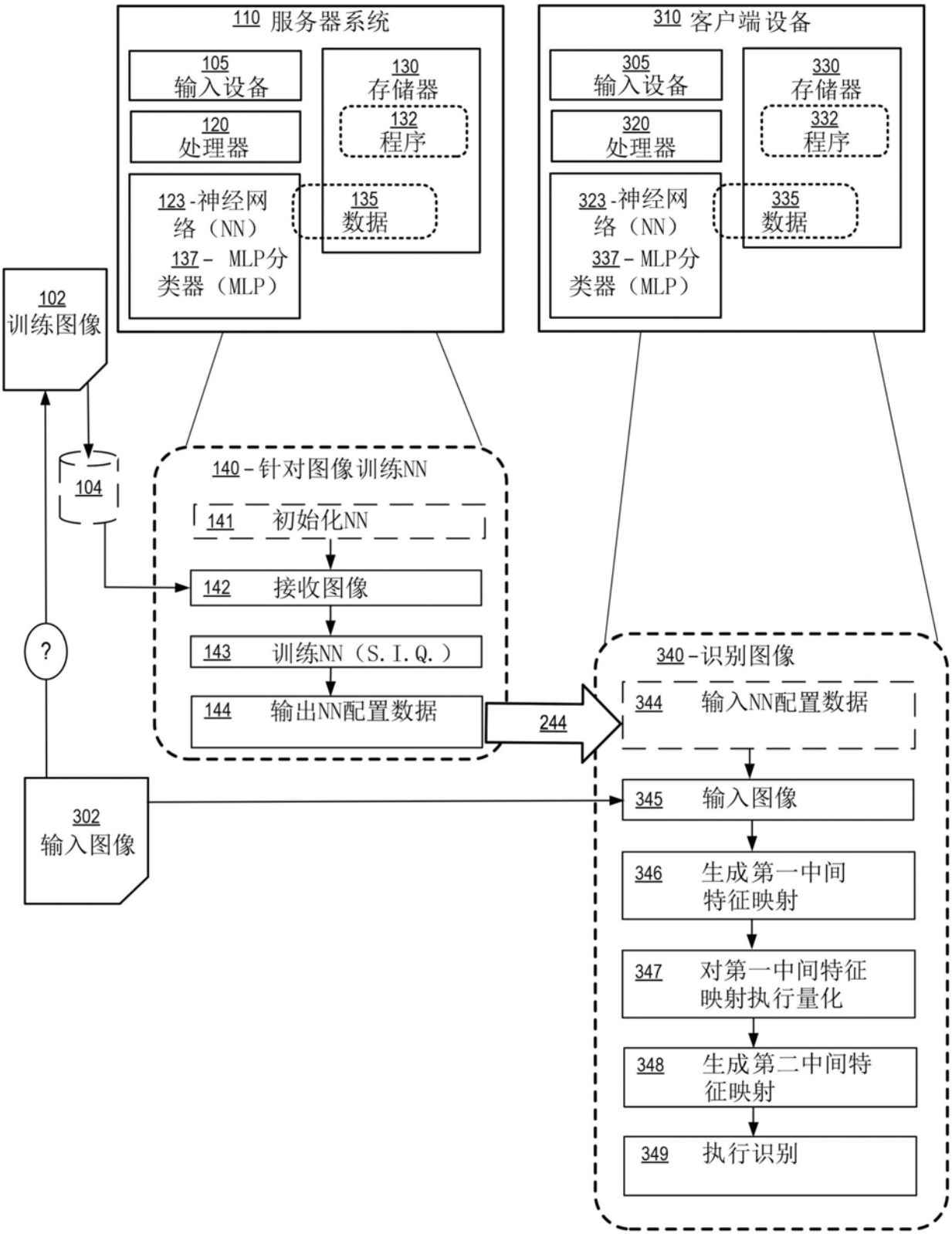

110 服务器系统
105 输入设备
120 处理器
123 -神经网络（NN）
137 - MLP分类器（MLP）
130 存储器
132 程序
135 数据
310 客户端设备
305 输入设备
320 处理器
323 -神经网络（NN）
337 - MLP分类器（MLP）
330 存储器
332 程序
335 数据
102 训练图像
104
140 - 针对图像训练NN
141 初始化NN
142 接收图像
143 训练NN（S. I. Q. ）
144 输出NN配置数据
244
?
302 输入图像
340 - 识别图像
344 输入NN配置数据
345 输入图像
346 生成第一中间特征映射
347 对第一中间特征映射执行量化
348 生成第二中间特征映射
349 执行识别

图1

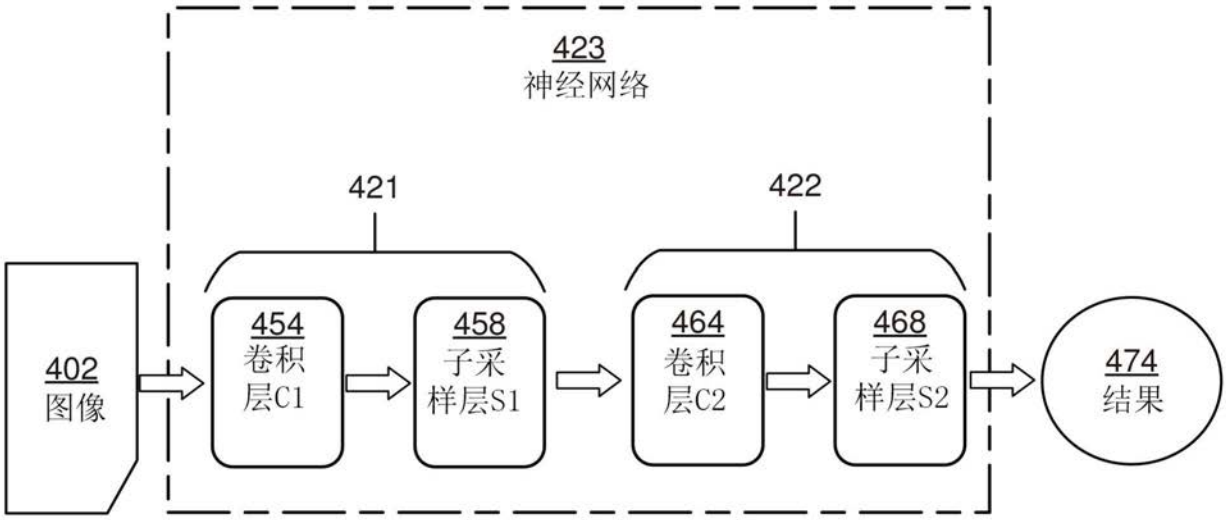

423
神经网络
421
422
402
图像
454
卷积
层C1
458
子采
样层S1
464
卷积
层C2
468
子采
样层S2
474
结果

图2

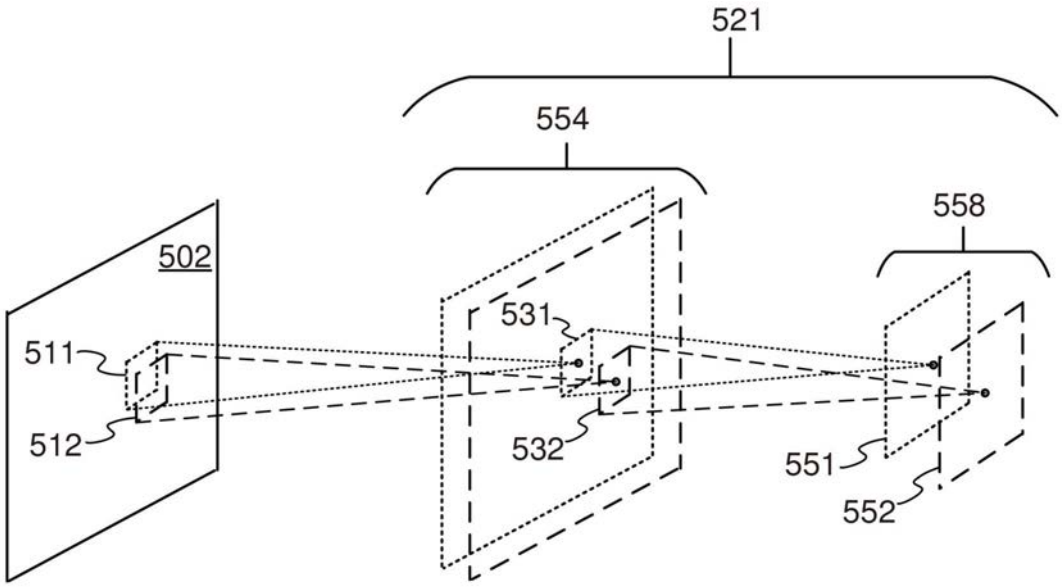

521
554
558
502
511
512
531
532
551
552

图3

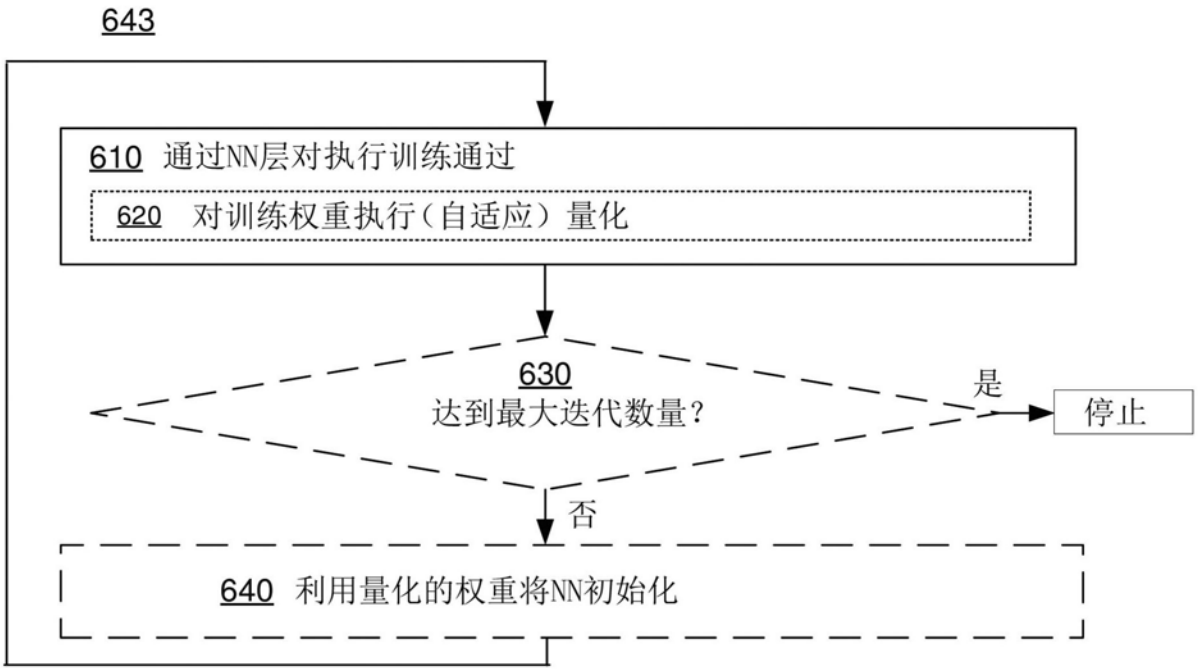

643
610 通过NN层对执行训练通过
620 对训练权重执行（自适应）量化
630
达到最大迭代数量？
是
停止
否
640 利用量化的权重将NN初始化
服务器训练—监督迭代量化

图4

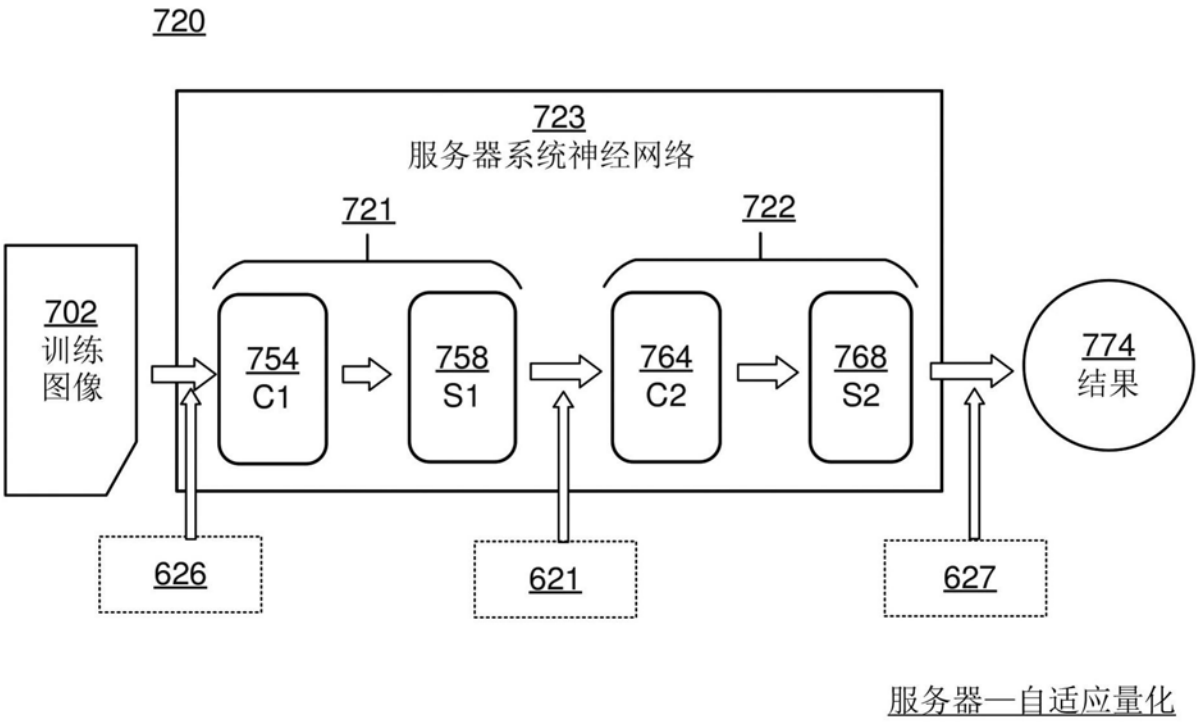
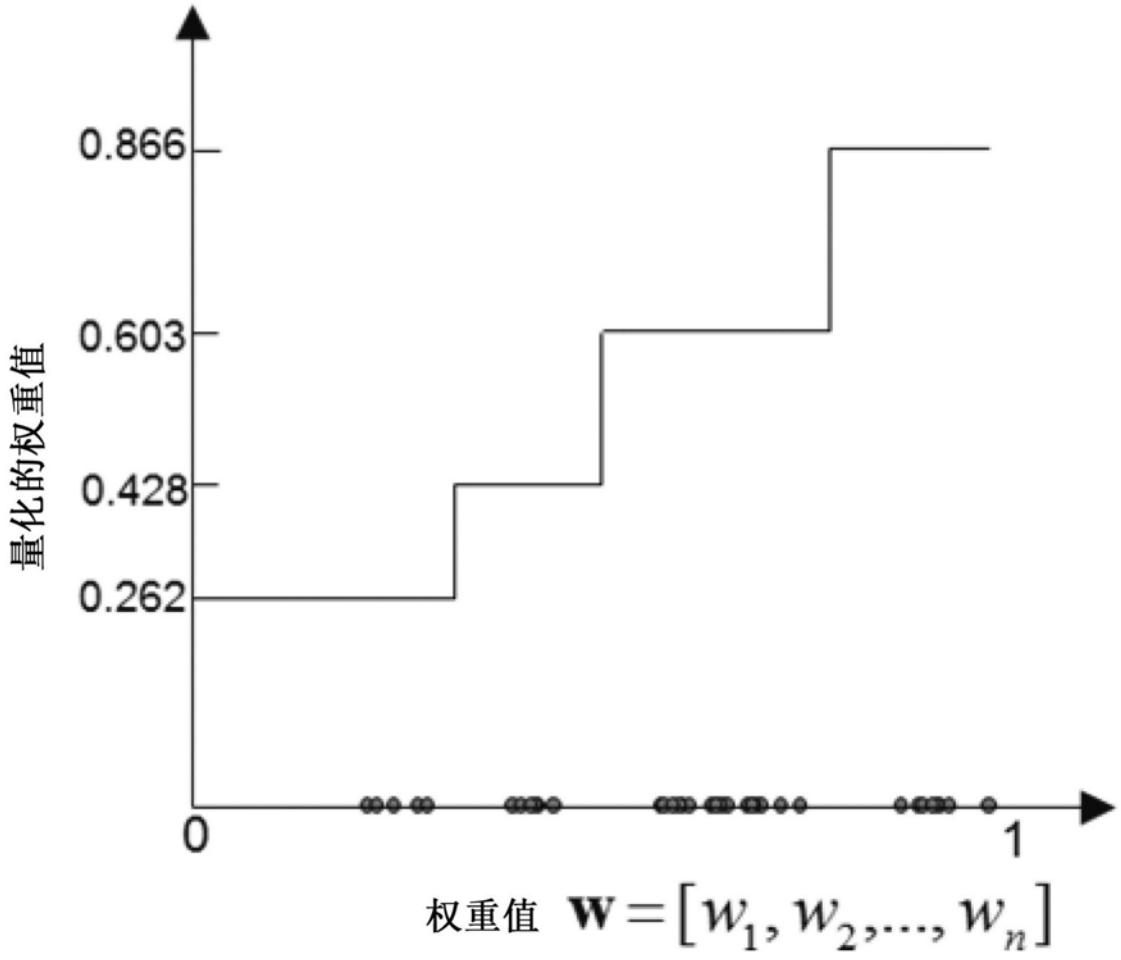

720
723
服务器系统神经网络
721
722
702
训练
图像
754
C1
758
S1
764
C2
768
S2
774
结果
626
621
627
服务器—自适应量化

图5

$$\min_{Q}\left\|Q(\mathbf{w})-\mathbf{w}\right\|_2^2$$

自适应量化

图6



自适应量化

图7

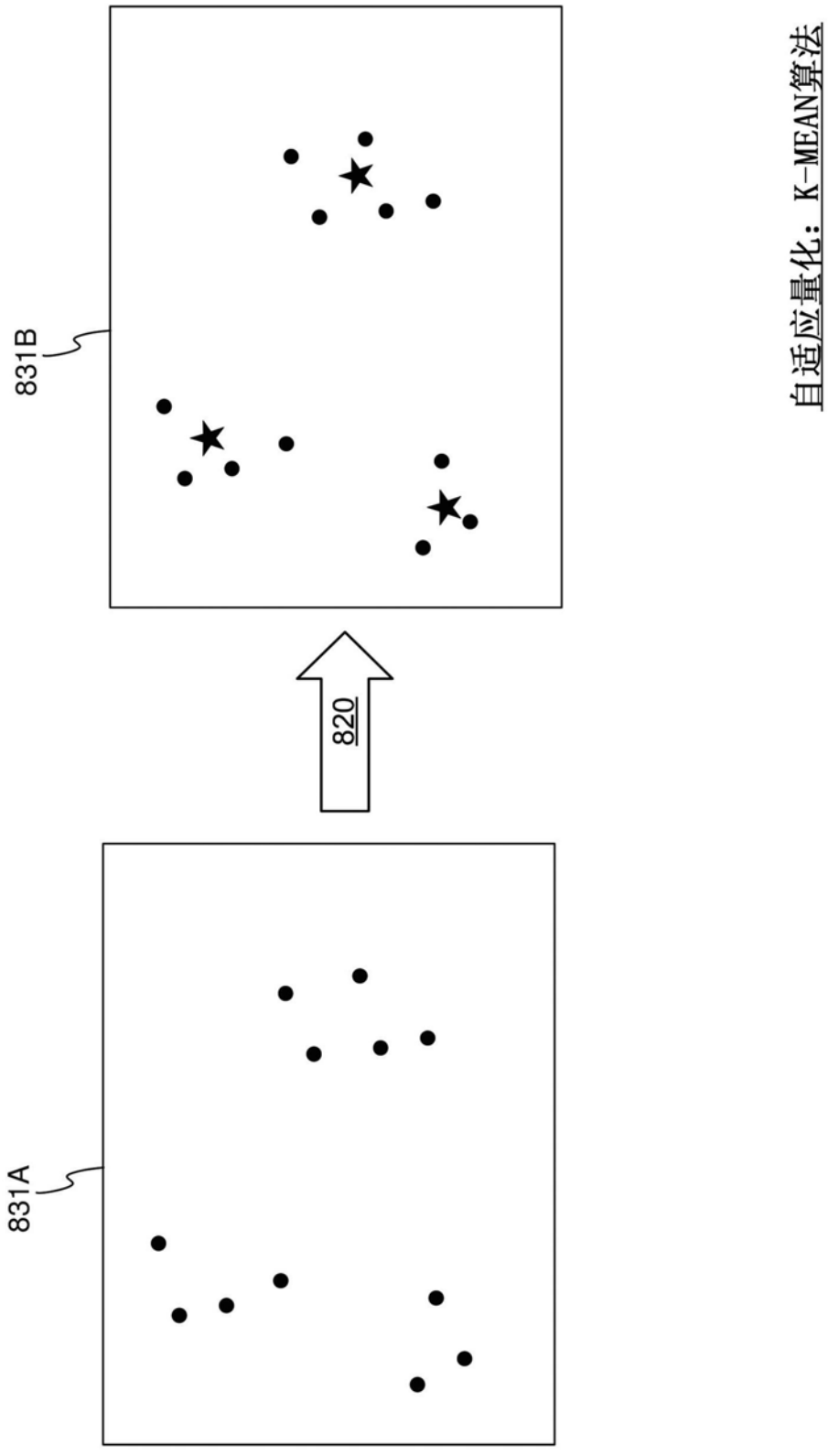


图8

(9A)

$$\min_{Q_l} \left\| Q_l(\mathbf{w}) - \mathbf{w} \right\|_2^2$$

(9B)

$$\min_{Q_l} \sum_t \left(y^{(t)} - Class\left(F_2\left(Q_2(\mathbf{w}), F_1\left(Q_1(\mathbf{w}), \mathbf{x}^{(t)} \right) \right) \right) \right)^2$$

(9C)

$$\min_{Q_l} \sum_t \alpha \left(y^{(t)} - Class\left(F_2\left(\mathbf{w}, F_1\left(\mathbf{w}, \mathbf{x}^{(t)} \right) \right) \right) \right)^2 + \sum_l \beta \left\| Q_l(\mathbf{w}) - \mathbf{w} \right\|_2^2,$$

图9

$$Q(\mathbf{w}) = \Delta \bullet \left(\left\lfloor \frac{\mathbf{w}}{\Delta} \right\rfloor + \frac{1}{2} \right)$$

均匀量化

图10

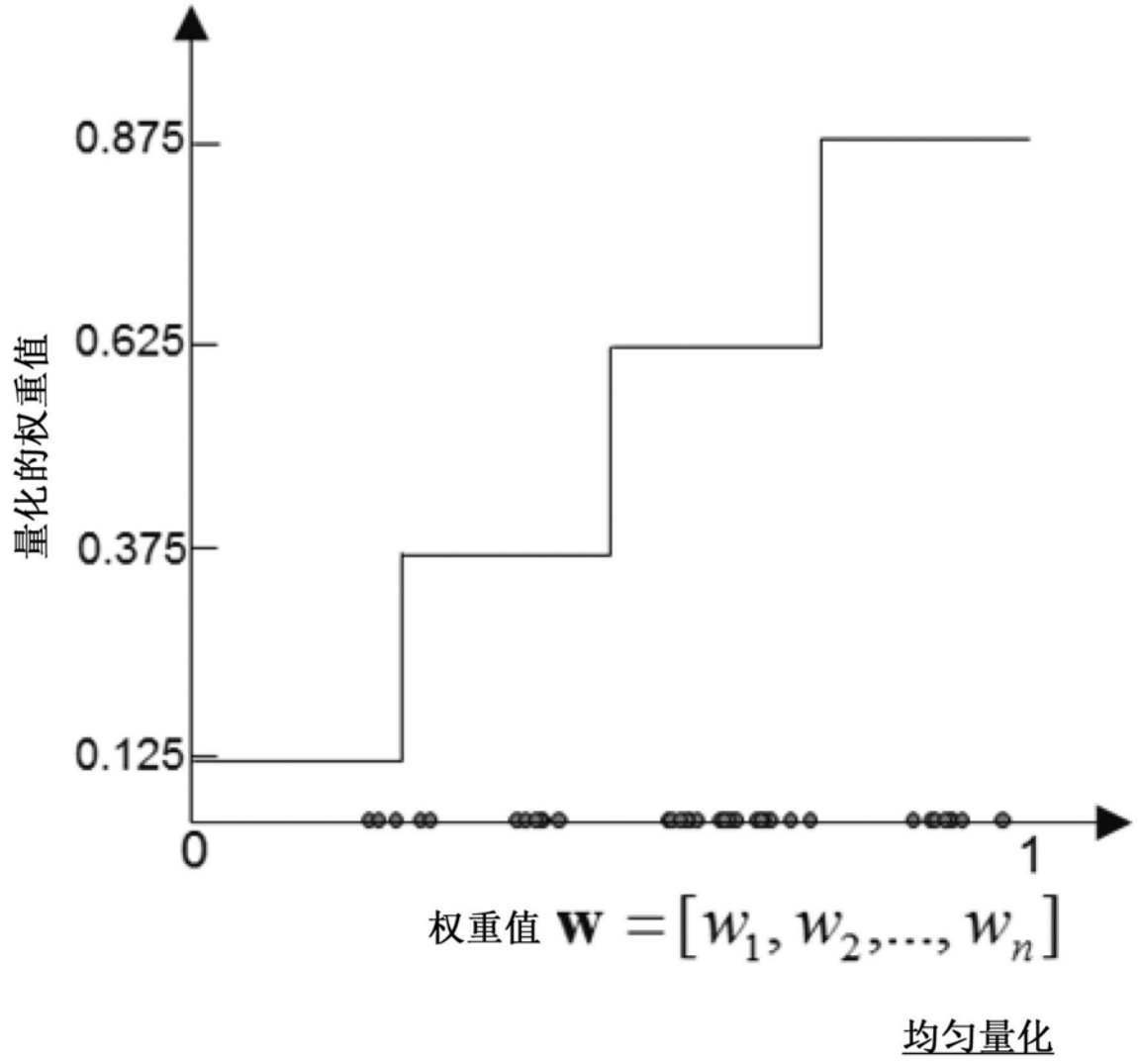
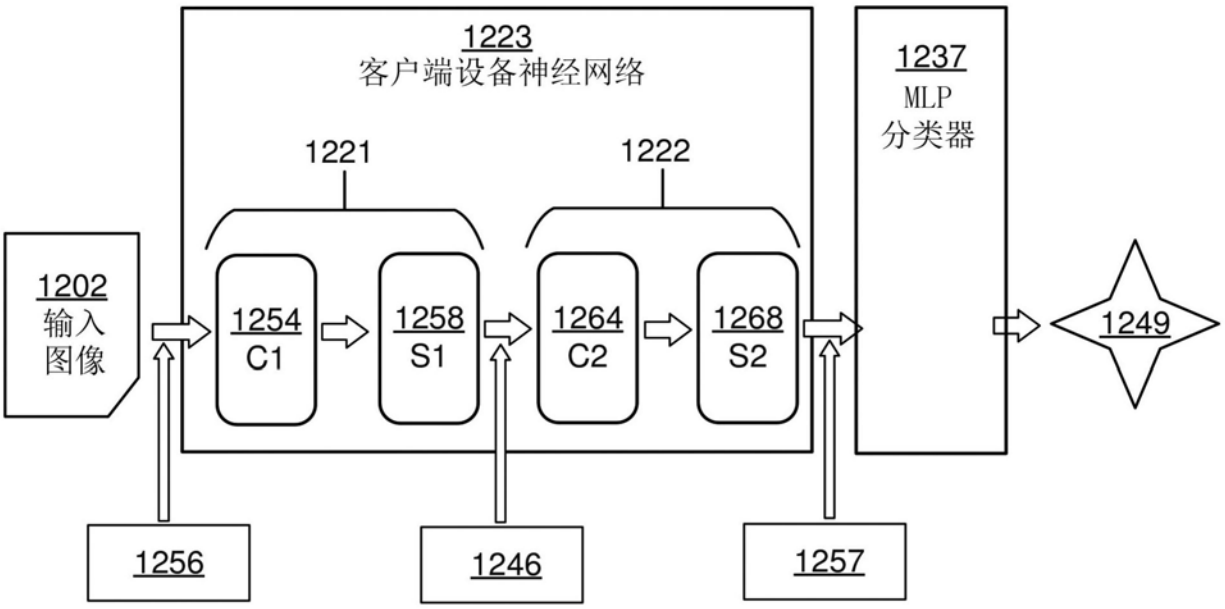


图11



客户端—利用均匀量化的识别

图12

样本训练图像（手写）

图13

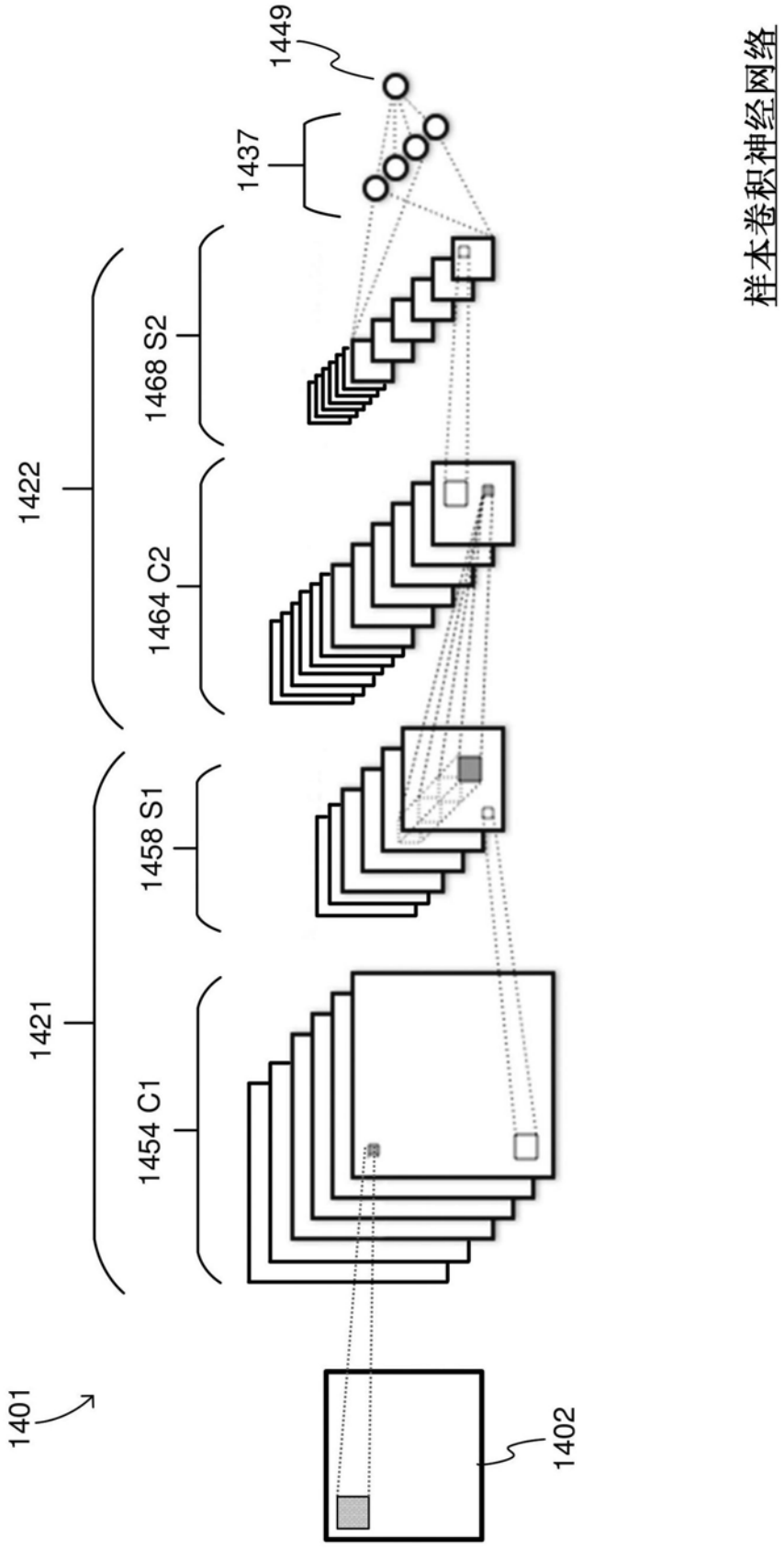
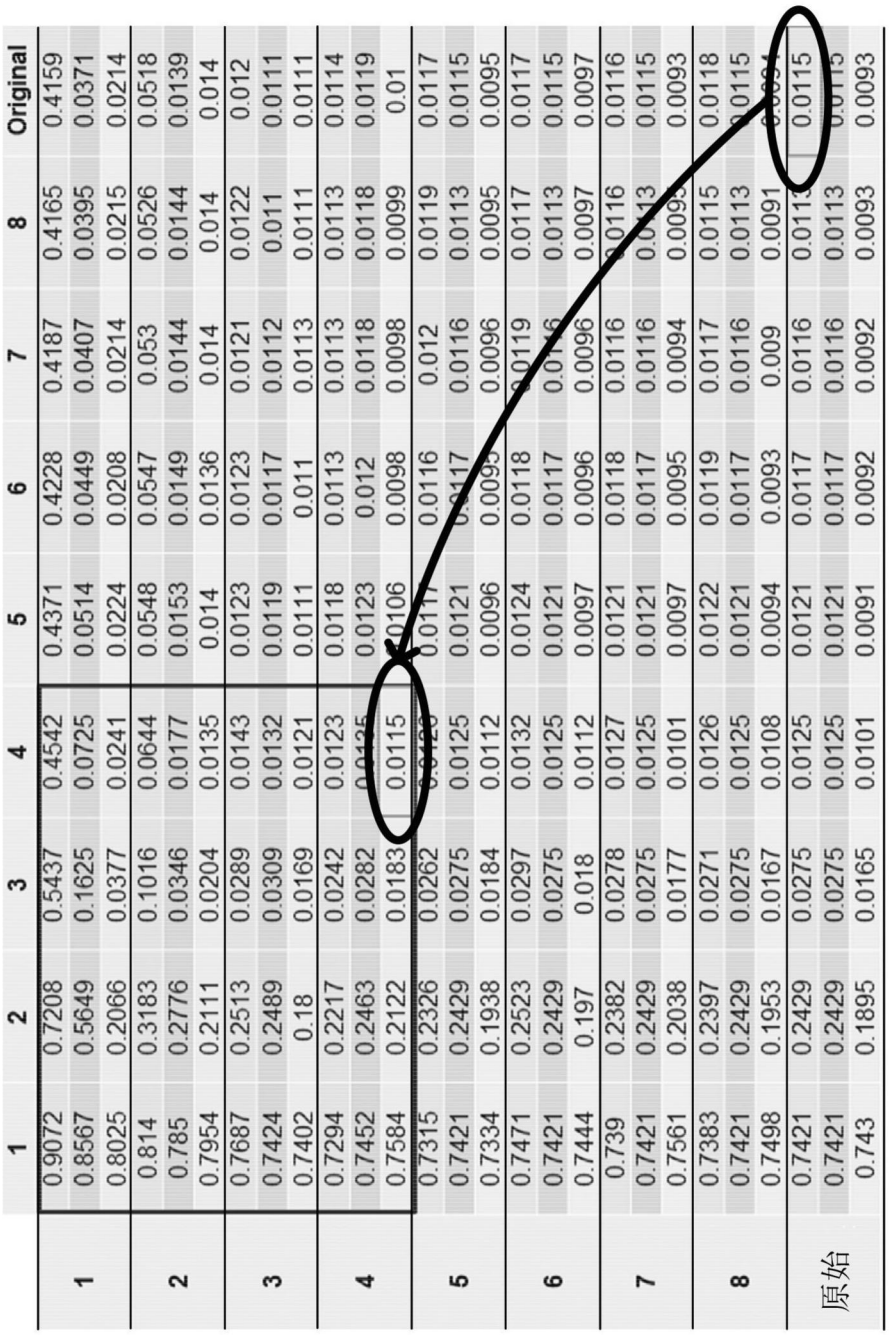

样本卷积神经网络
1449
1437
1468 S2
1422
1464 C2
1458 S1
1421
1454 C1
1401
1402

图14



	1	2	3	4	5	6	7	8	Original
1	0.9072	0.7208	0.5437	0.4542	0.4371	0.4228	0.4187	0.4165	0.4159
	0.8567	0.5649	0.1625	0.0725	0.0514	0.0449	0.0407	0.0395	0.0371
	0.8025	0.2066	0.0377	0.0241	0.0224	0.0208	0.0214	0.0215	0.0214
2	0.814	0.3183	0.1016	0.0644	0.0548	0.0547	0.053	0.0526	0.0518
	0.785	0.2776	0.0346	0.0177	0.0153	0.0149	0.0144	0.0144	0.0139
	0.7954	0.2111	0.0204	0.0135	0.014	0.0136	0.014	0.014	0.014
3	0.7687	0.2513	0.0289	0.0143	0.0123	0.0123	0.0121	0.0122	0.012
	0.7424	0.2489	0.0309	0.0132	0.0119	0.0117	0.0112	0.011	0.0111
	0.7402	0.18	0.0169	0.0121	0.0111	0.011	0.0113	0.0111	0.0111
4	0.7294	0.2217	0.0242	0.0123	0.0118	0.0113	0.0113	0.0113	0.0114
	0.7452	0.2463	0.0282	0.0125	0.0123	0.012	0.0118	0.0118	0.0119
	0.7584	0.2122	0.0183	0.0115	0.0106	0.0098	0.0098	0.0099	0.01
5	0.7315	0.2326	0.0262	0.0123	0.0117	0.0116	0.012	0.0119	0.0117
	0.7421	0.2429	0.0275	0.0125	0.0121	0.0117	0.0116	0.0113	0.0115
	0.7334	0.1938	0.0184	0.0112	0.0096	0.0095	0.0096	0.0095	0.0095
6	0.7471	0.2523	0.0297	0.0132	0.0124	0.0118	0.0119	0.0117	0.0117
	0.7421	0.2429	0.0275	0.0125	0.0121	0.0117	0.0116	0.0113	0.0115
	0.7444	0.197	0.018	0.0112	0.0097	0.0096	0.0096	0.0097	0.0097
7	0.739	0.2382	0.0278	0.0127	0.0121	0.0118	0.0116	0.0116	0.0116
	0.7421	0.2429	0.0275	0.0125	0.0121	0.0117	0.0116	0.0113	0.0115
	0.7561	0.2038	0.0177	0.0101	0.0097	0.0095	0.0094	0.0093	0.0093
8	0.7383	0.2397	0.0271	0.0126	0.0122	0.0119	0.0117	0.0115	0.0118
	0.7421	0.2429	0.0275	0.0125	0.0121	0.0117	0.0116	0.0113	0.0115
	0.7498	0.1953	0.0167	0.0108	0.0094	0.0093	0.009	0.0091	0.0094
原始	0.7421	0.2429	0.0275	0.0125	0.0121	0.0117	0.0116	0.0113	0.0115
	0.7421	0.2429	0.0275	0.0125	0.0121	0.0117	0.0116	0.0113	0.0115
	0.743	0.1895	0.0165	0.0101	0.0091	0.0092	0.0092	0.0093	0.0093

结果

图15

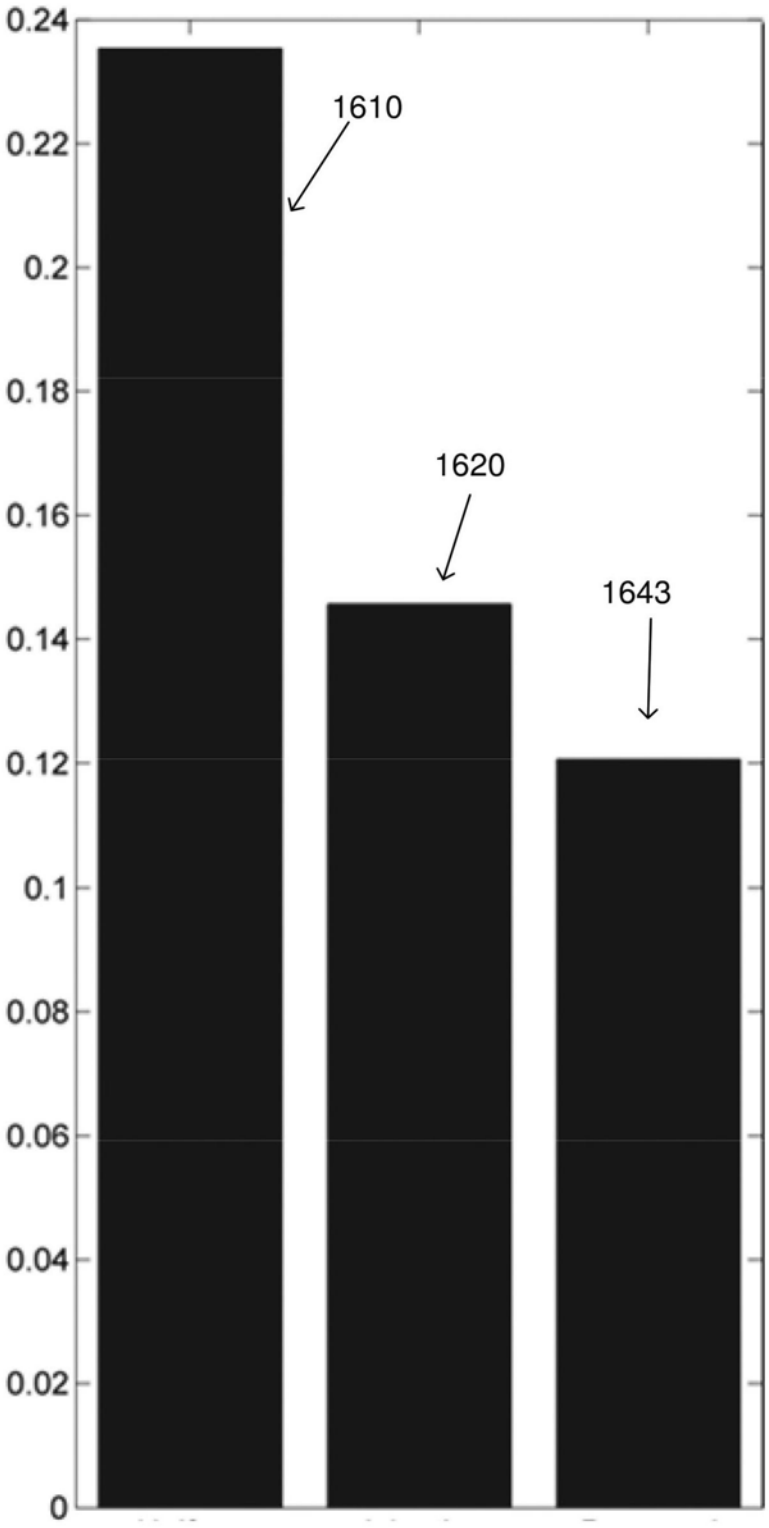

0.24
0.22
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1610
1620
1643
比较平均测试误差

图16

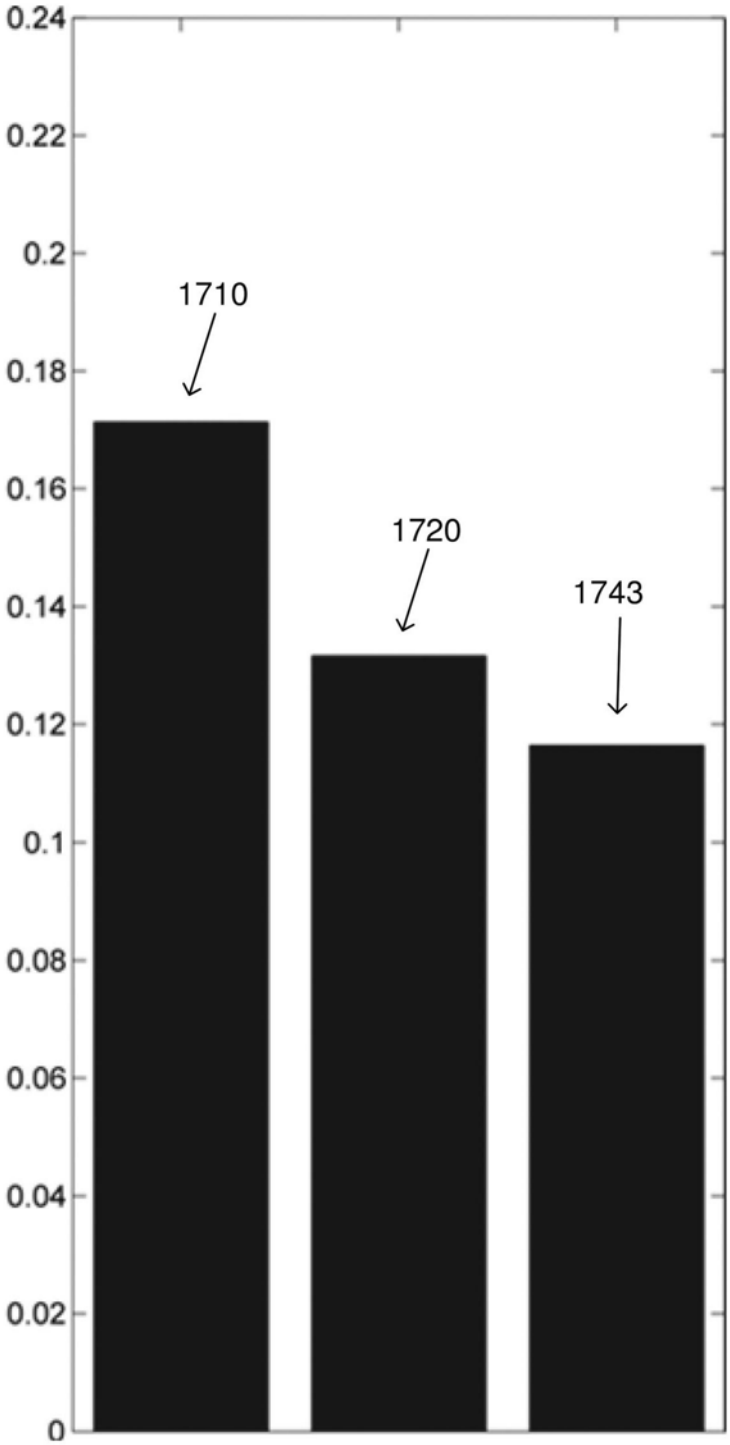

0.24
0.22
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1710
1720
1743
比较平均测试误差

图17

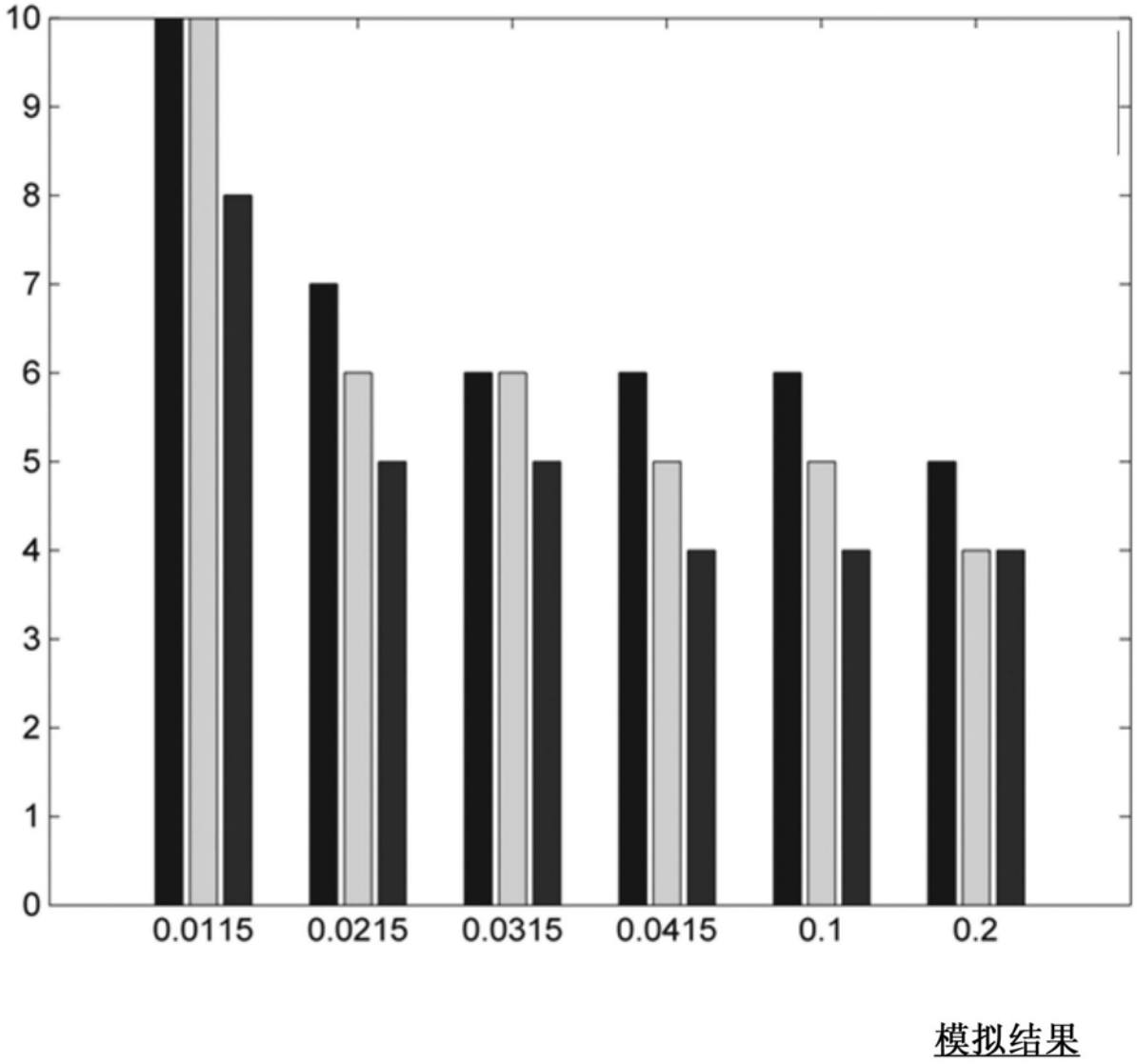


图18