基于判别模型与生成模型的层叠图像自动标注*

柯 逍[1,2,3] 李绍滋[1,2] 曹冬林[1]

[1](厦门大学 信息科学与技术学院 智能科学与技术系 厦门 361005)
[2](厦门大学 福建省仿脑智能系统重点实验室 厦门 361005)
[3](福州大学 数学与计算机科学学院 福州 350108)

摘 要 图像自动标注是模式识别与计算机视觉等领域中重要而又具有挑战性的问题. 针对现有模型存在数据利用率低与易受正负样本不平衡影响等问题,提出了基于判别模型与生成模型的新型层叠图像自动标注模型. 该模型第一层利用判别模型对未标注图像进行主题标注,获得相应的相关图像集; 第二层利用提出的面向关键词的方法建立图像与关键词之间的联系,并使用提出的迭代算法分别对语义关键词与相关图像进行扩展; 最后利用生成模型与扩展的相关图像集对未标注图像进行详细标注. 该模型综合了判别模型与生成模型的优点,通过利用较少的相关训练图像来获得更好的标注结果. 在 Corel 5K 图像库上进行的实验验证了该模型的有效性.

关键词 图像自动标注,层叠模型,相关图像扩展,语义扩展
中图法分类号 TP 391

Hierarchical Image Automatic Annotation Based on Discriminative and Generative Models

KE Xiao[1,2,3], LI Shao-Zi[1,2], CAO Dong-Lin[1]

[1](*Cognitive Science Department*, *School of Information Science and Technology*, *Xiamen University*, *Xiamen* 361005)
2(*Fujian key laboratory of the Brain-like Intelligent Systems*, *Xiamen University*, *Xiamen* 361005)
3(*College of Mathematics and Computer Science*, *Fuzhou University*, *Fuzhou* 350108)

ABSTRACT

Image automatic annotation is a significant and challenging problem in pattern recognition and computer vision. Aiming at the problems that the existing models have low utilization and they are affected by unbalanced positive and negative samples, a hierarchical image annotation model is proposed. In the first layer, discriminative model is used to assign topic annotations to unlabeled images, and then the corresponding relevant image sets are obtained. In the second layer, a keywords-oriented method is proposed to establish links between images and keywords, and then the proposed iterative algorithm is used to expand semantic words and relevant image sets. Finally, a generative model is used to assign

* 国家自然科学基金项目(No. 60873179, 60803078)、高等学校博士学科点专项科研基金项目(No. 20090121110032)和深圳市科技计划基础研究项目(No. JC200903180630A)资助
收稿日期: 2010 - 04 - 26; 修回日期: 2010 - 08 - 16
作者简介 柯逍,男,1983 年生,博士研究生,主要研究方向为模式识别与计算机视觉. E-mail: kevinkexiao@ sina. com. 李绍滋,男,1963 年生,教授,博士生导师,主要研究方向为计算机视觉、多媒体信息检索等. E-mail: szlig@ xmu. edu. cn. 曹冬林,男,1977 年生,讲师,博士,主要研究方向为信息检索与模式识别等.

detailed annotations to unlabeled images on expanded relevant image sets. Hierarchical model uses less relevant training images to obtain better annotation results. Experimental results on Corel 5K datasets verify the effectiveness of proposed hierarchical image annotation model.

Key Words Image Automatic Annotation, Hierarchical Model, Relevant Image Expansion, Semantics Expansion

1 引 言

随着互联网相关技术的高速发展,互联网上图像与视频等多媒体信息飞速增长. 如何对图像与视频信息进行有效地组织与管理成了当前急需解决的热点问题. 图像与视频检索技术是解决上述问题的有效途径,而图像自动标注(Image Automatic Annotation) 是实现图像检索的关键步骤,其过程是根据图像视觉内容,由计算机系统自动产生图像对应的标注关键词. 早期的图像自动标注所关心的大多是图像的整体类别,更接近于图像分类领域. 而现在的图像自动标注主要针对图像中不同的视觉内容进行标注. 目前针对图像自动标注的方法主要有 2 大类: 一类是商业化图像搜索引擎采取的方式(如 Google, Yahoo! 等) ,即采用自然语言处理的相关技术,利用网页中图像的文件名、URL、ALT 标签以及上下文信息作为图像的标注,这类方法并没有使用图像内部特征,加之网络信息存在随意性与不确定性等特点,效果并不理想; 另一类方法主要从图像的视觉内容出发,根据图像的视觉内容产生相应标注词,可称为基于内容的图像自动标注,这类方法可以很好地构建图像视觉内容与关键词之间的联系. 本文所研究的图像自动标注就属于此类.

基于内容的图像自动标注模型从结构上可大致分为 2 类,分别是判别模型[1-3]与生成式模型. 这两类模型各有优缺点. 判别模型将每个语义关键词看成一个类别标签,通过监督学习的分类方法来对图像进行标注,其缺点是容易受到正负样本不平衡的影响. 生成式模型通过估计训练集中待标注图像与语义关键词的联合概率来进行标注,其缺点是容易受到语义鸿沟问题的影响,表现在其标注过程易受到那些视觉相似而语义不同图像的干扰,此外,该类方法对数据的利用率较低. 从模型理论上,判别模型与生成式模型具有一定的互补性. 本文通过构建层叠图像标注模型,从一定程度上解决了判别模型与生成式模型存在的问题,并将这两种模型优点相结合用于提升图像标注性能. 此外,现有模型在图像标注时几乎都需要计算待标注图像与所有训练图像之间的相似度关系,通过累积视觉生成概率与词汇生成概率来确定最终的标注. 而实际上训练库里的大部分图像都与待标注图像无关,利用这些大量无关的训练图像进行概率估计不仅会影响最终的标注结果,而且也会影响标注速度. 所以本文针对如何从整个训练库中有效地选取相关图像子集进行较为深入研究.

本文的创新点包括: 1) 提出层叠图像标注模型,通过利用较少的相关训练图像来获得更好的标注结果; 2) 提出面向关键词(Keywords - Oriented) 的方法来建立图像与关键词之间的联系; 3) 提出一个迭代算法用于扩展语义关键词与相关图像; 4) 结合判别模型与生成式模型的优点,并将其用于图像自动标注中.

2 相关工作介绍

图像自动标注与目标识别[4-7]有相似之处,但又不同于目标识别. 图像自动标注并不关心每个目标在图像中出现的具体位置,如在图像自动标注中,系统会将“汽车”、“赛道”等词作为某一幅图像的标注,但并不会具体标出“汽车”与“赛道”在图像中的位置. 目标识别系统一般是寻找特殊的前景物体,如行人、车辆与人脸等. 通常都是针对不同目标分别构建不同的分类器. 而在图像自动标注中,背景物体也同等重要. 图像自动标注需要一次处理几百个甚至更多的目标,同时学习所有的标注词以及获得每幅图像的若干个标注词. 目前,图像自动标注与目标识别都是具有实际意义,且有挑战性的研究领域.

近年来,研究者们利用各种机器学习方法与统计模型,建立图像视觉内容与标注关键词之间的联系. 同一关键词对应的视觉特征应该具有相似性,如“马”,其颜色和纹理在视觉特征上应基本保持一致. 这样,图像可以被分割成一些具有不同语义的块,分割后的每个块对应一个或两个语义对象,通过对这些图像块与语义关键词进行建模,就可以达到对图像进行标注的目的. 2002 年,Duygulu 等[8]提出机器翻译模型(Translation Model) ,将图像自动标注

看作是两种语言之间的翻译问题: 一种语言由描述图像内容的视觉词汇构成; 另一种语言由文本词汇构成. 通过 Normalized Cut [9] 将图像进行分割,并对图像中的所有区域利用 K-Means 算法进行聚类,得到视觉词汇 blob,图像的标注问题就可以看作是从视觉词汇 blob 到语义关键词的翻译过程. Blei 等 [10] 提出相关 LDA (Relevant Latent Dirichlet Allocation) 模型,将 LDA 分布扩展到词汇与图像,模型假设 Dirichlet 分布可以产生一些隐变量,而这些隐变量可以用于生成图像区域与关键词. Jeon 等 [11] 提出跨媒体相关模型(Cross Media Relevance Model, CM-RM) ,利用视觉关键字与语义关键字之间的联合概率进行标注,采用与机器翻译模型一样的离散特征进行表征区域特征,所以不可避免地带来一定的信息缺失. Lavrenko 等 [12] 提出连续相关模型(Continuous Relevance Model, CRM) ,它直接利用了图像区域的连续特征值,利用非参数高斯核进行连续视觉特征生成概率的估计. Feng 等 [13] 提出多伯努利相关模型(Multiple Bernoulli Relevance Model, MBRM) ,将图像划分为规则的矩形区域来取代复杂的图像分割算法,同时引入多伯努利分布取代多项式分布来表示词汇生成概率分布. Kang 等 [14] 提出互相关标记传播模型 (Correlated Label Propagation model, CLP) ,利用标记的相关性在相邻的图像间同时传播多个标记(每个标记对应一个词汇) . Liu 等 [15] 提出 AGAnn 模型,利用自适应图(adaptive graph) 与不同词之间的相关性改善标注结果. Gustavo 等 [16] 将提出 SML 模型,半监督学习引入图像自动标注中,从而避免了图像的分割过程. Yong 等 [17] 将全局特征、区域特征与上下文特征相结合并应用于扩展的 CM-RM 模型中. Stefanie 等 [18] 利用视觉分众分类(Visual Folksonomy) 思想,对 Flickr 图像库(http: //www.flickr.com) 的部分水果与蔬菜图像进行标注.

3　层叠图像自动标注

图像自动标注是一个多标签学习问题,如一幅图像可同时赋予“天空”、“草地”和“老虎”三个标注. 多标签学习问题与传统的多分类问题不同,传统的分类问题本质上是排它性的,而图像自动标注中的不同标注词对于某幅图像而言,都只是对应于图像的某些区域,所以这些标注词对某幅图像来说并不相互排斥. 但图像自动标注本身会存在一些问题: 1) 高层语义的多义性与歧义性. 图像的语义内容往往表现为用户对图像的一种主观理解,因此对图像的概念指派缺少确切的判断准则. 2) 语义概念的重叠性. 不同概念之间往往存在高层语义的重叠性,如“水”、“湖”以及“海”等,因此清晰地学习不同语义概念的边界是十分困难的.

在图像自动标注的 2 大类模型中,判别模型的缺点是容易受到正负样本不平衡的影响,而生成式模型的缺点是容易受到语义鸿沟问题的影响以及数据的利用率低. 本文通过构建层叠图像标注模型,从一定程度上解决了判别模型与生成式模型存在的上述问题,并将这 2 种模型相结合用于提升图像标注性能. 此外,现有模型对某一幅图像进行标注时,几乎都是通过计算该图像与所有训练图像的某种隐含概率联系来确定最后的标注. 例如,使用相关模型进行图像自动标注时,需要累积计算某一幅测试图像与所有训练图像之间的视觉生成概率与词汇生成概率,而实际上训练库里的大部分图像都与该测试图像无关,利用这些大量无关的训练图像进行概率估计不仅会影响最后的标注结果,而且也会影响标注的速度,所以如何从整个训练图像库中选取与某测试图像相关的那部分图像进行训练是也是本文的一个重点.

3.1　层叠标注模型

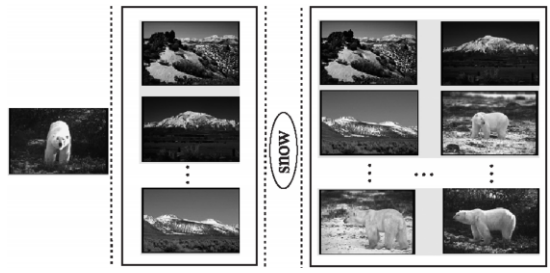
针对上述问题,本文提出基于判别模型与生成模型的层叠图像自动标注模型(Hierarchically Discriminative and Generative Annotation Model, HDGAM) ,其主要思想是给定一幅待标注图像,利用 K-means 算法对所有训练图像进行聚类,然后利用判别式模型判断这幅待标注图像所属的类别,取类别概率较高的几个聚类中的图像作为该待标注图像的相关图像集合,最后利用生成式模型与待标注图像的相关图像集合对该待标注图像进行标注.

SVM 作为判别分类模型的代表,已经广泛应用于文本分类与图像分类等诸多领域,并取得很好的效果. 由于 SVM 在分类问题上的出色性能,本文的判别式模型选用 SVM. 在本文的层叠自动标注模型中,SVM 主要用于图像分类,即将待标注的测试图像分配到应属的聚类中,这个步骤对应本文层叠标注模型中的第一层标注,即主题标注,主题标注并不针对待标注图像的所有内容,而是给待标注图像整体标注若干个主题. 需要注意的是,这边的主题是对所有训练图像进行聚类得到的,即每个聚类下所有图像的视觉内容都应该与相对应的主题保持一致. 也就是说,每个主题下的图像应具有一致的视觉特征分布,而某个主题只是针对同一视觉特征分布的一个抽象概念,即有些主题概念并不具有实际意义.

3.2 相关图像扩展

通过后面实验可发现,使用上节提出的层叠标注模型,直接利用几个最相似类别里的图像进行生成概率计算已经可以提升标注结果,但是提高的幅度并不明显. 主要原因是受限于现有聚类算法与分类算法本身的局限性,即在聚类过程中,很多主题里都会存在一些视觉上并不相关的图像. 在主题标注过程中,一些待标注的图像没有被正确分到应属的主题中. 这些聚类错误与分类错误会影响着第一层的标注结果,并会将错误传播到下一层标注,进而影响最终的标注结果. 所以必须考虑如何减小聚类以及分类错误对最终标注结果带来的影响.

现有的模型都是直接面向图像(Images-Oriented),主要是考虑图像到关键词之间的映射(Images→Keywords),而这种映射条件下一般都无法直接建立图像与关键词之间的联系. 所以本文提出以面向关键词(Keywords-Oriented)的方法来建立图像与关键词之间的关系,重点考虑关键词到图像之间的映射(Keywords→Images). 面向关键词的方法将每个关键词所对应的图像组成一个集合,即同一个集合的图像都至少包含某个相同的关键词,这样图像集合的数目就由标注词汇数量所决定. 由于同一个集合里的图像都至少共享一个语义概念,而现实生活中出现在同一场景的不同目标物体之间必然有着某种客观的联系,所以通过这个共同的语义目标建立起的同一集合下不同图像之间必然有着直接或间接的隐含联系,这样通过面向关键词的方法就可以构建起相似场景下关键词与图像之间的联系. 比如,"飞机"与"鸟"可以通过"天空"建立起联系,而"飞机"和"鱼"则很难通过某个语义概念建立起联系. 通过图像与关键词之间直接或间接的联系可以扩展相关图像集合. 图 1 是扩展相关图像集合的一个例子,假设出现一种极端情况,(a)的这幅待标注图像由于聚类与分类错误被分配到了如集合(b)所示的某类风景图像集合中,此时如果直接采用我们提出的层叠标注模型,由于错误的主题划分,利用相关图像集合(b)进行视觉生成概率与词汇生成概率的计算将会使得最终的标注结果无法出现如"bear"的正确标注. 而如果通过对集合(b)中存在的标注词"snow"进行扩展,可以得到相关图像集合(c),从集合(c)可以发现,扩展后的相关图像集合里出现了带有"bear"的训练图像,通过利用扩展后的相关图像集合计算未标注图像的生成概率,将很可能正确地标注出"bear"这个关键词.



(a) 待标注图像 (b) 相关图像集合 (c) 扩展后的相关图像集合

(a) Unlabeled image

(b) Relevant image set

(c) Extended relevant image set

图 1 相关图像集合扩展

Fig. 1 Expansion of relevant image sets

根据上面的分析,通过我们提出的面向关键词的方法就可以构建起相似场景下关键词与图像之间的联系. 同时,对相关图像进行扩展时,必须要设计一个迭代算法来选择最应该加入相关图像集合的那些图像. 在本文提出的层叠图像标注模型中,对未标注图像进行主题标注后得到的相关图像集合里每幅训练图像的权值:

$$w_{p_i}^{(0)} = \frac{1}{n_p^{(0)}},\ i = 1,2,\cdots,n_p^{(0)},$$

其中,p_i 为初始相关图像集合里的第 i 幅训练图像,$w_{p_i}^{(0)}$ 为第 i 幅相关训练图像的权值,$n_p^{(0)}$ 为初始相关图像数量,这里的(0) 表示第 0 次迭代的值,也就是初始值. 第 $t+1$ 次迭代时相关图像集合里每个标注词所对应的权值如下:

$$\widetilde{w}_{l_j}^{(t+1)} = \begin{cases} \mu \cdot \sum_{i=1}^{n_p^{(t)}} \zeta_{l_j} \cdot w_{p_i}^{(t)} \cdot \frac{1}{\#(p_i)}, & l_j \in Cor\_al(l^{(t)}),\ j = 1,2,\cdots,n_l^{(t+1)} \\ v \cdot \sum_{i=1}^{n_p^{(t)}} \zeta_{l_j} \cdot w_{p_i}^{(t)} \cdot \frac{1}{\#(p_i)}, & l_j \notin Cor\_al(l^{(t)}),\ j = 1,2,\cdots,n_l^{(t+1)} \end{cases}$$

其中,l_j 为相关图像集合里的第 j 个标注词,这里的标注词包括相关图像集合里所有图像对应的标注词,$\widetilde{w}_{l_j}^{(t+1)}$ 为第 $t+1$ 次迭代时相关图像集合里第 j 个标注词的权值,$\#(p_i)$ 为相关图像集合里第 i 幅图像包含的标注词个数,$n_l^{(t+1)}$ 为第 $t+1$ 次迭代时标注词的个数,$n_p^{(t)}$ 为第 t 次迭代时相关图像的数目. $Cor\_al(l^{(t)})$ 为第 t 次迭代时的标注词集合,当 l_j 属于第 t 次迭代的标注词集合时,对 l_j 赋予权值 μ; 当 l_j

不属于第 t 次迭代的标注词集合,即 l_j 是在第 $t+1$ 次迭代时产生的,对 l_j 赋予权值 $v.\mu$ 大于 v,这是为了避免随着迭代次数的增加,一些无关的标注词占有较高的权重. ζ_{l_j} 为二值函数,用于判断标注词 l_j 是否属于相关图像 p_i 的标注:

$$\zeta_{l_j} = \begin{cases} 1, \text{ if } l_j \in Anno(p_i) \\ 0, \text{ else} \end{cases}$$

其中 $Anno(p_i)$ 为相关图像 p_i 所对应的标注词.

对 $\widetilde{w}_{l_j}^{(t+1)}$ 进行规范化,得

$$w_{l_j}^{(t+1)} = \frac{\widetilde{w}_{l_j}^{(t+1)}}{\sum_{j=1}^{n_l^{(t+1)}} \widetilde{w}_{l_j}^{(t+1)}}. \quad (1)$$

第 $t+1$ 次迭代时每幅相关图像的权值:

$$w_{p_i}^{(t+1)} = \sum_{j=1}^{n_l^{(t+1)}} \tau_{p_i} \cdot w_{l_j}^{(t+1)} \cdot \frac{1}{\#(l_j)}, \quad (2)$$

$$p_i \in Cor\_il(p^{(t+1)}),\ i = 1,2,\cdots,n_p^{(t+1)},$$

其中,$w_{p_i}^{(t+1)}$ 为第 $t+1$ 次迭代时相关图像 p_i 的权值,$\#(l_j)$ 为所有训练图像中包含标注词 l_j 的图像数目. τ_{p_i} 为二值函数,用于判断相关图像 p_i 是否包含标注词 l_j:

$$\tau_{p_i} = \begin{cases} 1, \text{ if } l_j \in Anno(p_i) \\ 0, \text{ else} \end{cases}$$

$w_{l_j}^{(t+1)}$ 与 $w_{p_i}^{(t+1)}$ 满足

$$\sum_{j=1}^{n_l^{(t+1)}} w_{l_j}^{(t+1)} = 1,\ \sum_{i=1}^{n_p^{(t+1)}} w_{p_i}^{(t+1)} = 1.$$

扩展相关图像集合迭代算法的终止条件:

$$\sum_{m=1}^{K} \widehat{w}_{p_m}^{(t+1)} - \sum_{m=1}^{K} \widehat{w}_{p_m}^{(t)} > \psi, \text{ and } t \geqslant \phi, \quad (3)$$

其中,$\widehat{w}_{p_m}$ 为按降序排列的每幅相关图像的权值,K 为最终选定的相关图像集合大小,ψ 为连续 2 次迭代间 Top(K) 幅相关图像权值和之差的阈值,ϕ 为迭代次数的阈值. 算法 1 为本文提出的扩展相关图像的迭代算法:

算法 1

输入　$Cor\_il(p^{(0)})$ 为层叠标注模型对未标注图像进行主题标注得到的相关图像集合,$C(p_i)$ 为每幅训练图像 p_i 对应的标注词集合,$S(l_j)$ 为每个标注词 l_j 对应的训练图像集合,U 为未标注图像集合; ψ,ϕ 为阈值参数

输出　U^{Cor} 为扩展后所有未标注图像对应的相关图像集合

step 1　For $n = 1$ to $|U|$

step 1.1　利用式(1) 计算相关图像集合里每个标注词的权值;

step 1.2　利用式(2) 计算相关图像集合里每幅图像的权值;

step 1.3　If(满足式(3) 迭代终止条件)

step 1.3.1　对得到的相关图像进行按照权值降序排列;

step 1.3.2　得到第 n 幅未标注图像的相关图像集合;

step 1.3.3　n ++, Go to step 1;

step 1.4　Else

step 1.4.1　t ++, go to step 1.1.

4　HDGAM 图像自动标注模型

本节将重点介绍本文提出的基于判别模型与生成模型的层叠图像自动标注模型(Hierarchically Discriminative and Generative Annotation Model, HDGAM) 的详细标注部分,即第二层标注. HDGAM 模型的详细标注部分基于 MBRM 模型,MBRM 模型属于生成式模型,在图像自动标注领域已被证明为一个有着良好标注性能的模型. 我们对其进行改进,利用待标注图像的相关图像集合来进行视觉生成概率与词汇生成概率的计算,避免使用大量不相关图像进行训练所产生的问题,提升标注准确度,同时降低模型的计算复杂度.

4.1　基于生成模型的图像详细标注

将每幅图像 I 分割为若干个互不重叠的区域集合,$D_I = \{d_1,\cdots,d_{|Z|}\}$,这里采用同 MBRM 模型相同的分块策略,$|Z|$ 为区域的个数. 对每个区域 d_i 提取 m 维的特征向量 F^i,定义图像区域的视觉生成概率为 $P_F(\sim|I)$. 词汇生成概率采用多伯努利分布,假设标注词集合 W_I 是从 $|V|$ 个多伯努利分布 $P_V(\sim|I)$ 独立采样的结果,其中 $|V|$ 为标注词个数. 一幅图像 I 就可以由图像视觉生成概率与词汇生成概率这两部分独立的条件分布构成.

假设图像 T 为一幅未标注图像,其特征向量表示为 $F_T = \{F_T^1,\cdots,F_T^{|Z|}\}$,其中 F_T^i 为图像 T 中第 i 个区域的特征向量. W_L 为标注词集合 $|V|$ 的一个子集. 对图像 T 的视觉表示与词汇表示的联合概率进行建模,记为 $P(F_T,W_L)$. 假设联合概率 $P(F_T,W_L)$ 中,F_T 和 W_L 的隐含关系与图像 T 的相关图像集合里某幅图像的视觉特征与标注词的隐含关系相似,而我们并不清楚这个具体的隐含关系,所以我们对图像 T 的相关图像集合里每一幅图像都计算其与图像 T

联合概率 $P(F_T, W_L)$ 的期望. 联合生成 F_T 与 W_L 的过程如下.

1) 按照3.2节提出的迭代算法获得未标注图像 T 扩展后的相关图像集合.

2) 按照概率 $P_\Gamma(I)$ 从相关图像集合 Γ 里选取训练图像 I.

3) 对 $i = 1, 2, \cdots, |Z|$ ($|Z|$为图像区域个数): 按照条件概率密度函数 $P_F(\sim | I)$ 生成第 i 个区域的视觉描述.

4) 对每个标注词: 按照多伯努利分布 $P_V(\sim | I)$ 生成标注词集合 W_I.

这里每个图像区域与标注词之间并不存在一一对应的关系,目的只是要找出对于整幅图像最适合的若干个标注词. 根据上面的概率生成过程,本文提出的 HDGAM 模型中详细标注部分关于联合产生图像视觉生成概率与标注词生成概率的公式如下:

$$P(F_T, W_L) = \sum_{I \in \Gamma} \{ P_\Gamma(I) \times \prod_{i=1}^{|Z|} P_F(F_T^i | I) \times \prod_{w \in W_T} P_V(v | I) \times \prod_{w \notin W_T} (1 - P_V(v | I)) \}. \quad (4)$$

HDGAM 模型利用上式对图像进行详细标注. 在详细标注部分,本文的 HDGAM 模型只是利用未标注图像的相关图像集合来计算视觉生成概率与词汇生成概率,而传统的模型需要对训练集中的每一幅图像都计算其视觉特征与词汇的联合概率,所以本文的 HDGAM 模型可以去除大量不相关图像对标注结果的干扰,提高标注速度.

4.2 参数估计

本节主要针对式(4)的参数进行估计. $P_\Gamma(I)$ 为某一幅图像 I 在待标注图像 T 的相关图像集合里出现的概率,假设 $P_\Gamma(I)$ 服从均匀分布,即 $P_\Gamma(I) = 1/|\Gamma|$,其中 $|\Gamma|$ 为相关图像集合大小.

条件概率密度函数 $P_F(\sim | I)$ 用来估计图像区域的视觉生成概率. 对 $P_F(\sim | I)$ 的分布使用非参数高斯核密度函数进行估计. 假定 $F_I = \{F_I^1, \cdots, F_I^{|Z|}\}$ 为图像 I 每个区域的特征,$P_F(\sim | I)$ 的估计公式如下:

$$P_F(F_T^i | I) = \frac{1}{|Z|} \sum_{j=1}^{|Z|} \frac{\exp\{ -(F_T^i - F_I^j)^{\mathrm{T}} \Sigma^{-1} (F_T^i - F_I^j) \}}{\sqrt{2^m \pi^m |\Sigma|}}, \quad (5)$$

其中,$|Z|$ 为图像区域个数,m 为特征维数. 式(5)对图像 I 的每个区域 F_I^j 都采用高斯核函数进行估计. 高斯核的参数由协方差矩阵 Σ 来确定,$\Sigma = \alpha \cdot \Lambda$,其中 α 为高斯核的宽度,确定 P_F 在 F_I^i 附近的平滑程度,Λ 为单位矩阵.

$P_V(v | I)$ 是多伯努利分布的第 v 个元素,为待标注图像 T 的相关图像集合里某幅图像 I 产生标注集合 W_I 的概率. 对每个词采用贝叶斯估计:

$$P_V(v | I) = \frac{\delta \cdot \eta_{v,I} + N_v}{\delta + |\Gamma|},$$

其中,N_v 是标注词 v 在相关图像标注中出现的次数,$|\Gamma|$ 为相关图像集合的大小,$\eta_{v,I}$ 是一个二值函数,当关键词 v 属于图像 I 的标注时,$\eta_{v,I} = 1$,否则 $\eta_{v,I} = 0$. δ 为平滑参数,也可看作是 $\eta_{v,I}$ 的权重.

4.3 HDGAM 模型框架

本节将介绍如何利用文中提出的 HDGAM 标注模型进行图像自动标注,模型框架图如图2所示. 首先利用 K-Means 算法将所有训练图像按照视觉内容进行聚类,得到带主题类别的训练图像集合,即每幅训练图像都隶属于某个主题类别. 利用判别模型 SVM 对所有主题类别里的图像进行训练,得到训练好的判别模型分类器. 给定一幅待标注图像,利用判别模型分类器对该图像进行多类分类,这一步对应于我们层叠标注模型中的第1层标注,即主题标注,这是对未标注图像进行弱标注的过程. 利用本文提出的面向关键词的方法构建相似场景下图像与关键词之间的联系,并使用本文提出的迭代算法对语义关键词以及相关图像进行扩展,得到扩展后每幅未标注图像对应的 K 幅相关图像集合. 对每一幅未标注的图像,利用生成式模型与 K 幅相关图像计算其最终的标注,这里的最终标注就是对未标注图像进行详细标注的结果,对应于本文层叠标注模型中的第2层标注.

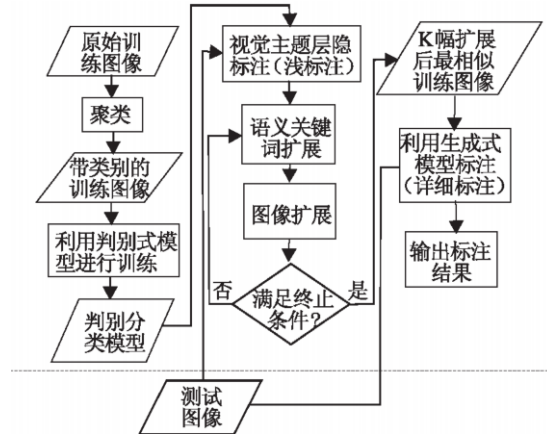


图2 HDGAM 模型框架图

Fig. 2 Framework of HDGAM model

5　实　验

5.1　实验设置

为验证文中 HDGAM 模型的有效性，并同其它模型进行公平比较，本文采用图像自动标注中普遍使用的 Corel 5K 数据集. 这个图像库共 5 000 张图片，每幅图像有 1 ~ 5 个词作为其标注，词汇总数为 374. 本文将数据集分为 3 个部分: 训练库共4 000幅图像，验证库共 500 幅图像，测试库共 500 幅图像. 其中，验证图像库主要用于模型参数的确定，等参数确定后，将验证图像库全部加入训练库中形成新的训练图像库. 其中，阈值 ψ 取 0.05，阈值 ϕ 取 20. 这样就与其它模型采用 4 500 幅训练图像，500 幅测试图像相一致，每幅图像固定返回 5 个标注词.

本文的特征提取分为 2 部分: 第 1 部分提取整幅图像的特征用于第 1 层标注，由于这部分特征是针对整幅图像进行提取，并非提取图像的分块特征，所以采用的特征维数会相对较多，具体包括 72 维的 HSV 直方图特征、9 维的 RGB 空间颜色矩、32 维的灰度共生矩阵特征、36 维的 Gabor 纹理特征，7 维的 Hu 不变矩特征，共计 156 维的特征; 第 2 部分提取的是图像的分块特征，用于第 2 层标注，即详细标注部分，为方便比较，本文采用与 MBRM 相同的 30 维特征，具体包括 9 维的 RGB 空间颜色矩，9 维的 Lab 空间颜色矩，12 维的 Gabor 纹理特征，包括 3 个尺度与 4 个方向. 本文模型中第 1 层视觉主题层标注使用 LIBSVM [19].

同其它图像标注模型一样，我们使用查准率、查全率与 F 度量来验证标注结果. 假设 w 为某个关键词，N_c 为正确标注的图像数，N_s 为检索返回的图像数，N_t 为测试图像库中包含标注词 w 的图像数，则

$$Precision(w) = \frac{N_c}{N_s},\ Recall(w) = \frac{N_c}{N_t},$$

$$F(w) = \frac{2 \times Precision(w) \times Recall(w)}{Precision(w) + Recall(w)}.$$

对所有出现在测试库中的关键词分别计算，最后把得到每个词的查准率、查全率以及 F 度量取平均作为最终的评价指标. 此外，我们还统计了至少被正确标注一次的关键词数量，记作 *"NZR"*. 它反映了模型对标注词的覆盖程度，这也是一个很重要的标注性能评价指标.

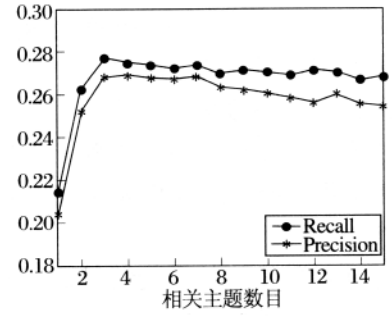
5.2　实验对比

5.2.1　确定相关主题数

这部分实验主要用于确定第 1 层标注中未标注图像所属的主题数，并没有使用我们提出的迭代扩展相关图像集的方法. 聚类数取 50，即共有 50 个主题类别. 实验结果如图 3 所示. 从图 3(a) 可以发现，相关主题数目为 3 时，标注结果最好. 当测试图像使用更多主题类别下的图像作为相关图像时，标注结果的查准率和查全率并没有提高，反而会下降一些，这是因为随着主题类别的增加，相关图像集合里的不相关图像也相应增加. 在后面的实验中，对于每 1 幅测试图像，第 1 层主题标注时都取 3 个最相关主题中的图像构成相关图像集合.

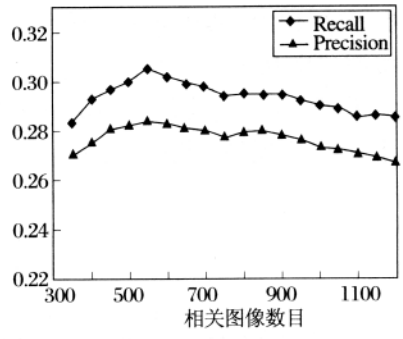
5.2.2　确定扩展后的相关图像数目

这部分实验主要用于验证我们提出的迭代扩展相关图像算法的有效性以及确定扩展后的相关图像集合大小. 实验结果如图 3(b) 所示，从(b) 可以发现，当最扩展后的相关图像集合大小取 550 时可以达到最好的标注效果. 随着相关图像数目的增加，查准率和查全率也相应地逐渐下降，这是因为不相关图像也随之增加，与上一小节的结论类似.



(a) 相关主题数目

(a) Relevant topic number



(b) 相关图像数目

(b) Relevant image number

图 3　2 个参数对标注结果的影响

Fig. 3　Influence of relevant topic and image numbers on annotation results

5.2.3 与其它模型的对比

本节将我们的 HDGAM 标注模型同其它经典的标注模型相对比，包括 TM、CMRM、MBRM、CLP、SML，实验结果见表 1.

表 1 各模型性能对比

Table 1 Performance comparison of different models

模型	Precision	Recall	F-measure	NZR
TM	0.06	0.04	0.05	49
CMRM	0.10	0.09	0.09	66
MBRM	0.24	0.25	0.24	122
CLP	0.21	0.25	0.22	125
SML	0.23	0.29	0.26	137
HDGAM	0.28	0.30	0.29	143

从表 1 可以看出，我们提出的基于判别模型与生成模型的层叠图像自动标注模型(HDGAM)是有效的，标注结果要好于现在流行的几种模型. 查准率达到 0.28，比前面模型中查准率最高的 MBRM 模型高出 17%；查全率高达 0.30，比前面模型中查全率最高的 SML 模型要高出约 3%；F 度量达到 0.29，比前面模型中 F 度量最高的 SML 高出约 12%. 此外，在衡量标注词覆盖程度的“NZR”这个指标上，我们的模型中有 143 个词至少被正确标注过一次，比前面模型中该指标最高的 SML 模型高出 4%.

5.2.4 结果分析

为更好地对各模型的实验结果进行比较分析，我们绘制了更为直观的柱状图，如图 4 所示. 从图 4 可以更容易看出，我们的 HDGAM 模型在各方面指标上都比现有的一些模型要好.

表 2 是 4 幅示例图的自动标注结果，将我们的 HDGAM 模型与 MBRM 模型进行比较，每个标注词的顺序是按照概率从大到小排列. 通过表 2 中 4 幅图的标注结果可以看出，我们的 HDGAM 模型标注效果要好于 MBRM 模型. 此外，还可以发现一些图像的标注结果虽然与数据集的原始标注不同，但是这些词也可以描述图像内容，或者这些词本来就反映图像的视觉内容. 只是在构建图像库的标注时，被一些标注者所忽略. 如第一幅图像标注结果中的“water”与“sand”并不属于原始标注，但是这两个词可以用来描述第一幅图像的内容. 第四幅图像的“coral”与“sea”也有类似情况.

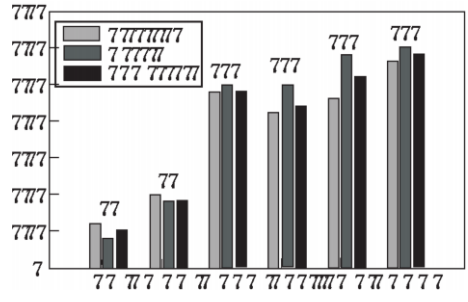


图 4 各模型标注结果对比

Fig. 4 Performance comparisons of different models

5.3 模型时间复杂度分析

使用 K-Means 算法进行聚类的所需要的时间为 $T(n) = K \cdot t \cdot n$，其中 K 为聚类的个数，t 为迭代的次数，n 为训练图像数. 由于 $K/n, t/n \leqslant 0.01$，所以 K 与 t 要远小于 n，故 K-Means 算法的渐进时间复杂度为 $O(n)$. 利用 SVM 将第一层标注模型训练好之后，对视觉主题层标注进行预测的时间复杂度为 $O(n)$. 第二层标注中，利用生成模型对图像进行详细标注时，由于只用了相关图像集合里的训练图像，所以详细标注时所需的时间为 $T(n) = l \cdot n'$，其中 l 为标注词总数，n' 为相关图像数量，详细标注部分时间复杂度不超过 $O(n)$. 因此，我们的 HDGAM 模型时间复杂度为 $O(n)$，与单独使用判别模型或生成模型进行图像标注的时间复杂度相同，这就说明我们的模型在提高标注准确率的同时并没有增加时间复杂度.

表 2 2 种模型的标注结果对比

Table 2 Comparisons of annotation results between 2 models

				
原始标注	beach palm people tree	birds nest	clouds mountain sky water	fish grouper ocean people
MBRM	clouds stone tree ruins sky	flowers plants leaf grass birds	sky grass wall palace water	sky tree mountain snow people
HDGAM	water people sand beach tree	grass birds nest baby leaf	water bridge sky tree clouds	people ocean coral sea fan

6　结 束 语

现有图像自动标注模型在计算图像与词汇联合概率时,几乎都采用所有训练图像进行生成概率估计,这就会给未标注图像带来大量不相关图像,一方面影响标注准确率,另一方面增加计算时间. 针对上述问题,本文提出了基于判别模型与生成模型的层叠图像自动标注模型 HDGAM,该模型结合了判别模型与生成模型的优点,分 2 层对图像进行自动标注. 在第 1 层标注中,利用判别模型对未标注图像进行主题标注,并获得未标注图像的相关图像集. 在第 2 层标注中,利用我们提出的面向关键词的方法建立图像与关键词之间的联系,并使用我们提出的迭代算法对相关图像进行扩展. 最终利用生成式模型与扩展后的相关图像集合对未标注图像进行详细标注. 实验结果表明,本文提出的 HDGAM 标注模型在查准率、查全率等指标上相比以往模型均有所提高. 不仅如此,通过分析计算复杂性可知,我们的层叠标注模型时间复杂度与单独使用判别模型或生成模型进行图像标注的时间复杂度相同,并没有增加时间复杂度. 下一步的研究工作将考虑如何利用基于内容的图像自动标注方法对大规模的网络图像进行标注,以及将图像自动标注应用到图像检索中.

参 考 文 献

[1] Zhao Yufeng, Zhao Yao, Zhu Zhenfeng. TSVM-HMM: Transductive SVM Based Hidden Markov Model for Automatic Image Annotation. Expert Systems with Applications: An International Journal, 2009, 36(6): 9813 - 9818

[2] Qi Xiaojun, Han Yutao. Incorporating Multiple SVMs for Automatic Image Annotation. Pattern Recognition, 2007, 40(2): 728 - 741

[3] Barrat S, Tabbone S. Modeling, Classifying and Annotating Weakly Annotated Images Using Bayesian Network // Proc of the 10th International Conference on Document Analysis and Recognition. Barcelona, Spain, 2009: 1201 - 1205

[4] Andriluka M, Roth S, Schiele B. People-Tracking-By-Detection and People-Detection-By-Tracking // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Alaska, USA, 2008: 1 - 8

[5] Mutch J, Loue D G. Object Class Recognition and Localization Using Sparse Features with Limited Receptive Fields. International Journal of Computer Vision. 2008, 80(1): 45 - 57

[6] Kalanit G, Rory S. Object Recognition: Insights from Advances in fMRI Methods. Current Directions in Psychological Science, 2009, 17(2): 73 - 79

[7] Dollar P, Wojek C, Schiele B, *et al*. Pedestrian Detection: A Benchmark // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 304 - 311

[8] Duygulu P, Barnard K, de Freitas J F G, *et al*. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary // Proc of the 7th European Conference on Computer Vision. Copenhagen, Denmark, 2002: 97 - 112

[9] Shi Jianbo, Malik J. Normalized Cuts and Image Segmentation. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888 - 905

[10] Blei D M. Ng A Y, Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, 3(5): 993 - 1022

[11] Jeon J, Lavrenko V, Manmatha R. Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models // Proc of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada, 2003: 119 - 126

[12] Lavrenko V, Manmatha R, Jeon J. A Model for Learning the Semantics of Pictures [EB/OL]. [2010-02-15]. http://citeseerx.ist.psu.edu/viewdoc/summary? doi - 10.1.1.70.3629&rank = 1

[13] Feng S L, Manmatha R, Lavrenko V. Multiple Bernoulli Relevance Models for Image and Video Annotation // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, USA, 2004: 1002 - 1009

[14] Kang Feng, Jin Rong, Sukthankar R. Correlated Label Propagation with Application to Multi-Label Learning // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA, 2006: 1719 - 1726

[15] Liu Jing, Li Mingjing, Ma Weiying, *et al*. An Adaptive Graph Model for Automatic Image Annotation // Proc of the 8th ACM International Workshop on Multimedia Information Retrieval. Santa Barbara, USA, 2006: 61 - 69

[16] Gustavo C, Antoni B C, Pedro J M, *et al*. Supervised Learning of Semantic Classes for Image Annotation and Retrieval. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(3): 394 - 410

[17] Wong Yong, Mei Tao, Gong Shaogang, *et al*. Combining Global, Regional and Contextual Features for Automatic Image Annotation. Pattern Recognition, 2009, 42(2): 259 - 266

[18] Lindstaedt S, Nörzinger R, Sorschag R, *et al*. Automatic Image Annotation Using Visual Content and Folksonomies. Multimedia Tools and Applications, 2009, 42(1): 97 - 113

[19] Chang C C, Lin C J. LIBSVM: A Library for Support Vector Machines [EB/OL]. [2010-03-01]. http://www.csie.ntu.edu.tw/~cjlin/libsvm