基于蒙特卡罗数据集均衡与鲁棒性增量极限学习机的图像自动标注

柯 逍[1,2],邹嘉伟[1,2],杜明智[1,2],周铭柯[1,2]
(1. 福州大学数学与计算机科学学院,福建福州 350116;
2. 福建省网络计算与智能信息处理重点实验室(福州大学),福建福州 350116)

摘 要: 针对传统图像标注模型存在着训练时间长、对低频词汇敏感等问题,该文提出了基于蒙特卡罗数据集均衡和鲁棒性增量极限学习机的图像自动标注模型.该模型首先对公共图像库的训练集数据进行图像自动分割,选择分割后相应的种子标注词,并通过提出的基于综合距离的图像特征匹配算法进行自动匹配以形成不同类别的训练集.针对公共数据库中不同标注词的数据规模相差较大,提出了蒙特卡罗数据集均衡算法使得各个标注词间的数据规模大体一致.然后针对单一特征描述存在的不足,提出了多尺度特征融合算法对不同标注词图像进行有效的特征提取.最后针对传统极限学习机存在的隐层节点随机性和输入向量权重一致性的问题,提出了鲁棒性增量极限学习,提高了判别模型的准确性.通过在公共数据集上的实验结果表明:该模型可以在很短时间内实现图像的自动标注,对低频词汇具有较强的鲁棒性,并且在平均召回率、平均准确率、综合值等多项指标上均高于现流行的大多数图像自动标注模型.

关键词: 蒙特卡罗数据集均衡;多尺度特征融合;极限学习机;图像自动标注
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2017)12-2925-11
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.12.014

The Automatic Image Annotation Based on Monte-Carlo Data Set Balance and Robustness Incremental Extreme Learning Machine

KE Xiao[1,2], ZOU Jia-wei [1,2], DU Ming-zhi [1,2], ZHOU Ming-ke [1,2]
(1. *College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian* 350116, *China*;
2. *Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, Fuzhou University, Fuzhou, Fujian* 350116, *China*)

Abstract: Aiming at the problem that the traditional image annotation model has long training time, sensitive to low-frequency words and other issues, this paper proposes a new automatic image annotation method based on Monte-Carlo data-set balance and robustness incremental extreme learning machine. First of all, training images of the public image library are segmented into different areas by this model and corresponding seed markup words are selected after segmentation, the areas are matched automatically based on comprehensive distance algorithm and the different keywords represent different areas. Then, for the huge difference of different annotated words' sizes in the public database, the Monte Carlo data set equalization algorithm is proposed to make the data size of each annotated word much the same. And a multi-scale feature fusion algorithm is proposed to extract effective features from different annotated images. Finally, the robustness incremental limit learning is proposed to improve the accuracy of the discriminant model for the problems of the consistency of the hidden layer nodes and the input vector weights existing in the traditional limit learning machine. The experimental results show that: compared with traditional algorithms of image automatic annotation, the methods proposed in this paper can implement the automatic image annotation quickly, and it is robust to low frequency words, and it is higher than most popular models of automatic image annotation in terms of average recall rate, average accuracy rate, comprehensive value and so on.

Key words: Monte-Carlo data set balance; multi-scale feature fusion; extreme learning machine; automatic image annotation

收稿日期:2016-03-14;修回日期:2017-03-27;责任编辑:梅志强
基金项目:国家自然科学基金(No.61502105);福建省科技引导性项目(No.2017H0015);福建省中青年教师教育科研项目(No.JA15075)

1 引言

面对如此海量的图像信息,如何高效的组织和管理这些图像,使得人们可以在海量图像信息中快速准确的获取所需信息,是当今世界一个十分棘手的问题.图像的自动标注技术因其有着广泛的应用场景,所以也成为近年来在模式识别领域的一个重要研究方向.自动图像标注需要对图像中的所有内容进行标注,即需要给出图像内容对应的所有标签[1].图像的自动标注是图像理解和图像检索的重要步骤.

近年来随着人工智能技术以及机器学习的快速发展,出现了多种图像自动标注模型.根据在建模方法上采用的不同思想,可以把图像自动标注模型大致分为两个类别.第一类图像自动标注模型是基于统计概率思想,第二类图像自动标注模型是基于分类思想.基于统计概率的方法主要是通过诸如图像的纹理、形状、颜色、边缘等底层视觉特征的提取,计算出其与某类或者某些关键词之间的相关性或联合概率,最后将概率最大或者关联程度最高的一个或者多个关键词作为该图的标注词.代表性的方法有:共生模型图像自动标注[2],其思想是通过计算标注词与栅格之间的共存关系来对每个图像栅格进行标注.共生模型具有实现架构简单,并且计算简单的优点.但是标注的准确率很低,并且分割方法较为粗糙.文献[3]从概率统计的角度出发并结合高斯图特征提取过程中的通用背景模型,提出了新的一种图像自动标注模型,该方法主要是通过引入受限的对称 Dirichlet 分布来描述高斯混合模型中权重参数的先验分布,然后利用贝叶斯最大后验概率对高斯混合模型参数进行估计,该方法在含有噪声的大规模数据集上得到了较好的标注效果.在基于分类思想的算法中,图像中的每一个类别或者说标签都是一个语义关键词,将图像的自动标注过程和图像的多分类做一个等价的转换.代表性的方法有:Corr-LDA(Correspondence-Latent Dirichlet Allocation)[4]对图像、类标签和标注进行联合建模,将类标签作为图像的全局描述,将标注词作为图像的局部描述,每个标注词会假设被与图像块相关联的主题所描述,首先生成模型主题对应的图像,然后由一个受生成图像主题约束的模型生成图像的标注词或类标签,这类方法的难点在于如何准确地划分主题类别. JEC(Joint Equal Contribution)[5]利用全局底层图像特征和基本距离的简单组合来寻找给定图像的最近邻,然后使用一种贪心的标签传递机制将关键词赋予给定图像,取得了很好的标注精度性能,该方法容易实现,但高维特征间两两计算视觉距离,带来了难以忽略的时间开销.文献[6]对 SML 模型存在的问题做了进一步的改进,提出了一种弱监督多标签图像标注系统.相比传统的 SML 方法性能上有了进一步的提升,但该模型较为复杂,需要消耗极大的计算机资源.

本文着重从机器学习的角度出发,重点对基于分类思想的图像自动标注模型进行了进一步研究,提出了一种基于蒙特卡罗数据集均衡和鲁棒性增量极限学习机的图像自动标注模型(MB_IELM).简单的说,该模型首先对分割后的图像进行本文所提出的基于综合距离的特征匹配,解决了传统方法上大量人工干预匹配所带来的不足,其次对匹配后的不同标注词图像集进行本文所提出的多尺度特征融合提取算法,完善了传统方法单一特征提取所带来的描述欠缺,再次通过本文提出的蒙特卡罗数据集均衡算法,很好的解决了公共图像库上的训练数据集规模不平衡而导致标注结果倾向于高频词而淹没低频词的问题.最后通过将提取的特征向量输入到本文提出的鲁棒性增量极限学习机中进行多标注词机器学习,弥补了传统多分类方法训练时间长,泛化能力差等问题.本文提出的图像自动标注模型,其标注效果明显优于现流行的大多数图像自动标注模型.

2 基于蒙特卡罗数据均衡和鲁棒性增量极限学习机的图像自动标注框架

该模型的整体框图如图 1 所示,具体的实现步骤如下.

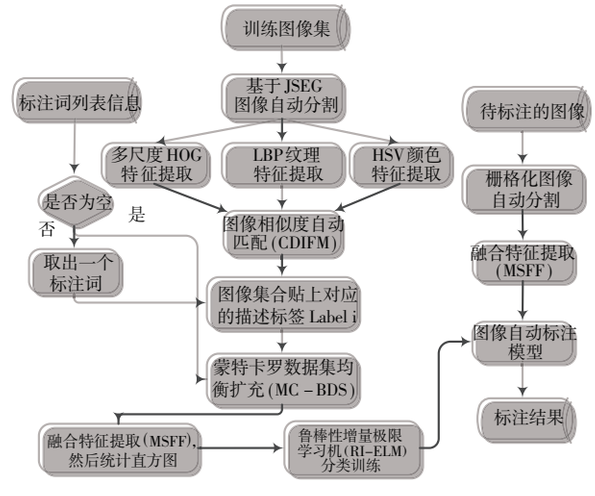


图1 本文图像自动标注模型整体框图

(1)通常来说,一张图像的场景是由多个标注词(关键词)来表述的,为使得各个标注词能够对符合其语义的图像区域进行描述,在选取合适的公共图像库中的训练集图像时,需要对训练集中的图像进行区域

分割,分割后具有不同特征信息的区域可以分别与不同的标注词相对应. 例如,从视觉和图像理解角度来说,“花”,“海”,“天空”应该对应不同的区域.

(2)选择对应于某个标注词的种子图像,例如可以选择标注词为“花”的一张或者多张区域图像作为种子图像,根据相应的图像匹配算法(CDIFM),将所有描述为“花”的图像串联在一起,放入一个集合中,并且将该集合的标签贴上花,实验中可以用阿拉伯数字 1 表示“花”,2 表示“天空”,3 表示“飞机”等等.

(3)通常来说,选择的训练集图像可能存在的数据不平衡问题,即有的标签集合中可能有上百张训练图像,而有的标签集合中可能只有两三张训练图像. 根据统计学思想,若是采用大规模相差巨大的不同数据集建模时,测试或者实验结果往往倾向于量大的数据集类别. 本文提出了蒙特卡罗数据集均衡算法(MC-BDS),使得各个标签之间的训练图像分布相对均衡.

(4)分别对各个标签类的图像进行多尺度特征融合的提取(MSFF),即结合多尺度下的 HOG 局部特征、LBP 纹理特征和 HSV 颜色作为该标签类的图像的描述符. 将提取后的特征输入到鲁棒性增量极限学习机(RI-ELM)中进行分类训练,求出相应的参数,最后得出该训练模型.

(5)利用 4.3 节的图像分割方法对待标注图像进行区域分割,分别对每个分割后的区域提取其多尺度融合特征. 并将该特征描述符输入到上述 RI-ELM 的图像自动标注模型之中,给待标注的图像进行标注.

3　多尺度特征融合与鲁棒性增量极限学习机

3.1　多尺度特征融合提取算法(MSFF)

图像的底层特征,例如颜色,形状,纹理等能够在一定程度上反映出图像所表达的内容,因此研究图像的底层特征是图像标注中关键的一步. 根据具体应用场景的不同,选用的底层特征也应有所不同. 本文提出多尺度特征提取算法包括以下几部分内容.

3.1.1　特征提取

方向梯度直方图(HOG)是一种在图像识别和机器视觉领域中得到广泛应用的特征描述算子[7]. 它表示的是边缘结构特征,可以很好的描述对象的局部形状信息. 它对图像的几何拉伸和光照变化有很强的鲁棒性[8].

局部二值模式(Local Binary Pattern,LBP)是用来表达图像局部纹理特征信息的一种常用算子. 在图像的旋转不变性和灰度不变性上表出现了较强的鲁棒性[9]. 本文采用了多级 LBP 特征提取.

最常用的颜色特征表达方法是颜色直方图,因其不受图像的尺度变化、背景色彩、噪声干扰、平移变换以及旋转等的影响得到广泛的应用. 本文根据人的颜色感知系统将 HSV 空间中的颜色分量进行非等间隔量化,通过计算颜色落在每个小区间内的像素数量可以得到 HSV 颜色直方图.

3.1.2　基于多尺度的特征描述

传统的视觉特征分析往往是基于单尺度的,很多情况下没法接近图像本质特征,并且很难进行推广应用. 针对该问题,本文提出了图像的多尺度特征描述. 其思想是:原始图像可以通过一个尺度变量在其空间领域进行一系列变换,在不同的尺度下得到图像的描述特征. 该尺度变量可以模拟人和观察物体在距离变化的情况下,物体特性在人体视网膜上的形成过程. 本文将一幅二维图像的尺度空间定义为:

$$L(x,y,\sigma)=G(x,y,\sigma)*I(x,y) \tag{1}$$

式中:$I(x,y)$表示的是原始图像,$G(x,y,\sigma)$表示的是尺度可变高斯函数, 数学定义为:

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)}/2\sigma^2 \tag{2}$$

式中(x,y)表示的是像素点的空间坐标,σ 表示的是相应的尺度坐标. 当 σ 比较大的时候则图像的平滑度会越高,这样对应的是图像的整体概貌特征. 当 σ 比较小的时候,则图像的平滑程度会比较低,与此相对应的是图像的更多细节特征.

从上述分析我们可以得知,传统的 HOG 特征描述更多的是对象的细节,轮廓边缘信息,一定程度上会忽略一些平坦的对象表面信息,结合空间信息概念,引入多尺度的方法之后在一定的程度上可以缓和这一不足. 此外,当对象的形状边缘存在较多的噪声时,会导致在识别或者分类问题上处理能力下降,我们通过结合 LBP 特征可以很大程度上克服这一劣势,LBP 特征可以有效的滤除边缘噪声带来的影响. 另外,HSV 颜色特征作为一种全局性的特征,对图像几何形态变化具有较强的抗干扰能力,可以弥补在全局信息上的损失. 全局信息用来描述对象的整体信息,而局部信息用来描述对象的细节,两者的结合在对象的识别或者是分类上是很有必要的,而且能达到良好的效果.

综上分析,本文提出基于上述三种特征相互融合的特征提取算法(MSFF),该特征可以描述为多尺度下的 HOG 特征和多级 LBP 特征以及 HSV 特征的结合,即 Multi-H-L-S 特征. 该特征的实现步骤如算法 1 所示.

3.2　鲁棒性增量极限学习机(RIELM)

极限学习机(ELM Extreme Learning Machine)是一种特殊的、有效的、基于单隐层前馈神经网络的学习算法[10]. 与传统的神经网络算法(如 BP 等算法)相比,极限学习机可以在不牺牲模型准确率的基础上,极大的降低因为网络参数的调节所消耗的时间,具有学习速

度快以及泛化能力强的优点.

和传统的神经网络相似,ELM 的输入值往往是样本的特征描述符,本文的前面部分也分析过,单一的特征描述、局部的特征描述或者全局的特征描述无法对目标进行有效的刻画. 因此,实际上往往需要融合多个特征来对目标进行刻画.

算法 1 多尺度特征融合提取算法(MSFF)

输入:分割后的图像区域集合 $\psi(\boldsymbol{Z})=\{\boldsymbol{Z}_1,\boldsymbol{Z}_2\cdots\boldsymbol{Z}_n\}$,其中 n 为图像集中的元素个数.

输出:$\{\varphi\{Z_1\},\varphi\{\boldsymbol{Z}_2\}\cdots\varphi\{Z_n\}\}$,其中 $\boldsymbol{\varphi}\{\boldsymbol{Z}_k\}(1\leqslant k\leqslant n\}$ 表示的是图像的多尺度融合特征向量.

初始化:

1. 选择合适的提取窗口(window),块(block)和单元(cell),大小分别为 $w_1\times w_2,b_1\times b_2,c_1\times c_2$.
2. 令窗口滑动步长(step)大小为 $s_1\times s_2$.
3. 令高斯尺度参数为 σ,多尺度下的图像组数(octave)和层数(interval)分别为 O 和 I.

开始:

(1)for $i=0:n-1$ //图像集中的元素个数

(2) for $j=0:O$//多尺度下图像组数

(3) for $k=0:I$//多尺度下图像层数

(4) for $s=0:h$//多尺度下的窗口个数

//其中 $h=(((w_1-b_1)/s_1)+1)\times(((w_2-b_2)/s_2)+1)\times(b_1\times b_2)/(c_1\times c_2)$

(5) compute $\varphi(\boldsymbol{Z}_{i\text{-}j\text{-}k\text{-}s\text{-}\mathrm{hog}}),s=s+1$

//计算当前尺度下当前窗口内图像局部特征向量

(6) if($s\leqslant h$) goto(5), else {goto(3), $k=k+1$。}

(7) repeat(1) ~ (6),compute $\varphi(\boldsymbol{Z}_{i-\mathrm{multi\text{-}hog}})$.

//计算该图像多尺度下的 HOG 特征向量

(8) if($j\leqslant O$) goto(3),else goto(9).

//计算该图像的多级 LBP 特征向量

(9) for $p=0:h'$

//其中满足 $h'=(w_1\times w_2)/(b_1\times b_2)$,和 HOG 特征提取方式的不同在于:此时的各个 block 之间没有重叠.

(10) compute $\varphi(Z_{i-\mathrm{lbp}}),p=p+1$.

//计算该图像的 LBP 纹理特征.

(11) if($p\leqslant h'$) goto(10), else goto(12)

(12) compute $\varphi(\boldsymbol{Z}_{i-\mathrm{hsv}}),i=i+1$ //计算该图像基于全局信息的 HSV 颜色特征.

(13) if($i\leqslant n-1$) repeat (2) ~ (12), else end for.

//计算出图像的多尺度特征向量为

$$\varphi(\boldsymbol{Z}_k\}=\{\varphi(\boldsymbol{Z}_{i-\mathrm{multi\text{-}hog}}),\varphi(\boldsymbol{Z}_{\mathrm{lbp}}),\varphi(\boldsymbol{Z}_{\mathrm{hsv}})\}(1\leqslant k\leqslant n\}$$

(14) output $\{\varphi(\boldsymbol{Z}_1\},\varphi\{\boldsymbol{Z}_2\}\cdots\varphi\{\boldsymbol{Z}_n\}\}$,end for.

此时便会涉及到一个问题,以本文提出的多尺度特征融合为例子,其中 HOG 特征是采用多尺度来描述,而 LBP 特征采用的是基于滑动窗口的多级描述,HSV 因为是基于全局的,所以是以整个目标的颜色直方图为基础. 那么可能会存在着这么一种隐患:维数规模较大的特征可能会淹没维数较小的特征,这样的话,带来的直接结果就是尽管有多个不同类型的特征融合,但是实际上影响结构的却是其中维数较大的那部分特征. 此外,ELM 隐层节点的个数是随机产生的,并没有一个较精确的估计隐层节点个数的方法,若是隐层节点个数选择的不够恰当,对计算结果会产生较大的波动. 针对上述的问题,结合正则化的思想[11]提出了鲁棒性增量极限学习机,该算法的数学描述如下:

$$\varphi(\boldsymbol{Z}_{i\text{-multi-hog}})=[x_{1,1}..x_{1,p^1}|x_{2,1}..x_{2,p^2}|\cdots x_{d,p^d}] \tag{3}$$

式中 $\varphi(\boldsymbol{Z}_{i\text{-multi-hog}})$ 表示样本 $\boldsymbol{Z}_i$ 的多尺度 HOG 特征向量,d 表示 HOG 特征的尺度规模,p^d 为 d 尺度下的 HOG 特征维数.

$$\boldsymbol{V}=[v_1,v_2\cdots v_d] \tag{4}$$

式中 $v_u(1\leqslant u\leqslant d)$ 表示的是不同尺度下的 HOG 特征影响因子.

$$\varphi(\boldsymbol{Z}_{i\text{-lbp}})=\{x_{\mathrm{lbp},1},x_{\mathrm{lbp},2}\cdots x_{\mathrm{lbp},o}\} \tag{5}$$

式中 $\varphi(\boldsymbol{Z}_{i\text{-lbp}})$ 表示的是样本 $\boldsymbol{Z}_i$ 的多级 LBP 特征向量,其维数为 o.

$$\varphi(\boldsymbol{Z}_{i-\mathrm{hsv}})=\{x_{\mathrm{hsv},1},x_{\mathrm{hsv},2}\cdots x_{\mathrm{hsv},y}\} \tag{6}$$

式中 $\varphi(\boldsymbol{Z}_{i-\mathrm{hsv}})$ 表示的是样本 $\boldsymbol{Z}_i$ 的全局 HSV 特征向量,此外令 LBP,HSV 的特征影响因子分别为 $v_{\mathrm{lbp}},v_{\mathrm{hsv}}$.

特征影响因子的主要目的在于,防止高维数的特征对低维数的特征产生特征淹没,通过采用不同的特征影响因子,可以使得各个特征之间能够得到均衡的,有效的利用,避免了特征掩盖行为的发生.

$$\hat{\boldsymbol{Z}}_i=\{\boldsymbol{V}\circ\varphi(\boldsymbol{Z}_{i-\mathrm{multi\text{-}hog}}),v_{\mathrm{lbp}}\circ\varphi(\boldsymbol{Z}_{i-\mathrm{lbp}}),v_{\mathrm{hsv}}\circ\boldsymbol{Z}_{i-\mathrm{hsv}})\} \tag{7}$$

式中 $\hat{\boldsymbol{Z}}_i$ 表示的样本 $\boldsymbol{Z}_i$ 在特征影响因子作用下的空间向量.

$$F_{\text{RI-ELM}}=\frac{1}{2}\|\beta\|^2+\frac{1}{2}k\sum_{i=1}^{n}\varepsilon_i^2 \tag{8}$$

$$\text{s.t}\quad h(\hat{\boldsymbol{Z}}_i)\boldsymbol{\beta}+\boldsymbol{\varepsilon}_i^{\mathrm{T}}=\boldsymbol{t}_i^{\mathrm{T}}i=1,2\cdots n \tag{9}$$

式(8)中 $F_{\text{RI-ELM}}$ 的第 1 项表达式表示的是经验风险,第 2 项表达式表示的则是置信风险,k 则是用于调节置信风险的因子,$h(\hat{\boldsymbol{Z}}_i)$ 表示的是输入层的样本到隐藏层的一个映射,对于隐层节点个数为 l 的单隐层 ELM 网络,$h(\hat{\boldsymbol{Z}}_i)$ 的维数为 L 维. 用 β_j 表示的是链接隐层和第 j 个节点的输出权值,m 表示的是标注词的个数,因此对应 m 个输出节点. 此时对应的 $\boldsymbol{\beta}=[\beta_1,\beta_2\cdots\cdots\beta_m]$. 若某一个图像样本 $\boldsymbol{Z}_i$,对应的标注词是第 $p(1\leqslant p\leqslant m)$ 个,则其输出向量可以表示为 $\boldsymbol{t}_i=[t_{i,1},t_{i,2},\cdots..t_{i,p}\cdots\cdots.t_{i,m}]^{\mathrm{T}}$,其中满足 $t_{i,p}=1$,其余位置上的值为 0. $\boldsymbol{\varepsilon}_i=[\varepsilon_{i1},\varepsilon_{i2}\cdots\cdots\cdots.\varepsilon_{im}]^{\mathrm{T}}$ 表示的是对应于样本 $\boldsymbol{Z}_i$ 的置信风险传递误差. 本文之中右上角的 T 均代表的是向量的转置. 为求解上述问题我们引进广义拉格朗日函数:

$$\boldsymbol{L}_{\text{RI-ELM}} = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{1}{2}k\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i^2 - \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_{i,j}(h(\hat{\boldsymbol{Z}}_i)\beta_j + \varepsilon_{i,j} - t_{i,j}) \quad (10)$$

设 $\boldsymbol{\alpha}_i = [\alpha_{i,1},\alpha_{i,2}\cdots\cdots\alpha_{i,m}]^{\mathrm{T}}$,表示拉格朗日乘子. 利用拉格朗日的对偶性,原问题的求解等价于其对偶优化问题的求解有:

$$\boldsymbol{\beta} = (\frac{1}{k}\boldsymbol{I} + \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H} \quad (112)$$

式中 $\boldsymbol{I}$ 表示的是单位矩阵,因此最终的标注词输出函数为:

$$f(\boldsymbol{Z}) = h(\boldsymbol{Z})(\frac{1}{k}\boldsymbol{I} + \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H} \quad (12)$$

其中 $f(\boldsymbol{Z}) = [f_1(\boldsymbol{Z}),f_2(\boldsymbol{Z})\cdots\cdots f_m(\boldsymbol{Z})]^{\mathrm{T}}$,针对多分标注的问题,其标注标签为:

$$\text{label}(\boldsymbol{Z}) = \arg\max f_i(\boldsymbol{Z})\ i = 1,2\cdots m \quad (13)$$

当训练样本的个数 n 较大的时候,且 $n \gg l$,计算关于训练样本 $n\times n$ 的矩阵可以转化为计算关于隐层节点 $l\times l$ 的矩阵,大大的减少了计算量.

关于隐层节点个数 l 的确定,本文提出一种增量反馈方法,通过不断调整隐层节点的个数使得整个模型能够达到最优,具体的步骤如算法 2 所示.

算法 2　隐层节点个数增量反馈算法

输入:随机产生一个由输入层到隐层的权重矩阵 $\boldsymbol{W}_{\text{input}}$,生成一个具有 l_{begin} 个节点的极限学习机网络($l_{\text{begin}} \ll n$).

输出:最佳隐层节点个数 l_{best}.

初始化阶段:计算此时网络对应的残差 $E(\text{error}_1) = \|\boldsymbol{H}\boldsymbol{\beta}_1 - \boldsymbol{T}\|$. // $\boldsymbol{T}$ 表示的是极限网络的输出向量.

学习阶段:

(1) 设置初始计数值 $c=1$,确定一个残差因子 σ,σ 为一个很小的实验双精度数. 确定一个步长因子 s,$s\in\mathbf{N}^+$.

(2) for $l_c = l_{\text{begin}}:n$

(3) if($l_c < n$) $c = c+1$, else $l_{c+1} = l_{\text{begin}} + sl_c$.

(4) compute β_2, $E(\text{error}_2) = \|\boldsymbol{H}\boldsymbol{\beta}_2 - \boldsymbol{T}\|$.

(5) if(($l_{c+1} < n$) && $\|E(\text{error}_2) - E(\text{error}_1)\| > \sigma$) goto(3)

(6) else output $l_{\text{best}} = l_c$

(7) end for.

该方法从多个角度对极限学习机进行了完善和扩展,使其具有更强的学习和泛化能力,进而可以较大的程度上提高整个图像自动标注模型的性能.

4　基于蒙特卡罗数据集均衡的图像自动标注

4.1　基于综合距离的图像特征匹配(CDIFM)

对于分割后的图像,机器并不知道分割后的区域是究竟对应的是哪个标注词. 针对该问题,本文提出一种基于综合距离的图像特征匹配方法(CDIFM),将具有同一个标注词描述的分割区域自动串联在一起,该方法简单有效,具有较高的准确率. 该方法的描述如下:

步骤 1　首先选择一个分割后的区域作为种子图像,例如我们选择了分割后标注词为 bear 的一个分割区域作为种子图像,并令该种子图像为 $\boldsymbol{X}_{\text{bear}}$. 设置一个置信因子 τ,当两个区域的特征相似度距离小于置信因子时,我们认为这两个区域同属于一个类别或者说两者应该具有相同的关键词描述.

步骤 2　从图像库中选择含有 bear 这一标注词的图像集 $\varphi(\boldsymbol{Y}) = \{\boldsymbol{Y}_1,\boldsymbol{Y}_2\cdots\boldsymbol{Y}_k\}$ $(1\leqslant k\leqslant n)$,其中 n 表示的是图像库中的图像个数. 假设满足该要求的某一张图像 $\boldsymbol{Y}_k$,其分割后的区域为 $\varphi(\boldsymbol{Y}_k) = \{\boldsymbol{X}_1,\boldsymbol{X}_2\cdots\boldsymbol{X}_c\}$ $(1\leqslant c\leqslant 5)$,因为每张主题照片的关键词不超过 5 个,因此分割区域一般控制在 5 个之内,并且分割的准则根据该张图片的标注词个数.

步骤 3　设 $\boldsymbol{X}_{\text{bear}}$ 的多尺度融合特征向量为 $\boldsymbol{X}_{\text{bear}} = \{x_{\text{bear1}},x_{\text{bear2}}\cdots x_{\text{bearm}}\}$,$m$ 代表其特征维度. 其中多尺度下的 HOG 特征向量为 ω 维,LBP 的特征向量维数为 σ 维,HSV 的特征向量维数为 δ 维(满足 $\omega+\sigma+\delta=m$),因此该样本的特征向量亦可以描述为:

$$\boldsymbol{X}_{\text{bear}} = [x_{\text{bear},1}\cdots x_{\text{bear},\omega}|x_{\text{bear},\omega+1}\cdots x_{\text{bear},\omega+\sigma}|x_{\text{bear},\omega+\sigma+1}\cdots x_{\text{bear},\omega+\sigma+\delta}].$$

同上述,我们可以采用同样的方式表示 $\boldsymbol{X}_c$ $(1\leqslant c\leqslant 5)$ 区域的特征向量. 因此,我们不妨设 $\boldsymbol{X}_c = [x_{c,1}\cdots x_{c,\omega}|x_{c,\omega+1}\cdots x_{c,\omega+\sigma}|x_{c,\omega+\sigma+1}\cdots x_{c,\omega+\sigma+\delta}]$,并且给出以下的数学定义:

$$M_{X_{\text{bear}},X_c} = \lambda_1 d_1 + \lambda_2 d_2 + \lambda_3 d_3 \quad (14)$$

式(14)定义 M_{X_{bear},X_c} 表示 $\boldsymbol{X}_{\text{bear}}$,$\boldsymbol{X}_c$ 两者的综合特征距离,$\lambda_1,\lambda_2,\lambda_3$ 分别表示不同特征距离的影响因子.

$$d_1 = \sum_{i=1}^{\omega}\sqrt{(x_{\text{bear},i} - x_{c,i})^2} \quad (15)$$

式(15)定义 d_1 表示的是两者多尺度 HOG 特征下的欧式特征距离.

$$d_2 = \sum_{i=1}^{\sigma}|x_{c,\omega+i} - x_{\text{bear},\omega+i}| \quad (16)$$

式(16)定义 d_2 表示的是两者 LBP 特征下的曼哈顿特征距离

$$d_3 = \sum_{i=\omega+\sigma+1}^{\omega+\sigma+\delta}\min(x_{\text{bear},i},x_{c,i})/\min(\sum_{i=\omega+\sigma+1}^{\omega+\sigma+\delta}x_{\text{bear},i},\sum_{i=\omega+\sigma+1}^{\omega+\sigma+\delta}x_{c,i}) \quad (17)$$

式(17)定义 d_3 表示两者 HSV 特征下的直方特征距离. 当 $M_{X_{\text{bear}},X_c} < \tau$ 时,我们认为 $\boldsymbol{X}_c$ 区域为 $\boldsymbol{X}_{\text{bear}}$ 的最佳匹配分割区域,并将该区域加入该种子图像集之中. 重复步骤 2 直到 $\varphi(\boldsymbol{Y})$ 中的元素均计算过.

本文分别尝试了不同的距离作为特征匹配的统一准则,但是通过实验证明了在不同特征下采用不同特征距离衡量的这样组合方式,能够很好的进行图像之

间的特征匹配,计算量较小,精度高,能够实现图像区域的快速匹配. 欧式距离在基于图像局部梯度信息的特征匹配上具有良好的性能,曼哈顿距离在基于图像纹理信息的特征匹配上具有良好的性能,直方距离在基于颜色信息的特征匹配上具有良好的性能.

4.2 蒙特卡罗数据集均衡(MC-BDS)

大多数的图像自动标注模型,主要是通过模型系统本身的完善和改进来提高图像自动标注的效果,往往忽略数据集本身对模型存在的影响. 此外大多数的图像标注系统对于训练集中较少的某个标注词,往往在测试的时候无法自动标注出来,更多的都是倾向于训练集中较多的或者大量出现的某类或者某些标注词. 这就是所谓的数据集不平衡的问题.

本文针对以上的问题,提出蒙特卡罗数据集均衡算法(MC-BDS)来平衡不同类别之间的数据集规模,以达到提高图像自动标注的整体性能.

本文基于上述思想所提出的蒙特卡罗数据集均衡算法(MC-BDS)可以描述为:设 $\varphi(\boldsymbol{X})=\{\boldsymbol{X}_1,\boldsymbol{X}_2\cdots\cdots\boldsymbol{X}_n\}$,表示某一类别的训练集图像集合. $\boldsymbol{X}_i(1\leqslant i\leqslant n)$表示的是该类别中的某个图像样本,$\boldsymbol{X}_i=[x_{i1},x_{i2}\cdots\cdots..x_{im}]$,并给出以下数学定义:

算法 3 蒙特卡罗数据集均衡算法(MC-BDS)

输入:$\varphi(\boldsymbol{S})=[s_1,s_2\cdots..s_c]$. //表示的是图像分类集合,其中 c 表示的是标注词分类数,φ 表示的是集合.

输出:$\varphi(\hat{\boldsymbol{S}})=[\hat{s}_1,\hat{s}_2\cdots..\hat{s}_c]$. //蒙特卡罗数据集均衡之后的图像集合,其中 c 表示的是标注词分类数.

初始化阶段:

1. 选择合适的 μ,η,ρ,该参数分别表示不同特征度量下的调节因子.

2. $\tilde{s}=\text{count}(s_1)+\text{count}(s_2)+\cdots\cdots\text{count}(s_c)/c$(分类均衡数). //其中 $\text{count}(s_j)(1\leqslant j\leqslant c)$表示的是该类别对应的图像集数目.

3. 选择合适的碰撞因子 β.

数据均衡阶段:

(1) for $i=0:c$

//对公共图像库中所有类别的标注词进行合理的均衡.

(2) compute $\text{count}(s_i)$

//$count(s_i)$表示的是各个不同类别的图像数目.

(3) if($\text{count}(s_i)<\tilde{s}$){$s_e=\tilde{s}-\text{count}(s_i)$, goto(4), else{$i=i+1$, goto(1)}}

//若是某个类别或者某些类别对应的图像集数目低于分类均衡数,则需要进行数据集扩充,并且扩充的大小为

$$s_e=\tilde{s}-\text{count}(s_i)(1\leqslant i\leqslant c,s_e>0).$$

(4) compute $T_{\min}=\min\{T_{X_{ip}}\}(1\leqslant p\leqslant\text{count}(s_i))$, $T_{\max}=\max\{T_{X_{ip}}\}(1\leqslant p\leqslant\text{count}(s_i))$ save $\boldsymbol{X}_{\min}=[x_{\min1},x_{\min2}\cdots\cdots x_{\min m}]$, $\alpha=x_{\min1}+x_{\min2}\cdots\cdots+x_{\min m}/m$.

//计算不同类别标注词的最大加权复合特征距离和最小加权复合特征距离,并且保存最小复合特征距离对应的特征向量作为基准向量. 其中:$\boldsymbol{T}_{X_{ip}}$表示的是第 i 个类别的图像在其对应的集合中第 p 个元素的加权复合特征距离. $\alpha=x_{\min1}+x_{\min2}\cdots\cdots+x_{\min m}/m$,表示图像特征的均衡步长.

(5) for $k=0:s_e$

(6) random $Y_i=[y_{i1},y_{i2}\cdots\cdots.y_{im}]$ st. $\mu(\boldsymbol{Y}_i)=0,\sigma(\boldsymbol{Y}_i)=1$.

//使用 Box-Muller 算法产生高斯随机向量 $\boldsymbol{Y}_i=[y_{i1},y_{i2}\cdots\cdots.y_{im}]$,本步骤的核心思想是先得到服从均匀分布的随机数,再将服从均匀分布的随机数转变为服从高斯分布,该算法能够在极短时间内产生所需的随机数.

(7) $\boldsymbol{X}_e=\boldsymbol{X}_{\min}+\alpha e^{\beta}\boldsymbol{Y}_i,\boldsymbol{X}_e=[x_{e1},x_{e2}\cdots\cdots x_{em}]$

//其表示的是一个与训练图像具有同样特征维数大小的向量,例如本文中图像的多尺度融合特征的大小为 m 维,则该向量的大小也为 m 维.

(8) compute T_{X_e}//计算此时的加权复合特征距离

(9) if($\boldsymbol{T}_{\min}<\boldsymbol{T}_{X_e}<\boldsymbol{T}_{\max}$)

(10) $\varphi(\hat{s}_i)=\varphi(s_i)+\boldsymbol{X}_e,k=k+1$

//如果 $\boldsymbol{T}_{\min}<\boldsymbol{T}_{X_e}<\boldsymbol{T}_{\max}$,那么则将 $\boldsymbol{X}_e$ 称为合理的均衡样本,并把该样本加入该类图像的训练集合之中. 即此时 $\varphi(\boldsymbol{X})=\{\boldsymbol{X}_1,\boldsymbol{X}_2\cdots\cdots\boldsymbol{X}_n,\boldsymbol{X}_e\}$.

(11) else goto(6).

(12) if($k>s_e$) end for k {$i=i+1$; goto(2)}

(13) if($i>c$) end for i

//当所有需要均衡的类别图像采用上述算法合理的扩充后,均衡算法结束.

$$\boldsymbol{X}_i=[x_{i1},x_{i2}\cdots..x_{i\xi}|x_{i,\xi+1},x_{i,\xi+2}\cdots x_{i,\xi+\psi}|x_{i,\xi+\psi+1},x_{i,\xi+\psi+2}\cdots x_{i,\xi+\psi+\zeta}] \tag{18}$$

上式中 $\boldsymbol{X}_i$ 表示的是样本图像,其中多尺度下的 HOG 特征向量为 ξ 维,LBP 的特征向量为 ψ 维,HSV 的特征向量为 ζ 维(满足 $\zeta+\xi+\psi=m$).

$$d_{\text{i-ho}}=\sqrt{x_{i1}^2+x_{i2}^2\cdots+x_{i\xi}^2} \tag{19}$$

$d_{\text{i-ho}}$表示的是多尺度下 HOG 特征的空间度量.

$$d_{\text{il-z}}=\sum_{p=i,\xi+1}^{i,\xi+\psi}|x_{i,p}| \tag{20}$$

$d_{\text{il-z}}$表示的是 LBP 特征的空间度量.

$$d_{\text{iv-f}}=\min\{x_{i,r}\}/\sum_{r=i,\xi+\psi+1}^{i,\xi+\psi+\zeta}x_{i,r}(\xi+\psi+1\leqslant r\leqslant\xi+\psi+\zeta) \tag{21}$$

$d_{\text{v-f}}$表示的是 HSV 特征的空间度量.

$$T_{X_i}=\mu d_{\text{i-ho}}+\eta d_{\text{il-z}}+\rho d_{\text{v-f}} \tag{22}$$

T_{X_i}为本文提出的关于图像特征的一个全新度量标准-加权复合特征距离,μ,η,ρ 分别表示不同特征度量下的调节因子.

4.3 图像分割策略

将图像进行不同栅格数的区域分割. 栅格划分中心区域一般可以作为一个关键词描述,此外可能存在对多个区域的描述是同一个关键词. 针对该问题,本文的解决方法是将每一个栅格化的区域作为一个独立描述,计算其多尺度下的特征融合向量,输入到训练好的自动标注模型之中,选择得分在前 $k(k\geqslant5)$名的关键词

作为潜在的可行性标注词. 最后累计加权累计整幅图像之中得分前 $k(k\geqslant 5)$ 名的关键词作为该幅图像的标注词. 该算法的描述如下:

步骤 1　从待测标注数据集之中选择一张图像 $\boldsymbol{P}_i$ $(1\leqslant i\leqslant N_t)$,$N_t$ 表示待标注图像的个数. 对该图像进行栅格化划分,划分后的区域为 $\boldsymbol{P}_i=\{\boldsymbol{p}_{i1},\boldsymbol{p}_{i2}\cdots\cdots\boldsymbol{p}_{in}\}$. 这里的 n 表示栅格区域的个数.

步骤 2　分别计算出区域 $\boldsymbol{p}_{i1},\boldsymbol{p}_{i2}\cdots\cdots\boldsymbol{p}_{in}$ 多尺度下的特征融合向量,将其作为输入值输入到训练好的图像自动标注模型之中. 求出各个区域在得分前五名的关键词以及其对应的得分. 不妨设 $\varphi(p_{in}^k)=\{\text{key}_{in1}-\text{score}_{in1},\text{key}_{in2}-\text{score}_{in2},\cdots\text{key}_{ink}-\text{score}_{ink}\}$,分别代表该区域得分前 $k(k\geqslant 5)$ 名的关键词以及其对应的得分.

步骤 3　按照步骤 2 的方法,计算出该图像其余分割区域得分前五的关键词及其得分. 若是分割的区域为中间区域或者中间领域我们对其赋值一个权重 λ,中间区域或者中间领域的最后得分为该权重乘以原来的得分值. 其余的非中间区域得分不变.

步骤 4　该幅图像分割的区域里,若是关键词相同的话则对应的得分进行累加,并且关键词不变. 若是关键词均不同的话,则保留相应的关键词以及其对应的得分. 统计该幅图像中最后得分前 5 名的关键词作为该幅图像的描述.

步骤 3　重复步骤 1 ~ 4 的方法选择测试集中的其余图像,直到测试集为空,算法停止.

4.4　基于 TF-IDF 的图像自动标注改善

一般来说,在一幅标注了关键词的图像之中,关键词与关键词之间可能存在着各式各样的语义关系. 我们采用了自然语言处理领域中文本检索中使用比较广泛的 TF-IDF 思想来对共生关系做一个度量[5],其数学表达式如下:

$$K(w_1,w_2)=K_c(w_1,w_2)\times\log\left(\frac{N}{n_1}\right)\tag{23}$$

上述式子中:我们使用 w_1,w_2 来表示图像标注模型中的所有关键词,$K_c(w_1,w_2)$ 表示的是关键词 w_1,w_2 两者在测试集图像中共同出现的次数,n_1 表示的是关键词 w_1 在测试图像集中出现的次数,N 表示的是训练集中的图像总和. 特别要注意的是;上述的表达式并不是一个对称式,也就是说 $K(w_1,w_2)$ 的值和 $K(w_2,w_1)$ 的值并不恒等. 因为如果 w_1,w_2 是两个在词频上差异较大的关键词,w_1 所体现出来的是一种比较具体的概念,w_2 所体现出来的是相对抽象的概念,当两者共同出现的频率较高的时候. 我们可以从 w_1 的存在来推断出 w_2 有较大的可能性也会存在. 但是反之则未必.

5　实验

5.1　实验设置

个人 PC 机一台,系统为 Windows7(64 位)系统,运行内存 12G,CPU 核心数为双核. 采用公共图像库 Corel-5K、EspGame 和 Iaprtc12. 这三个数据集是目前为止使用的最为广泛的自动图像标注公共图像库. 图像库由三个部分所组成:第一部分称为训练部分,主要用于训练模型所使用的. 第二部分称为参数调整部分,主要标注模型中参数的调优. 第三部分称为测试部分,用于对图像自动标注模型的性能进行测试. 此外,该图像库中每一幅图像上含有 1 ~ 5 个标注词不等,该标注词共同来表达本幅图像所反映的语义信息.

5.2　图像自动标注的评价标准

我们用准确率 P、召回率 R、F_1 值和被准确预测过的标签数 N^+ 进行评测. 各评价指标按如下公式计算:

$$P=\frac{1}{M}\sum_{j=1}^{M}\frac{\text{Correct}(w_j)}{\text{Predicted}(w_j)},\tag{24}$$

$$R=\frac{1}{M}\sum_{j=1}^{M}\frac{\text{Correct}(w_j)}{\text{Ground}(w_j)},\tag{25}$$

$$F_1=\frac{2\cdot P\cdot R}{P+R},\tag{26}$$

其中,$\text{Correct}(w_j)$ 为关键词 w_j 的正确预测数,$\text{Predicted}(w_j)$ 为关键词 w_j 总预测数,$\text{Ground}(w_j)$ 为关键词 w_j 实际标注数. 被准确预测过的标签数 N^+ 作为图像自动标注泛化性能的衡量标准. 其含义是至少被正确标注过一次的标注词数量,这个数值反映了图像自动标注模型对标注词的覆盖程度,是图像自动标注算法性能好坏的一个主要标准.

5.3　实验参数的设置

5.3.1　蒙特卡罗数据集均衡数 s_e 的确定

本文统计了训练集中其部分标注词所关联的图像出现的频数,以及在数据集未均衡的情况下,该部分标注词在标注结果中的准确率和召回率,实验结果如图 2 所示:

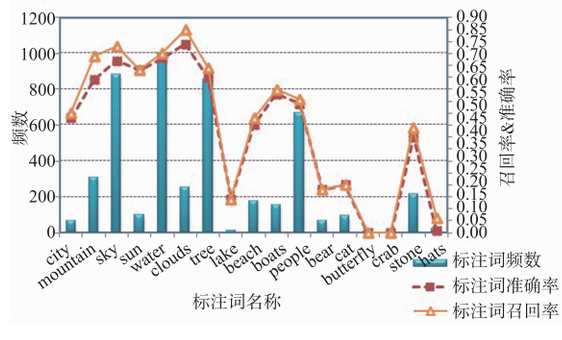


图2　Corel5k未进行数据集均衡情况下标注词的频数,准确率和召回率的关系

从图中可以看出,图像库中不同类别的标注词之间所关联的图像集,其数据规模具有较大的差别,例如 water 标注词与其关联的图像有 1000 多张,lake,crab 等标注词与其关联的图像只有 10 来张. 相互之间的数据集规模相差从几倍到上百倍不等,这样想要构建出一个合理的分类模型是相当困难的. 例如在训练集中出现词频较高的关键词 sky,water,people,tree 往往能够获得较好的标注准确率和召回率,而在训练集中词频出现较少的关键词,例如 butterfly,crab 甚至一次标注正确的都没有.

本文提出蒙特卡罗数据集均衡算法(MC-BDS),使得均衡之后的各个不同类别之间的数据集规模大致平衡或者之间差别较小. 本文采用加权复合特征距离作为图像扩充的准则,以下为部分标注词其加权复合特征距离的分布情况.

表 1 不同标注词的加权复合特征距离

标注词名称	加权复合特征距离下限	加权复合特征距离上限
city	5. 6782	6. 9325
bear	11. 6356	12. 7411
mountain	6. 8217	7. 6452
cat	12. 3641	13. 2829
sun	9. 2418	10. 6722
butterfly	14. 2025	16. 7318
lake	10. 6325	11. 7187
crab	17. 2150	18. 6336
hats	13. 3702	14. 1923

由表 1 可知:不同标注词之间的加权复合特征距离具有一定差异性. 有部分不同类别的标注词之间的加权复合特征距离上限和下限可能会存在少部分的重叠,但是并不会影响到数据集的均衡扩充. 其原因主要有以下两点:

(1)本文中提出的数据均衡扩充算法(MC-BDS),其扩充的数据的加权复合特征距离是均匀分布在其类别的下限和上限之间,不会集中在下限领域或者上限领域,因此能保证扩充之后的数据的特征距离,能够较大程度上远离重叠区域.

(2)假设标注词 A 的特征为 $\boldsymbol{X}_A=\{x_1,x_2\cdots\cdots.\ x_m\}$,新样本的特征的产生是基于原始样本,也就是说新产生的数据特征 $\boldsymbol{X}_{\text{new}}=\{x_1\circ\varepsilon_1,x_2\circ\varepsilon_2\cdots\cdots.\ x_m\circ\varepsilon_m\}$. 其中,$\circ$ 表示分别作用在各个特征上的运算.

本文分别通过选择不同的均衡扩充数 s_e,从平均准确率和平均召回率以及召回数来确定合理的均衡数,实验结果如图 3 和图 4 所示:

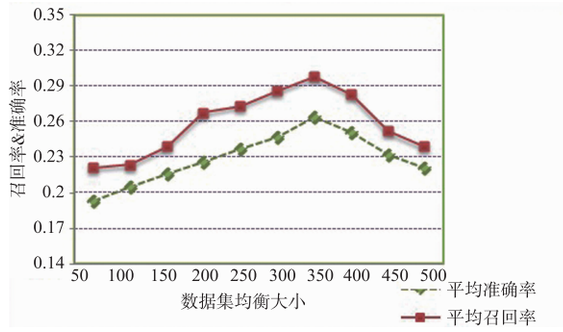


图3 Corel5k数据集均衡数与准确率, 召回率的关系

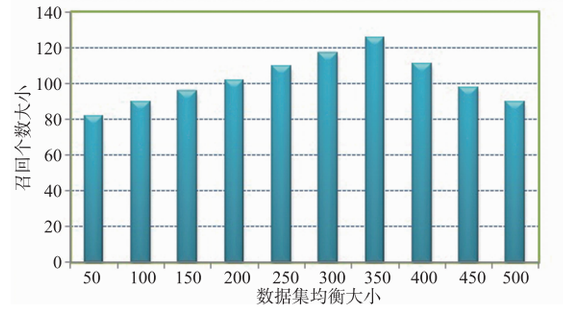


图4 Corel5k数据集均衡数与召回数的关系

从上图中我们可以看出,如果能够选择合理的均衡数 s_e 的话,可以从标注的平均准确率、召回率以及召回个数等方面较大的提高整个标注模型的性能. 若是均衡数 s_e 选择不合理的话会对标注模型带来一定的影响.

从实验的结果我们可以分析,合理的均衡数应该在训练集中所有词频次数的均值附近浮动,这也就是本文的提出的蒙特卡罗数据集均衡算法(MC-BDS),在均衡数上选择的指导原则. 采用数据均衡算法之后,再次统计了训练集中的部分标注词与其关联的图像的频数以及均衡之后的准确率和召回率,如图 5 所示.

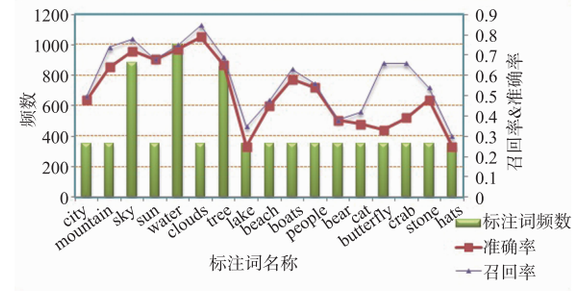


图5 Corel5k数据均衡之后标注词的频数, 准确率和召回率的关系

从图中我们可以得知,本文提出的数据集均衡扩充算法对低频词具有较好的鲁棒性,并且原有高频词汇的标注性能并没有出现较大的浮动,仍然可以维持

在一个较高的水平. 从而使得整个图像自动标注模型的性能得到综合的提高.

5.3.2 关于待标注图像栅格化数 K 和区域个数 N 的讨论

(1)待标注图像的栅格划分区域数 n 应该选择哪一个值比较合适并没有严格的数学依据,如果区域划分过大的话,可能会带来将两个甚至多个不同类别的对象划分在同一个区域中. 若是分割区域过细,那么每个区域内的特征描述信息量将会非常少,不利于机器对其的学习.

(2)参数 k 表示的是对分割后的区域选择前 k 个最大得分值的关键词作为标注的可行依据. 当 k 选择过小的时候会存在中心信息淹没背景信息,即中心区域可以得到很好的判别,但是边缘区域可能得不到累加就被去除. 当 k 选择过大的时候,区域分割之后的投票标注便会转化为对该幅图像的直接标注,失去的区域分割的优势. 最后,本文通过实验集不断对 k 和 n 进行交互调整,得出当 $k=8$,$n=2*3$ 时,可以取得最优值.

5.4 不同特征实验对比

本文在 Corel5K 数据集中做了如下实验论证. 分别测试单独使用不同特征的实验结果、采用融合特征的实验结果以及传统极限学习机与鲁棒性增量极限学习机的实验对比,实验结果分别如表 2 所示. 从实验结果可以看出,采用本文的鲁棒性增量极限学习机相比传统极限学习机在图像标注效果有了一定的提升. 因此本文鲁棒性增量极限学习机是有效的.

表 2　不同特征在 Corel5k 上的实验对比

特征类别	准确率(P)	召回率(R)	调和值(F_1)	$N+$
RIELM(HOG)	13	14	13	87
RIELM(LBP)	11	13	12	81
RIELM(HSV)	18	22	20	98
ELM(HOG + LBP + HSV)	21	22	21	103
RIELM(HOG + LBP + HSV)	23	26	24	112

5.5 实验结果对比

我们将通过本文提出的基于蒙特卡罗数据集均衡和鲁棒性增量极限学习机的图像自动标注模型(MB_IELM)与国内外已发表的经典图像自动标注模型在公共数据集 Corel-5k、Esp Game 和 IAPRTC-12 图像库上进行比较,以此来验证本文提出的模型框架的可行性以及有效性. 这里涉及的比较方法包括:Corr-LDA[4], JEC[5], SML[6], MBRM[12], MLR-GL[13], GMM-Mult[14], GMM[15],这些都是国外已发表的期刊/会议论文中的经典图像自动标注模型,实验结果如表 3 所示.

从表 3 我们可以看出,本文提出的图像自动标注模型是有效可行的,标注的综合性能要好于目前使用比较广的几种标注模型,其中在 Corel5K 数据集上的准确率 P 达到 26%,均不低于上述模型;召回率 R 达到 29%,略低于上述模型最高值 33%(Corr-LDA);F_1 值达到 28%,略低于上述模型最高值 29%(JEC);在衡量标注词覆盖程度的召回数有 124 个关键词至少被正确的标注过一次,具有较好的覆盖率,低于 Corr-LDA、JEC 和 SML 模型. 在 ESPGame 数据集上的准确率 P 达到 33%,均高于上述模型;召回率 R 达到 22%,低于 JEC 和 GMM-Mult 模型,但高于其它模型;F_1 值达到 26%,略低于上述模型的最高值 27%(GMM-Mult),但均高于上述其它模型;在衡量标注词覆盖程度的召回数有 240 个关键词至少被正确的标注过一次,具有上述模型最好的覆盖率. 在 IAPRTC-12 数据集上的准确率 P 达到 34%,均高于上述模型;召回率 R 达到 24%,低于 JEC 模型的 29%,但高于其它模型;F_1 值达到上述模型最高的 28%;在衡量标注词覆盖程度的召回数有 248 个关键词至少被正确的标注过一次,具有上述模型较好的覆盖率,只略低于最高值 250(JEC). 除此之外,本文提出的自动标注模型相对于低频词汇以及传统的人工标注亦表现出一定的优良性能,此外,为了说明这个问题我们给出了些反应本模型特点的标注结果,如表 4 所示.

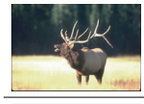
表 3　Corel-5k、ESPGame、IAPRTC-12 上本文提出的模型与其它模型实验结果对比

Dataset→	Corel-5k				ESP Game				IAPRTC-12			
Method ↓	P	R	$F1$	$N+$	P	R	$F1$	$N+$	P	R	$F1$	$N+$
MBRM	24	25	24	122	18	19	18	209	24	23	23	233
MLR-GL	15	13	13	74	19	15	16	181	19	13	15	169
Corr-LDA	22	33	26	138	21	19	20	201	24	21	22	207
GMM-Mult	19	31	24	104	29	26	27	224	28	22	25	227
JEC	27	32	29	139	22	25	23	224	28	29	28	250
SML	23	29	25	137	—	—	—	—	—	—	—	—
RIELM (HOG + LBP + HSV)	23	26	24	112	30	19	23	231	32	22	26	241
MB_RIELM	26	29	28	124	33	22	26	240	34	24	28	248

表 4 的 5 幅图像是本文提出的图像自动标注模型的标注结果,每一幅图像中标注的五个结果分别按照得分大小降序排列. 在图像自动标注结果的一栏中,我们用黑色加粗的字体表示的是自动标注的结果之中具有和原始图像中人工标注结果相同含义的标注词,而使用斜体字体来表示自动标注结果之中和原始人工标注结果不同的标注词. 在这里,我们并没有选择完全被

标注正确的那些图像,而是选择了部分能够比较好反应本文模型特点的一些图像. 从表中可以看出,本文一些图表的标注结果虽然和原始图像上的人工标注结果有区别,但是确实对原始图像标注结果的有益补充,能够更加准确的描述图像的语义信息. 例如第一幅图像人工标注上并未将 bird 这一关键词给标注上,而从图像的场景来看,bird 显然要作为一个重要的关键词来描述该幅图像的场景. 在第三幅图像中,从人的视觉角度分析,显然用 sea 这个关键词相比原始图片中的 water 更有说服力,并且原始图像中也疏漏了 sky 等从图像中可以直接得到的关键词. 此外,在对抽象概念 maui, kauai 等描述上,原始图像中的信息并不能对其进行准确的描述,或者说,单从人的视觉角度来出发,图像上无法得到这些信息. 因此,也从另一个角度说明了人工标注存在的一些问题,可能存在漏标注,并且不同人对同一幅图像的认识也存在一定的主观差异.

表 4 本文模型的自动标注结果

图像	人工标注结果	本文标注结果
	sun,sea	sun,sky,cloud,sea, bird
	grass,antlers,elk	grass,antlers,elk,tree, plant
	water,beach, people,hawaii	sea,people,beach,sky, tree
	mountain,jet, plane	jet,piane,mountain, sky,tree
	sky,sand,elephant, desert	sky,sand,ground, elephant,desert

6 总结

本文重点对基于分类思想的图像自动标注模型进行了进一步研究,提出了一种基于蒙特卡罗数据集均衡和鲁棒性增量极限学习机的图像自动标注模型(MB_IELM). 简单的说,针对传统的基于分类的图像标注算法的求解往往导致局部最优问题,本文提出了一种鲁棒性增量极限学习机进行学习训练. 针对传统的特征提取算法对图像的描述相对单一,本文提出了多尺度特征融合提取算法对其进行准确描述. 针对实现不同类别标注词的大规模自动匹配,本文提出了基于综合距离的图像特征匹配算法. 最后针对公共图像库上的训练数据集规模不平衡性导致标注结果往往倾向于高频词而淹没低频词,本文提出了蒙特卡罗数据集均衡算法使得不同类别标注词之间具有合理数据规模. 通过公共数据集验证了本文提出的图像自动标注模型的合理性和有效性. 然而,在现有的图像自动标注模型中,针对例如 garden, autumn, street 等表达抽象的事物或者场景的关键词,在标注效果上仍然不尽人意,如何提高抽象词汇的标注准确率可以作为下一步的研究方向.

参考文献

[1] 罗惠兰,郭敏杰,孔繁胜. 一种基于多级空间视觉词典集体的图像分类方法[J]. 电子学报,2015,43(4):684-693.
Luo Hui-lan, Guo Min-jie, Kong Fan-sheng. An image classification method based on multiple level spatial visual dictionary ensemble[J]. Acta Electronica Sinica,2015,43(4):684-693. (in Chinese)

[2] Y Mori, H Takahashi, R Oka. Image-to-word transformation based on dividing and vector quantizing images with words[A]. First International Workshop on Multimedia Intelligent Storage and Retrieval Management[C]. Florida, USA:MISRM, 1999. 405-409.

[3] 杨栋,周秀玲,郭平. 基于贝叶斯通用背景模型的图像标注[J]. 自动化学报,2013,39(10):1674-1680.
YANG Dong, ZHOU Xiu-ling, GUO Ping. Image annotation with bayesian universal background model[J]. Acta Automatica Sinica, 2013, 39(10): 1674-1680. (in Chinese)

[4] Wang C, Blei D, Li F F. Simultaneous image classification and annotation[A] IEEE Conference on Computer Vision & Pattern Recognition[C]. Miami, FL, USA: IEEE, 2009. 1903-1910.

[5] Makadia A, Pavlovic V, Kumar S. Baselines for image annotation. [J]. International Journal of Computer Vision, 2010, 90(1): 88-105.

[6] Cabral, Ricardo, T Fernando De la, et al. Matrix completion for weakly-supervised multi-label image classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI), 2015, 37(1): 121-135.

[7] U Mlakar, B Potocnik. Automated facial expression recognition based on histograms of oriented gradient feature vector differences[J]. Signal, Image and Video Processing, 2015, 9(1): 245-253.

[8] 刘威,段成伟,遇冰,等. 基于后验 HOG 特征的多姿态行人检测[J]. 电子学报,2015,43(2):217-224.
Liu Wei, Duan Cheng-wei, Yu Bing. Multi-pose pedestrian detection based on posterior HOG Feature[J]. Acta Electronica Sinica, 2015, 43(2): 217-224. (in Chinese)

[9] 刘军,景晓军,孙松林,谭有恒. 一种基于可变长起主导

作用特征(VLDF)的人脸识别算法[J]. 电子学报,2015,43(3):544 – 549.
Liu Jun, Jing Xiao-jun, Sun Song-lin, Tan You-heng. A variable length dominant feature (VLDF) based algorithm for face recognition[J]. Acta Electronica Sinica, 2015, 43(3):544 – 549. (in Chinese)

[10] Z Bai, GB Huang, D Wang, et al. Sparse extreme learning machine for classification[J]. IEEE Transactions on Cybernetics, 2014, 44(10):1858 – 1870.

[11] Deng W, Zheng Q, Chen L. Regularized extreme learning machine[A]. CIDM '09 [C]. Nashville, Tennessee, USA:IEEE, 2009. 389 – 395.

[12] SL Feng, R Manmatha, V LaVrenko. Multiple bernoulli relevance models for image and video annotation[A]. Computer Vision and Pattern Recognition[C]. Washington, USA:IEEE, 2004. 1002 – 1009.

[13] Bucak S S, Jin R, Jain A K. Multi-label learning with incomplete class assignments[J]. Computer Visien & Pattern, 2011, 42(7):2801 – 2808.

[14] Barrat S, Tabbone S. Classification and automatic annotation extension of images using bayesian network[J]. Lecture Notes in Computer Science, 2008, 5342:937 – 946.

[15] F Shi, J Wang, Z Wang. Region-based supervised annotation for semantic image retrieval[J]. International Journal of Electronics and Communications, 2011, 65(11):929 – 936.

作者简介

柯　逍　男,1983 年 10 月生,福建省福州市人,博士,福州大学副教授,主要研究方向为计算机视觉、模式识别.
E-mail:kex@fzu.edu.cn.

邹嘉伟　男,1991 年 7 月生,福建省龙岩市人,硕士,主要研究方向为计算机视觉、模式识别.
E-mail:1468645610@qq.com.

杜明智　男,1988 年 2 月生,福建省泉州市人,硕士,主要研究方向为机器学习,图像处理.
E-mail:dmz1028@163.com.

周铭柯　男,1990 年 1 月生,福建省三明市人,硕士,主要研究方向为深度学习、计算机视觉.
E-mail:443810956@qq.com.