(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 108510067 A
(43)申请公布日 2018.09.07

(21)申请号 201810319586.6

(22)申请日 2018.04.11

(71)申请人 西安电子科技大学
地址 710071 陕西省西安市雁塔区太白南路2号

(72)发明人 张犁 黄蓉 陈治宇 赵博然 牛毅 石光明

(74)专利代理机构 陕西电子工业专利中心 61205
代理人 王品华 朱红星

(51)Int.Cl.
G06N 3/063(2006.01)
G06N 3/04(2006.01)
G06K 9/62(2006.01)

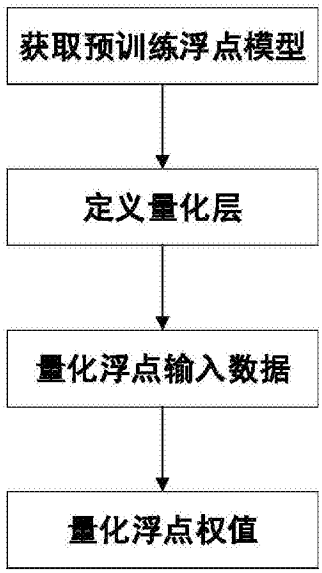
权利要求书1页 说明书5页 附图3页

(54)发明名称

基于工程化实现的卷积神经网络量化方法

(57)摘要

本发明公开了一种基于工程化实现的卷积神经网路量化方法，主要解决现有技术耗费时间长，准确率不高的问题，其实现方案是：1)下载已经预训练好的浮点格式的卷积神经网络模型；2)在下载的浮点网络中定义量化层；3)在下载的浮点网络中每一层批量归一化层后面调用2)定义的量化层，并构建输入数据的量化公式对浮点输入数据进行量化；4)在1)下载的浮点网络中，构建权值量化公式对浮点权值进行量化。本发明与现有技术相比，在保持识别准确率的同时降低了图像分类任务的时间成本和存储需求，可用于专用芯片FPGA/ASIC硬件平台的部署。



CN 108510067 A

1.一种基于工程化实现的卷积神经网络量化方法，包括：

(1)从互联网下载两个已经预训练好的浮点格式卷积神经网络模型；

(2)在(1)下载的预训练浮点模型的每一层卷积层和全连接层后面都添加一层自定义的量化层，并用该自定义的量化层对浮点形式的输入数据进行量化，量化的公式为：

$$Convert(x,\langle IL,FL\rangle)\begin{cases}-2^{IL-1} & 如果x\le -2^{IL-1}\\ 2^{IL-1}-2^{-FL} & 如果x\ge 2^{IL-1}-2^{-FL}\\ \dfrac{round(x\cdot 2^{FL})}{2^{FL}} & 其他\end{cases}$$

其中，Convert表示将浮点输入数据转化为定点输入数据，x为浮点输入数据，IL和FL分别表示定点输入数据的整数位宽和小数位宽，round为四舍五入函数，是编程语言的内置函数，2^{FL}表示量化成小数位宽为FL的定点数，-2^{IL-1}表示定点输入数据表示的数值范围的下限，$2^{IL-1}-2^{-FL}$表示定点输入数据表示的数值范围的上限；

(3)对(1)下载的预训练浮点模型中已经训练好的浮点权值进行量化，量化的公式为：

$$Convert(w,\langle IL',FL'\rangle)\begin{cases}-2^{IL'-1} & 如果w\le -2^{IL'-1}\\ 2^{IL'-1}-2^{-FL'} & 如果w\ge 2^{IL'-1}-2^{-FL'}\\ \dfrac{round(w\cdot 2^{FL'})}{2^{FL'}} & 其他\end{cases}$$

其中，Convert表示将浮点权值转化为定点权值，w为浮点权值，IL′和FL′分别表示定点权值的整数位宽和小数位宽，round为四舍五入函数，是编程语言的内置函数，$2^{FL'}$表示量化成小数位宽为FL′的定点数，$-2^{IL'-1}$表示定点权值表示的数值范围的下限，$2^{IL'-1}-2^{-FL'}$表示定点权值表示的数值范围的上限。

2.根据权利要求1所述的方法，其中步骤(2)中在步骤(1)下载的预训练浮点模型的每一层卷积层和全连接层后面都添加一层自定义的量化层，是利用编程语言python完成的，其步骤如下：

(2a)定义一个量化层，量化层对浮点输入数据进行量化，量化的定点数位宽用1+IL+FL表示，其中IL表示整数位宽，FL表示小数位宽；

(2b)在步骤(1)下载的预训练浮点网络的每一层卷积层和全连接层后面调用(2a)定义的量化层，用于网络的前向传播。

3.根据权利要求1所述的方法，其中步骤(1)从互联网下载两个已经预训练好的浮点格式卷积神经网络模型，包括：

由3层卷积层和2层全连接层组成的小型网络模型，

由13层卷积层和3层全连接层组成的大型网络模型。

基于工程化实现的卷积神经网络量化方法

技术领域

[0001] 本发明属于深度学习技术领域，具体涉及一种卷积神经网络量化方法，可用于专用芯片FPGA/ASIC硬件平台的部署。

背景技术

[0002] 深度学习近年来发展迅速，已经被广泛应用到各个领域，特别是计算机视觉、语音识别和自然语言处理领域。卷积神经网络是深度学习的代表，在计算机视觉领域掀起了热潮，凭借其强大的学习能力被广泛应用于图像分类任务中。为了提高图像分类任务的识别准确率，卷积神经网络的层数越来越多，结构越来越复杂。提高识别准确率的同时也付出了巨大的代价，计算复杂度和模型存储需求大量增加，这不利于卷积神经网络在功率预算有限的硬件平台的部署。因此，改进卷积神经网络的算法，降低卷积神经网络的存储需求已成为趋势，从而可以促进卷积神经网络在硬件平台FPGA和ASIC芯片上的应用。目前，将卷积神经网络使用的32位浮点数的数制量化成低位宽的定点数这种方法可以使得硬件资源占用和功耗更少。

[0003] Gupta,S.在其发表的论文“Deep learning with limited numerical precision”(《Computer Science》,2015)中提出了使用随机舍入的方法对卷积神经网络进行定点数的量化，该方法在网络量化位宽为16的时候也能取得与网络使用32位浮点数时几乎相同的性能。但是在硬件平台中随机数的实现特别复杂，所以该方法不易于部署在硬件平台上。

[0004] Rastegari M.在其发表的论文“XNOR-Net:ImageNet Classification Using Binary Convolutional Neural Networks”(European Conference on Computer Vision,2016:525-542)中提出了XNOR-Net，XNOR-Net将卷积神经网络量化成了二值网络，量化位宽为1，该方法虽说能最大程度地降低硬件占用的资源和消耗的功率，实现起来非常的高效。但是对大规模图像数据集imagenet做分类任务时，该方法的识别准确率与网络使用32位浮点数时得到的识别准确率相比下降超过了10%。

发明内容

[0005] 本发明的目的在于针对上述现有技术的问题，提出一种基于工程化实现的卷积神经网络量化方法，以在保持识别准确率的同时降低图像分类任务的时间成本和存储需求。

[0006] 本发明的基本思路是：根据硬件平台处理器的位宽将卷积神经网络量化为位宽为16、8的定点网络，对定点数整数和小数进行不同的位宽组合，对量化后的定点网络进行测试，根据测试准确率选择最适合部署在硬件平台的定点数位宽和表示格式，其实现方案包括如下：

[0007] (1)从互联网下载两个已经预训练好的浮点格式卷积神经网络模型；

[0008] (2)在(1)下载的预训练浮点模型的每一层卷积层和全连接层后面都添加一层自定义的量化层，并用该自定义的量化层对浮点形式的输入数据进行量化，量化的公式为：

[0009] $$Convert(x,\langle IL,FL\rangle)\begin{cases}-2^{IL-1} & 如果x\le -2^{IL-1}\\ 2^{IL-1}-2^{-FL} & 如果x\ge 2^{IL-1}-2^{-FL}\\ \dfrac{round(x\cdot 2^{FL})}{2^{FL}} & 其他\end{cases}$$

[0010] 其中，Convert表示将浮点输入数据转化为定点输入数据，x为浮点输入数据，IL和FL分别表示定点输入数据的整数位宽和小数位宽，round为四舍五入函数，是编程语言的内置函数，2^{FL}表示量化成小数位宽为FL的定点数，-2^{IL-1}表示定点输入数据表示的数值范围的下限，$2^{IL-1}-2^{-FL}$表示定点输入数据表示的数值范围的上限；

[0011] (3)对(1)下载的预训练浮点模型中已经训练好的浮点权值进行量化，量化的公式为：

[0012] $$Convert(w,\langle IL',FL'\rangle)\begin{cases}-2^{IL'-1} & 如果w\le -2^{IL'-1}\\ 2^{IL'-1}-2^{-FL'} & 如果w\ge 2^{IL'-1}-2^{-FL'}\\ \dfrac{round(w\cdot 2^{FL'})}{2^{FL'}} & 其他\end{cases}$$

[0013] 其中，Convert表示将浮点权值转化为定点权值，w为浮点权值，IL′和FL′分别表示定点权值的整数位宽和小数位宽，round为四舍五入函数，是编程语言的内置函数，$2^{FL'}$表示量化成小数位宽为FL′的定点数，$-2^{IL'-1}$表示定点权值表示的数值范围的下限，$2^{IL'-1}-2^{-FL'}$表示定点权值表示的数值范围的上限。

[0014] 本发明与现有技术相比有以下优点：

[0015] 第一、由于本发明将已经预训练好的的浮点卷积神经网络量化为定点卷积神经网络，数据经过量化由高位宽变为低位宽，充分地降低了时间成本。

[0016] 第二、由于只需对网络的前向传播过程进行量化，网络模型设计简单，易于实现。

附图说明

[0017] 图1为本发明的实现流程图；

[0018] 图2为现有vgg16网络模型结构图；

[0019] 图3为本发明使用位宽为16的定点网络测试cifar100得到的准确率；

[0020] 图4为本发明使用位宽为8的定点网络测试cifar100得到的准确率；

[0021] 图5为本发明使用位宽为16的定点网络测试imagenet得到的准确率；

[0022] 图6为本发明使用位宽为8的定点网络测试imagenet得到的准确率。

具体实施方式

[0023] 下面结合附图对本发明做进一步的描述。

[0024] 参照附图1，本发明的具体步骤如下。

[0025] 步骤1，获取预训练浮点模型。

[0026] 本发明从互联网下载两个已经预训练好的浮点格式的卷积神经网络模型，一个是由3层卷积层和2层全连接层组成的小型网络模型，另一个是由带13层卷积层和3层全连接

层组成的大型网络模型，其中：

[0027] 小型网络模型中每一个卷积层后面按顺序加了一层批量归一化层、激活层、池化层，每一个全连接层后面都按顺序加了一层批量归一化层、激活层、Dropout层，最后一层全连接层除外；该小型网络模型用于测试中规模的数据集cifar100，cifar100的测试集包括10000张测试图片，图片分为100类。

[0028] 该大型网络模型在vgg16模型的基础上在每层卷积层之后都加了一层批量归一化层；该大型网络模型用于测试大规模的数据集imagenet，imagenet是计算机视觉领域最大的数据库，本发明使用其中的分类数据集，分类数据集中验证集有50000张验证图片，图片分为1000类。

[0029] 所述常用的vgg16模型，如图2。该vgg16一共有13层卷积层和3层全连接层，13层卷积层分为5段，每段卷积之后紧接着最大池化层。

[0030] 步骤2，浮点模型定点化。

[0031] (2a)定义一个量化层，量化层对浮点输入数据进行量化，量化的定点数位宽用1+IL+FL表示，其中IL表示整数位宽，FL表示小数位宽；

[0032] (2b)利用编程语言python，在步骤1下载的小型预训练浮点模型的每一层批量归一化层后面调用(2a)定义的量化层；在步骤1下载的大型预训练浮点模型的每一层批量归一化层后面调用(2a)定义的量化层，由于全连接层后面没有批量归一化层，故直接在全连接层后面调用(2a)定义的量化层，最后一层全连接层不调用；

[0033] (2c)将(2b)中小型网络的每一层批量归一化层的输出作为(2a)定义的量化层的输入，将大型网络的每一层批量归一化层的输出和全连接层的输出作为(2a)定义的量化层的输入，量化层对这个浮点形式的输入进行量化，量化的公式为：

[0034]
$$Convert(x,\langle IL,FL\rangle)\begin{cases}-2^{IL-1} & \text{如果} x\le -2^{IL-1}\\ 2^{IL-1}-2^{-FL} & \text{如果} x\ge 2^{IL-1}-2^{-FL}\\ \dfrac{round(x\cdot 2^{FL})}{2^{FL}} & \text{其他}\end{cases}$$

[0035] 其中，Convert表示将浮点输入数据转化为定点输入数据，x为浮点输入数据，IL和FL分别表示定点输入数据的整数位宽和小数位宽，round为四舍五入函数，是编程语言的内置函数，2^{FL}表示量化成小数位宽为FL的定点数，-2^{IL-1}表示定点输入数据表示的数值范围的下限，$2^{IL-1}-2^{-FL}$表示定点输入数据表示的数值范围的上限；

[0036] (2d)对步骤1下载的预训练浮点模型中已经训练好的浮点权值进行量化，量化的定点数位宽用1+IL′+FL′表示，其中IL′表示整数位宽，FL′表示小数位宽，量化的公式为：

[0037]
$$Convert(w,\langle IL',FL'\rangle)\begin{cases}-2^{IL'-1} & \text{如果} w\le -2^{IL'-1}\\ 2^{IL'-1}-2^{-FL'} & \text{如果} w\ge 2^{IL'-1}-2^{-FL'}\\ \dfrac{round(w\cdot 2^{FL'})}{2^{FL'}} & \text{其他}\end{cases}$$

[0038] 其中，Convert表示将浮点权值转化为定点权值，w为浮点权值，IL′和FL′分别表示定点权值的整数位宽和小数位宽，round为四舍五入函数，是编程语言的内置函数，$2^{FL'}$表示

量化成小数位宽为FL′的定点数，$-2^{IL'-1}$表示定点权值表示的数值范围的下限，$2^{IL'-1}-2^{-FL'}$表示定点权值表示的数值范围的上限。

[0039] 本发明的效果可通过以下仿真实验做进一步说明。

[0040] 1.仿真条件：

[0041] 本发明的仿真实验是在基于python的深度学习框架pytorch下进行的。

[0042] 2.仿真内容：

[0043] 本发明通过使用定点量化后的小型网络模型对图像数据集cifar100进行测试，使用定点量化后的大型网络模型对图像数据集imagenet进行测试，验证本发明提出的量化方法的效果。

[0044] 仿真1，使用定点量化后的小型网络模型测试图像数据集cifar100。

[0045] 将输入图像数据的范围从0到255映射到0到1，并通过设置均值和方差将数据归一化到-1到1；

[0046] 根据硬件处理器的位宽，通过设置IL和FL将定点输入数据位宽分别设置成16、8，通过设置IL′和FL′将定点权值位宽分别设置成16、8，使用量化后的定点网络进行仿真测试，测试结果如图3和图4，其中：

[0047] 图3是定点网络位宽为16时测试得到的准确率，图4是定点网络位宽为8时测试得到的准确率，该图3和图4是一个三维图，图中x维代表定点输入数据的整数位宽，y维代表定点权值的整数位宽，z维代表测试准确率。

[0048] 从图3和图4可以看出：当定点输入数据整数位宽和小数位宽分别为4和11，定点权值整数位宽和小数位宽分别为2和13或3和12时，位宽为16的定点网络测试得到的准确率最高，为56.43％，比浮点型网络测试得到的准确率56.41％还要高。当定点输入数据整数位宽和小数位宽分别为4和3，定点权值整数位宽和小数位宽分别为3和4时，位宽为8的定点网络测试得到的准确率最高，为56.26％，比预训练的浮点型网络测试得到的准确率56.41％只低0.15％。

[0049] 仿真2，使用定点量化后的大型网络模型测试图像数据集imagenet。

[0050] 将输入图像数据的尺寸调整到256×256，再在图片的中间区域进行裁剪，将图片裁剪成224×224的尺寸，最后将裁剪后的图像数据的范围从0到255映射到0到1，并通过设置均值和方差将数据进行归一化处理。

[0051] 根据硬件处理器的位宽，通过设置IL和FL将定点输入数据位宽分别设置成16、8，通过设置IL′和FL′将定点权值位宽分别设置成16、8，使用量化后的定点网络进行仿真测试，测试结果如图5和图6，其中：

[0052] 图5是定点网络位宽为16时测试得到的准确率，图6是定点网络位宽为8时测试得到的准确率，图中x维代表定点输入数据的整数位宽，y维代表定点权值的整数位宽，z维代表测试准确率。

[0053] 从图5和图6可以看出：当定点输入数据整数位宽和小数位宽分别为6和9，定点权值整数位宽和小数位宽分别为2和13时，位宽为16的定点网络测试得到的准确率最高，为73.496％，比预训练的浮点型网络测试得到的准确率73.476％还要高。当定点输入数据整数位宽和小数位宽分别为2和5，定点权值整数位宽和小数位宽分别为0和7时，位宽为8的定点网络测试得到的准确率最高，为71.968％，只比浮点型网络测试得到的准确率73.476％

下降约1.5%。

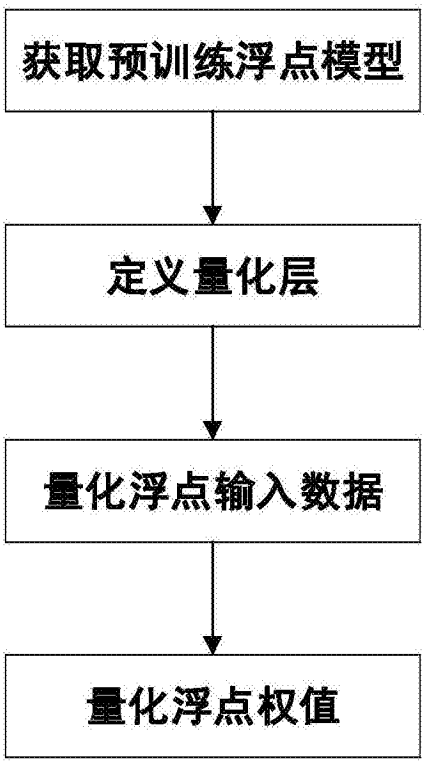

获取预训练浮点模型
定义量化层
量化浮点输入数据
量化浮点权值

图1

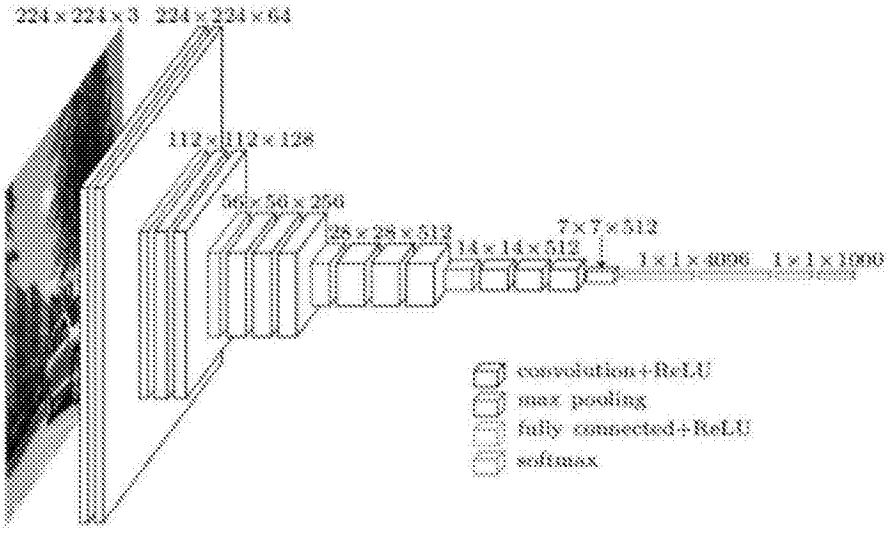

224×224×3
224×224×64
112×112×128
56×56×256
28×28×512
14×14×512
7×7×512
1×1×4096
1×1×1000
convolution+ReLU
max pooling
fully connected+ReLU
softmax

图2

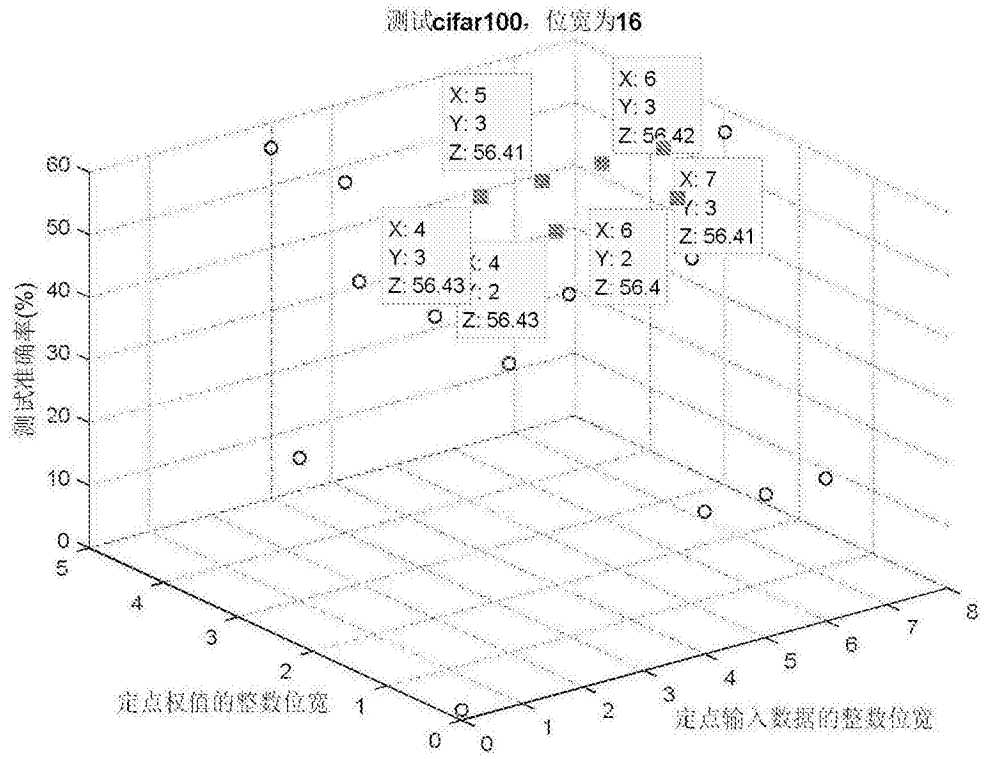

测试cifar100，位宽为16
X: 5
Y: 3
Z: 56.41
X: 6
Y: 3
Z: 56.42
X: 7
Y: 3
Z: 56.41
X: 4
Y: 3
Z: 56.43
X: 4
Y: 2
Z: 56.43
X: 6
Y: 2
Z: 56.4
测试准确率(%)
定点权值的整数位宽
定点输入数据的整数位宽

图3

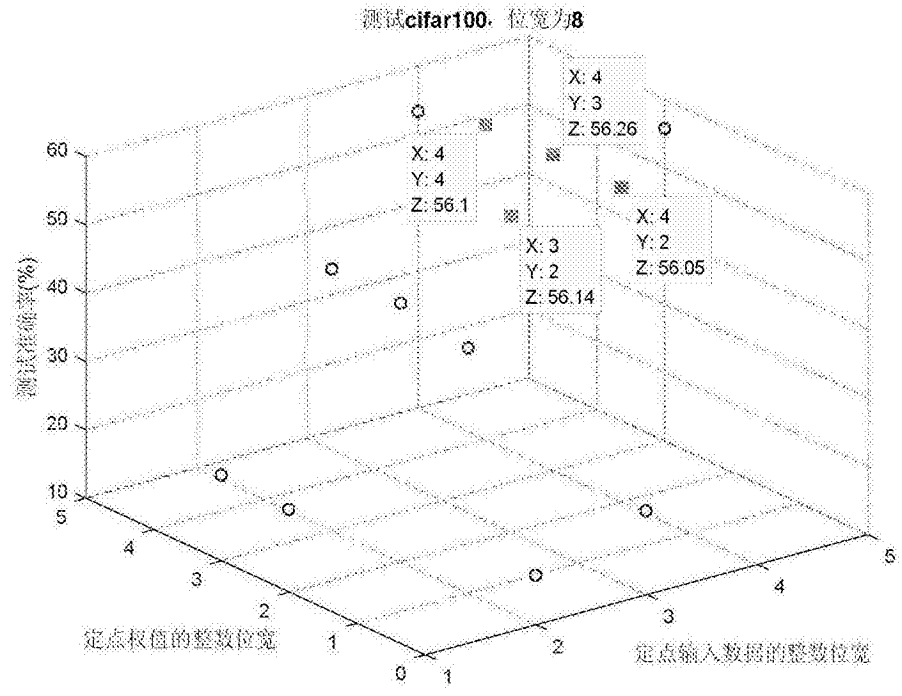

测试cifar100，位宽为8
X: 4
Y: 3
Z: 56.26
X: 4
Y: 4
Z: 56.1
X: 4
Y: 2
Z: 56.05
X: 3
Y: 2
Z: 56.14
测试准确率(%)
定点权值的整数位宽
定点输入数据的整数位宽

图4

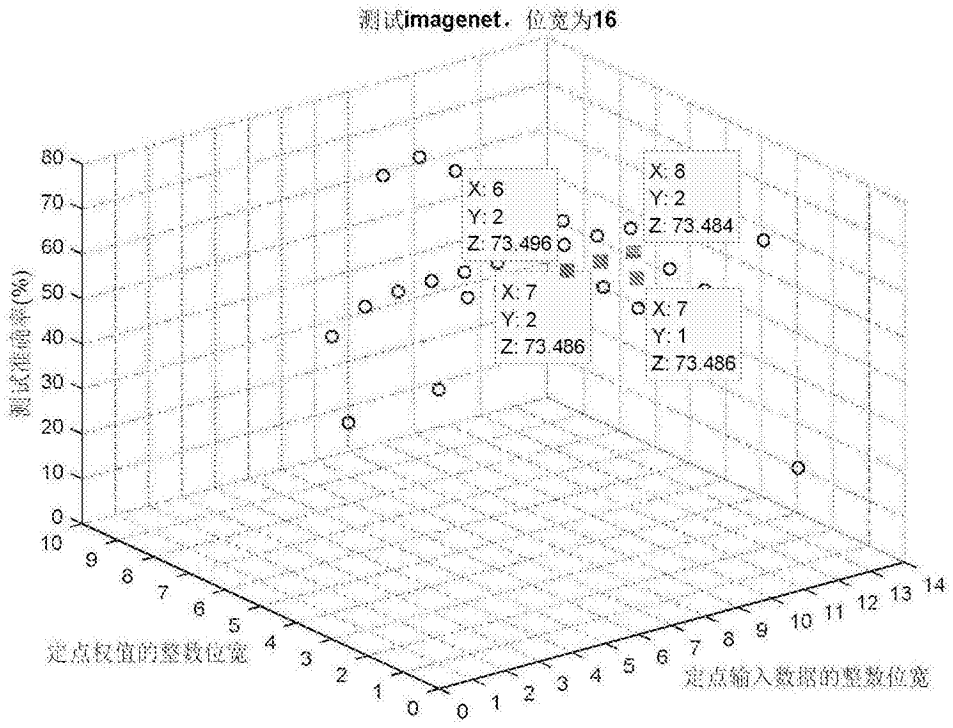

测试imagenet，位宽为16
测试准确率(%)
定点权值的整数位宽
定点输入数据的整数位宽
X: 6
Y: 2
Z: 73.496
X: 8
Y: 2
Z: 73.484
X: 7
Y: 2
Z: 73.486
X: 7
Y: 1
Z: 73.486

图5

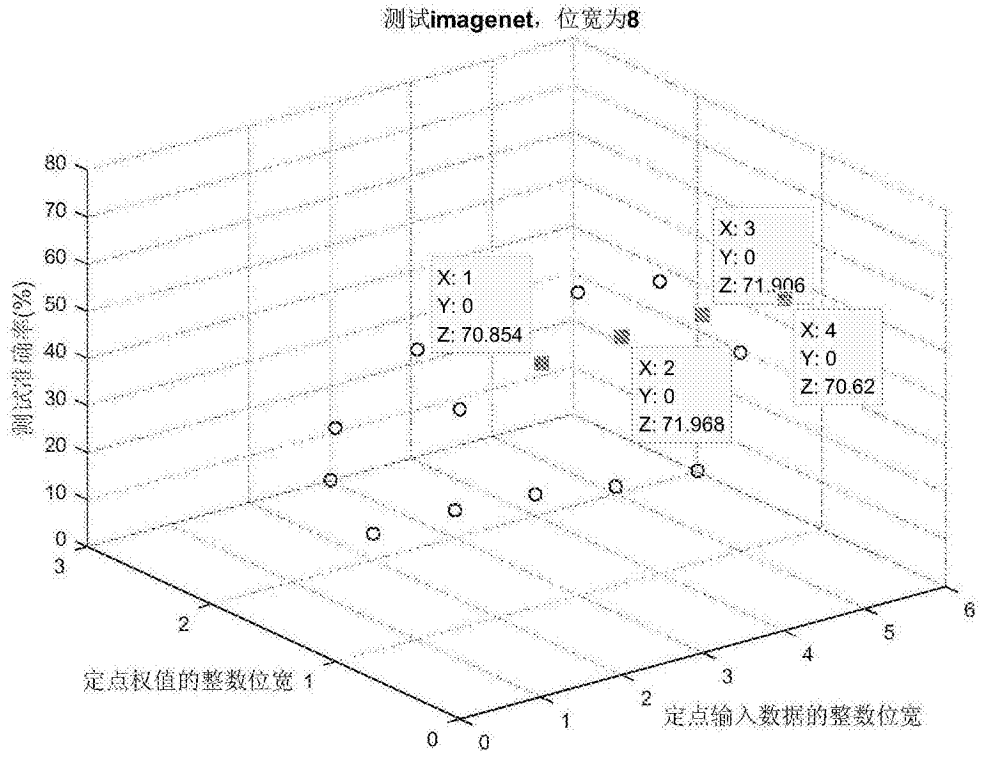

测试imagenet，位宽为8
测试准确率(%)
定点权值的整数位宽 1
定点输入数据的整数位宽
X: 1
Y: 0
Z: 70.854
X: 3
Y: 0
Z: 71.906
X: 2
Y: 0
Z: 71.968
X: 4
Y: 0
Z: 70.62

图6