(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 107423815 A
(43)申请公布日 2017.12.01

(21)申请号 201710665692.5

(22)申请日 2017.08.07

(71)申请人 北京工业大学
地址 100124 北京市朝阳区平乐园100号

(72)发明人 李玉鑑 余华擎

(74)专利代理机构 北京思海天达知识产权代理有限公司 11203
代理人 沈波

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

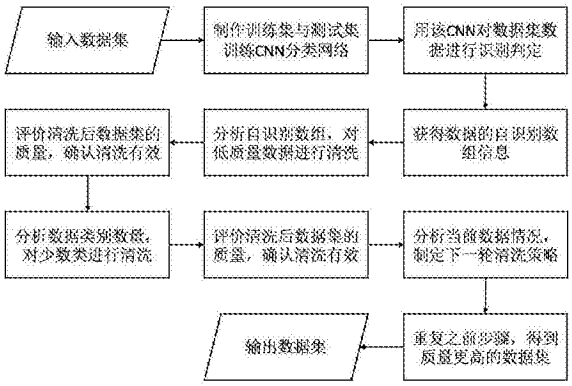
权利要求书2页 说明书5页 附图4页

(54)发明名称

一种基于计算机的低质量分类图像数据清洗方法

(57)摘要

本发明公开了一种基于计算机的低质量分类图像数据清洗方法，可以对从互联网批量收集的低质量分类图像数据进行有效清洗，从而获得更高质量的图像数据，用来训练一个识别率更高的分类模型。具体过程包括：先直接利用低质量分类图像数据训练一个初步的卷积神经网络，再用该网络对数据本身进行识别，清洗掉模型识别为本类的伪概率低到一定程度的图像或数量少于一定程度的图像类别，重复上述过程直到获得所有图像数据类型的识别率达到预设的标准。对比实验说明本发明能够有效提升图像数据的分类质量和识别水平。



CN 107423815 A

1.一种基于计算机的低质量分类图像数据清洗方法，其特征在于：本方法包括以下步骤，a)从互联网批量下载带有标签的图像数据，整理得到共计M类的图像数据集DataSet0，其中第i类包含的图像数目为N_i，i＝1,2,3…M；

b)用DataSet0训练一个卷积神经网络CNN0，具体步骤如下：

i.构建一个卷积神经网络模型，并固定这个网络模型的结构保持不变；

ii.对DataSet0随机取一定比例作为卷积神经网络的训练集；

iii.将DataSet0中非训练集的部分作为卷积神经网络的测试集；

iv.训练CNN0，迭代到指定次数后把网络测试识别率记为Acc0；

c)在DataSet0中，对第i类图像构造长度为N_i的一维图像自识别数组K_i，具体步骤如下：

i.用CNN0对DataSet0的图像数据进行识别，把其中第i类第j张图像识别为第k类的伪概率记为p_{ijk}，k＝1,2,3…M，并将这些伪概率从大到小排序；

ii.若排序后的前L个伪概率中存在k＝i，则记自识别率$K_{ij}=p_{ijk}$，否则记$K_{ij}=0$；

d)分析自识别数组K，清洗第i类图像数据里的低质量部分：

i.计算第i类图像自识别率的平均值：$\mu=\frac{\sum_{j=1}^{N_i}K_{ij}}{N_i}$

ii.计算第i类图像自识别率的标准差：$\sigma=\sqrt{\frac{1}{N_i}\sum_{j=1}^{N_i}(K_{ij}-\mu)^2}$

iii.计算第i类图像“低识别率”的分界值SepVal＝μ-σ*α，1≤α≤10且为整数，且SepVal>0；

iv.第i类图像中，若有K_{ij}<SepVal，则清洗掉第j张图像；清洗完成后得到数据集DataSet1；

e)使用DataSet1再次进行同样方式的卷积神经网络训练，得到网络测试识别率Acc1，记录并与Acc0比较并确认清洗是否有效；

f)在DataSet1中，重新对第i类图像数量进行统计，记每类图像数量为N'_i，对N'_i进行分析并清洗少数类别，以减少低质量数据类对卷积神经网络的影响：

i.计算当前M类别图像数量的平均值：$\mu=\frac{\sum_{i=1}^{M}N'_i}{M}$

ii.计算当前M类别图像数量的标准差：$\sigma=\sqrt{\frac{1}{M}\sum_{i=1}^{M}(N'_i-\mu)^2}$

iii.计算“少数类”图像数量的分界值SepVal＝μ-σ*α，1≤α≤10且为整数，且SepVal>0；

iv.统计M类图像中类别数量低于SepVal的类别共m类；

v.记m类数量总和为sum，M类数量总和为SUM；

vi.若m/M远大于sum/SUM，则判定该m类为少数类，需要清洗处理掉；

若m/M与sum/SUM数值接近，则认为m类数量正常，无需清洗处理；

g)用清洗过后的数据集DataSet2再次进行同样方式的卷积神经网络训练，得到网络测

试识别率Acc2，记录并与Acc1比较并确认清洗是否有效；

h)根据所得数据集情况，重复步骤(d)和(f)，得到清洗后的数据类别共m′类，m′<M；i)对于清洗后剩余的m′类共sum′张图像数据的质量进行评价：

i.获得该m′类在DataSet0里的所有数据，记总数量为SUM′，SUM′>sum′；

ii.对总量为SUM′和sum′的m′类图像数据进行同样方式的卷积神经网络训练，得到网络测试识别率Acc(SUM′)和Acc(sum′)，若Acc(SUM′)<Acc(sum′)，则说明清洗后的数据更有利于卷积神经网络的分类训练；

iii.从总量为sum′的m′类数据中随机或手动抽取一定的数据test作为公共测试集，以SUM′和sum′中除去test部分的数据作为训练集，进行同样方式的卷积神经网络训练，得到网络测试识别率为Acc(SUM′)和Acc(sum′)；若Acc(SUM′)<Acc(sum′)，则说明对于同样的测试集，使用经过清洗的数据作为训练集训练所得到的卷积神经网络泛化能力更强，测试识别率更高，即数据质量更高。

一种基于计算机的低质量分类图像数据清洗方法

技术领域

[0001] 一种基于卷积神经网络的低质量分类图像数据的清洗方法，该方法可以对从互联网批量收集的低质量分类图像数据进行有效清洗，从而获得更高质量的图像数据，用来训练一个识别率更高的分类模型，属于人工神经网络技术领域。

背景技术

[0002] 人工神经网络(Artificial Neural Network，即ANN)，是20世纪80年代以来人工智能领域兴起的研究热点。它从信息处理角度对人脑神经元网络进行抽象，建立某种简单模型，按不同的连接方式组成不同的网络。在工程与学术界也常直接简称为神经网络或类神经网络。神经网络是一种运算模型，由大量的节点(或称神经元)之间相互联接构成。每个节点代表一种特定的输出函数，称为激励函数(activation function)。每两个节点间的连接都代表一个对于通过该连接信号的加权值，称之为权重，这相当于人工神经网络的记忆。网络的输出则依网络的连接方式，权重值和激励函数的不同而不同。而网络自身通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达。

[0003] 卷积神经网络(Convolutional Neural Network，CNN)是一种前馈神经网络，也是人工神经网络的一种，它的人工神经元可以响应一部分覆盖范围内的周围单元，对于大型图像处理有出色表现。由于其独特的网络结构可以有效地降低反馈神经网络的复杂性，目前卷积神经网络已成为当前语音分析和图像识别领域的研究热点。它的权值共享网络结构使之更类似于生物神经网络，降低了网络模型的复杂度，减少了权值的数量。该优点在网络的输入是多维图像时表现的更为明显，使图像可以直接作为网络的输入，避免了传统识别算法中复杂的特征提取和数据重建过程。卷积神经网络相比于传统的神经网络，其特点如下：

[0004] 1.稀疏连接(Sparse Connectivity)

[0005] 卷积网络通过在相邻两层之间强制使用局部连接模式来利用图像的空间局部特性，在第m层的隐层单元只与第m-1层的输入单元的局部区域有连接，第m-1层的这些局部区域被称为空间连续的接受域。

[0006] 2.权值共享(Shared Weights)

[0007] 在卷积神经网络中，每个稀疏过滤器通过共享权值都会覆盖整个可视域，这些共享权值的单元构成一个特征映射，再加上与稀疏连接的配合，构成了特征提取层——卷积层。

[0008] 3.池化层(Pooling Layer)

[0009] 池化层是卷积神经网络的另一个构建块，它的功能是通过逐步减小表征的空间尺寸来减小参数量和网络中的计算。池化层在每一个特征图上独立操作。

[0010] 除此之外，卷积神经网络也包含有传统的神经网络的元素，如全连接层以及常见的非线性激活函数sigmoid、tanh、ReLU等。

[0011] 在卷积神经网络取得巨大成功的今天，一个良好的数据集是训练好的卷积神经网

络模型的关键所在。常见的数据集有PASCAL VOC、MNIST、ImageNet、CIFAR-10等，其中ImageNet有22K种共15M张高分辨率带标签图像，图像全被收集于网络，人工标记，常被用于卷积神经网络模型的分类性能检测。

[0012] 以上所述数据是通用而专业的，经过了大量的检验和人工标记。然而对于普通的应用级别的数据，所能获取到的某类图像可能来源于互联网爬虫，这其中必然夹杂着很多噪声，如何从其中清洗出质量较高的数据并给与一定的评价方式，是本发明的重点所在。从噪声数据中获取质量较高的数据之后，则可以将其用于卷积神经网络的训练，从而达到一定的实际应用的目的。

发明内容

[0013] 本发明采用的技术方案为一种基于卷积神经网络的低质量分类图像数据的清洗方法，包

[0014] 括以下步骤：

[0015] a)从互联网批量下载带有标签的图像数据，整理得到共计M类的图像数据集DataSet0，其中第i类包含的图像数目为N_i，i＝1,2,3…M；

[0016] b)用DataSet0训练一个卷积神经网络CNN0，具体步骤如下：

[0017] i.构建一个卷积神经网络模型，并固定这个网络模型的结构保持不变；

[0018] ii.对DataSet0随机取一定比例(如80％、90％)作为卷积神经网络的训练集；

[0019] iii.将DataSet0中非训练集的部分作为卷积神经网络的测试集；

[0020] iv.训练CNN0，迭代到指定次数后把网络测试识别率记为Acc0；

[0021] c)在DataSet0中，对第i类图像构造长度为N_i的一维图像自识别数组K_i，具体步骤如下：

[0022] i.用CNN0对DataSet0的图像数据进行识别，把其中第i类第j张图像识别为第k类的伪概率记为p_{ijk}，k＝1,2,3…M，并将这些伪概率从大到小排序；

[0023] ii.若排序后的前L个(如L＝10)伪概率中存在k＝i，则记自识别率$K_{ij}=p_{ijk}$，否则记$K_{ij}=0$；

[0024] d)分析自识别数组K，清洗第i类图像数据里的低质量部分：

[0025] i.计算第i类图像自识别率的平均值：$$\mu=\frac{\sum_{j=1}^{N_i}K_{ij}}{N_i}$$

[0026] ii.计算第i类图像自识别率的标准差：$$\sigma=\sqrt{\frac{1}{N_i}\sum_{j=1}^{N_i}(K_{ij}-\mu)^2}$$

[0027] iii.计算第i类图像“低识别率”的分界值SepVal＝μ-σ*α，1≤α≤10且为整数，且SepVal>0；

[0028] iv.第i类图像中，若有K_{ij}<SepVal，则清洗掉第j张图像；清洗完成后得到数据集DataSet1；

[0029] e)使用DataSet1再次进行同样方式的卷积神经网络训练，得到网络测试识别率Acc1，记录并与Acc0比较并确认清洗是否有效；

[0030] f)在DataSet1中,重新对第i类图像数量进行统计,记每类图像数量为N'i,对N'i进行分析并清洗少数类别,以减少低质量数据类对卷积神经网络的影响:

[0031] i.计算当前M类别图像数量的平均值:$\mu=\frac{\sum_{i=1}^{M}N_i'}{M}$

[0032] ii.计算当前M类别图像数量的标准差:$\sigma=\sqrt{\frac{1}{M}\sum_{i=1}^{M}(N_i'-\mu)^2}$

[0033] iii.计算"少数类"图像数量的分界值SepVal=μ-σ*α,1≤α≤10且为整数,且SepVal>0;

[0034] iv.统计M类图像中类别数量低于SepVal的类别共m类;

[0035] v.记m类数量总和为sum,M类数量总和为SUM;

[0036] vi.若m/M远大于sum/SUM,则判定该m类为少数类,需要清洗处理掉;

[0037] 若m/M与sum/SUM数值接近,则认为m类数量正常,无需清洗处理。

[0038] g)用清洗过后的数据集DataSet2再次进行同样方式的卷积神经网络训练,得到网络测试识别率Acc2,记录并与Acc1比较并确认清洗是否有效;

[0039] h)根据所得数据集情况,重复步骤(d)和(f),得到清洗后的数据类别共m'类,m'<M;

[0040] i)对于清洗后剩余的m'类共sum'张图像数据的质量进行评价:

[0041] i.获得该m'类在DataSet0里的所有数据,记总数量为SUM',SUM'>sum';

[0042] ii.对总量为SUM'和sum'的m'类图像数据进行同样方式的卷积神经网络训练,得到网络测试识别率Acc(SUM')和Acc(sum'),若Acc(SUM')<Acc(sum'),则说明清洗后的数据更有利于卷积神经网络的分类训练;

[0043] iii.从总量为sum'的m'类数据中随机或手动抽取一定的数据test作为公共测试集,以SUM'和sum'中除去test部分的数据作为训练集,进行同样方式的卷积神经网络训练,得到网络测试识别率为Acc(SUM')和Acc(sum');若Acc(SUM')<Acc(sum'),则说明对于同样的测试集,使用经过清洗的数据作为训练集训练所得到的卷积神经网络泛化能力更强,测试识别率更高,即数据质量更高。

附图说明

[0044] 图1是实验整体思路流程图。

[0045] 图2是初始数据集情况及其卷积神经网络测试识别率结果图。

[0046] 图3是当前数据集自识别数组示意图。

[0047] 图4是清洗掉低质量图像之后的数据集情况及其卷积神经网络测试识别率结果图。

[0048] 图5是第一次对数据进行少数类的分析以及类别清洗情况结果图。

[0049] 图6是一次清洗少数类别之后的数据集情况及其卷积神经网络测试识别率结果图。

[0050] 图7是第二次对数据进行少数类的分析以及类别清洗情况结果图。

[0051] 图8是二次清洗少数类别之后的数据集情况及其卷积神经网络测试识别率结果

图。

[0052] 图9是分别对清洗前后的类进行数据质量评价的对比结果图。

[0053] 图10是使用同一个测试集，数据清洗前后作为训练集，训练的卷积神经网络比较结果图。

具体实施方式

[0054] 下面结合附图及具体实施案例对本发明作进一步的描述：

[0055] 1.从互联网上批量下载植物花卉图像数据，整理可得775个分类共计161015张图，其中第i类包含的图像数目为N_i(i＝1,2,3…M)；

[0056] 2.用所得图像数据集训练一个卷积神经网络，具体步骤如下：

[0057] a)获取python caffe上AlexNet的网络模型文件，并获得其在ImageNet上的预训练模型文件，用于卷积神经网络的初始化；

[0058] b)在图像数据集中随机取约90％数据共144921张图作为训练集，剩下10％的16067张图作为测试集，使用caffe进行卷积神经网络训练，迭代10000次后，得到网络测试识别率为39％；

[0059] c)对数据集中第i类图像构造长度为N_i的一维图像自识别数组K_i(i＝1,2,3…M)，具

[0060] 体步骤如下：

[0061] i.用该训练好的卷积神经网络对初始图像数据集一一进行识别；

[0062] ii.对第i类的第j张图的伪概率识别结果做以下处理：

[0063] 1)若卷积神经网络返回的前10个伪概率识别结果里没有第i类，则记K_{ij}＝0；

[0064] 2)若返回的前10个伪概率识别结果里有第i类且概率为p，则记K_{ij}＝p；

[0065] d)分析自识别数组，对i类数据进行低质量图像数据清洗：

[0066] i.计算第i类图像自识别率的平均值：$\mu=\frac{\sum_{j=1}^{N_i}K_{ij}}{N_i}$；

[0067] ii.计算第i类图像自识别率的标准差：$\sigma=\sqrt{\frac{1}{N_i}\sum_{j=1}^{N_i}(K_{ij}-\mu)^2}$；

[0068] iii.计算“少数类”图像数量的分界值SepVal＝μ-σ*α(本实验取α＝1)；

[0069] iv.将第i类图像中，识别率低于SepVal的图像作为低质量图像删去。

[0070] e)清洗过后剩下79198张图，取约90％的71298作为训练集，剩下的7900张作为测试集，再次用同样的方法进行卷积神经网络训练，得到测试识别率为59.7％，相比初始的39％有不小提高；

[0071] f)对清理后这775类图像的数量重新统计，记为N'_i，对N'_i行分析并清洗少数类别，以减少低质量数据类对分类网络的影响：

[0072] i.计算当前M＝775类别图像数量的平均值：$\mu=\frac{\sum_{i=1}^{M}N'_i}{M}$

[0073] ii.计算当前M＝775类别图像数量的标准差：$\sigma=\sqrt{\frac{1}{M}\sum_{i=1}^{M}(N_i-\mu)^2}$

[0074] iii.计算“少数类”图像数量的分界值SepVal＝μ-σ*α(本实验取α＝1)；

[0075] iv.在低质量噪声图像清洗后，统计低于SepVal张图像的类别数共178个类，共1815张图；因为178/775远大于1815/79198，故判断这些类为少数低质量数据类，将其清洗掉；

[0076] g)清洗过后的77383张图像数据取70000张作为训练集，剩下7383张作为测试集，再次用同样的方法进行卷积神经网络训练，得到测试识别率为60.0％，比之前网络的测试识别率稍有提升；

[0077] h)根据所得数据集情况，再次清洗少数类，步骤同上，得到清洗后的数据类别共468类；

[0078] i)对于清洗后剩余的468个类共70755张图像数据的性能评价如下：

[0079] i.获得该468个类在原始数据里的所有数据，总数量为111290张；

[0080] ii.对总量为111290和70755的468个类数据各自进行同样方法的卷积神经网络训练，得到网络测试识别率为60.8％和62.6％，说明清洗后的数据更有利于卷积神经网络的分类训练；

[0081] iii.从总量为70755的468类数据中随机抽取10％作为公共的测试集test，以111290和70755中除去test的数据作为训练集，用同样的方法训练卷积神经网络，得到的网络测试识别率分别为59.6％和62.6％。这说明，对于同样的测试集，经过清洗的数据网络泛化能力更强，平均准确率更高，即数据性能更好。

[0082] 通过实验结果可以看出：

[0083] 1.本次数据清洗效果真实有效，按同样的评价方法较原始数据质量上有所提升。

[0084] 2.可根据当前数据集情况来决定下一步清洗策略，该方法较为灵活。

[0085] 3.在测试集相同的情况下，清洗后的数据得到卷积神经网络的识别率更高，说明清洗后数据质量有所提高。

[0086] 以上示例仅用以说明本发明，而并非限制本发明所描述的技术方案。因此，一切不脱离本发明的精神和范围的技术方案及其改进，均应涵盖在本发明的权利要求范围当中。

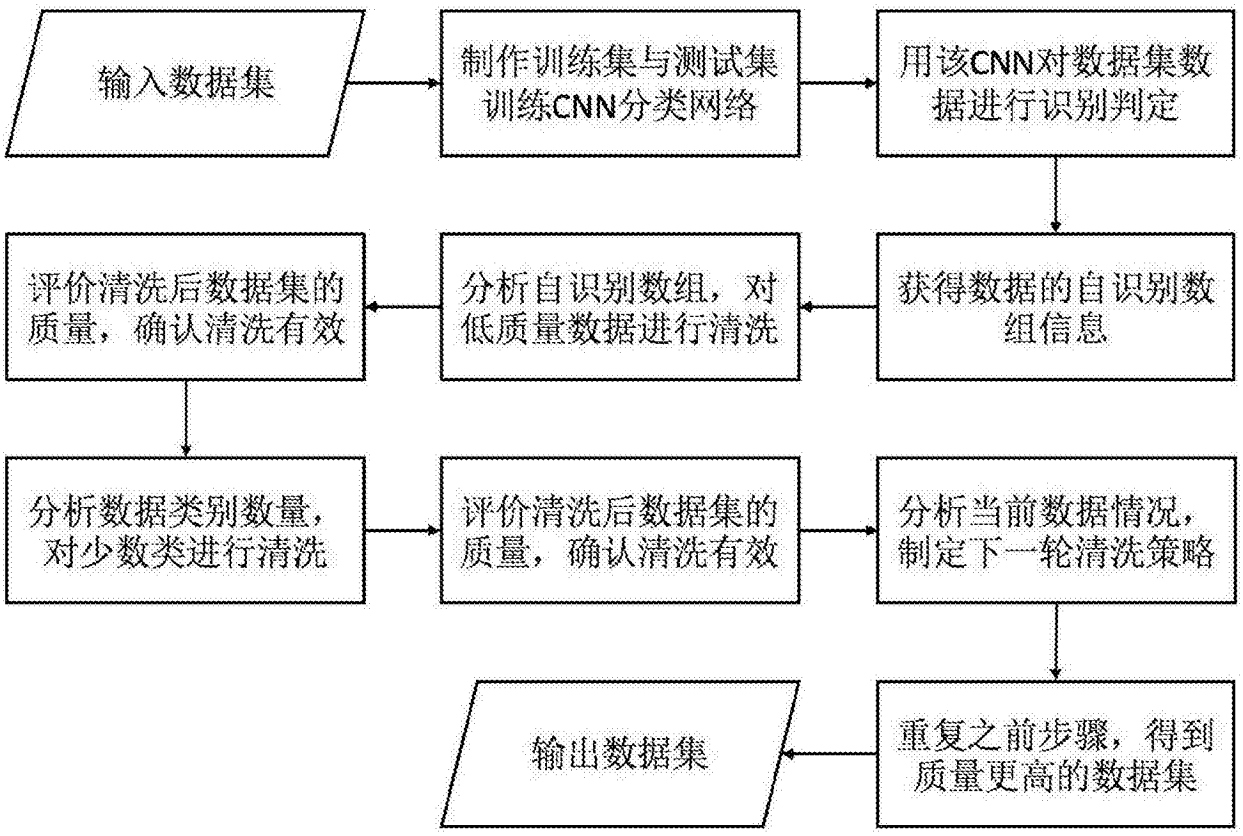


图1

- 初始图像数据: 161015张图, 775类
- 训练集: 144921张图
- 测试集: 16067张图
- 结果: 测试识别率均值为39% (比较低)

图2

…	(图像数)	1(数值)	2(数值)	3(数值)	4(数值)
类别 66	198	0.25	0.37	0.02	0.4
类别 67	233	0.0	0.12	0.5	0.0
…					

第67类图的第2张图像被模型识别为第67类的概率P= 0.37

图3

- 图像数据: 775类共计79198张，类均数量102.2
- 训练集: 71298张图
- 测试集: 7900张图
- 结果: 测试识别率均值为59.7% (显著提高)

图4

A	B	C	D	E
	图像总数	类别数	图像占比	类别占比
少数类	1815	178	2.29%	22.96%
多数类	77383	597	97.71%	77.04%

图5

- 图像数据: 597类共计77383张，类均数量129.6
- 训练集: 70000张图
- 测试集: 7383张图
- 结果: 测试识别率均值为60.0% (稍有提高)

图6

	图像总数	类别数	图像占比	类别占比
少数类	6628	129	8.56%	21.60%
多数类	70755	468	91.44%	78.40%

图7

- 图像数据: 468类共计70755张，类均数量151.2
- 训练集: 63680张图
- 测试集: 7075张图
- 结果: 测试识别率均值为62.6% (比597类高2%左右)

图8

初始数据集

- 总量: 111290张图 类均数量237.8张图
- 结果: 测试识别率均值为60.8%

清洗后当前数据集

- 总量: 70755张图 类均数量151.2张图
- 结果: 测试识别率均值为62.6%(比上述高约2%)

图9

- **公共测试集: 70755图像中的随机10%**

初始数据集

- 总量: 104215张图 类均数量222.7张图
- 结果: 测试识别率均值为59.6%

清洗后当前数据集

- 总量: 63680张图 类均数量136.1张图
- 结果: 测试识别率均值为62.6%(比上述高约3%)

图10