软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(4):801−812 [doi: 10.3724/SP.J.1001.2011.03742]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

融合语义主题的图像自动标注*

李志欣[1,2+], 施智平[1], 李志清[1,2], 史忠植[1]

[1](中国科学院 计算技术研究所 智能信息处理重点实验室,北京 100190)

[2](中国科学院 研究生院,北京 100049)

Automatic Image Annotation by Fusing Semantic Topics

LI Zhi-Xin[1,2+], SHI Zhi-Ping[1], LI Zhi-Qing[1,2], SHI Zhong-Zhi[1]

[1](Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: E-mail: lizx@ics.ict.ac.cn, http://www.intsci.ac.cn

Li ZX, Shi ZP, Li ZQ, Shi ZZ. Automatic image annotation by fusing semantic topics. *Journal of Software*, 2011,22(4):801−812. http://www.jos.org.cn/1000-9825/3742.htm

Abstract: Automatic image annotation has become an important issue, due to the existence of a semantic gap. Based on probabilistic latent semantic analysis (PLSA), this paper presents an approach to annotate and retrieve images by fusing semantic topics. First, in order to precisely model training data, each image is represented as a bag of visual words. Then, a probabilistic model is designed to capture latent semantic topics from visual and textual modalities, respectively. Furthermore, an adaptive asymmetric learning approach is proposed to fuse these semantic topics. For each image document, the topic distribution of each modality is fused by multiplying different weights, which is determined by the entropy of the distribution of visual words. Consequently, the probabilistic model can predict semantic annotations for an unseen image because it associates visual and textual modalities properly. This approach is compared with several other state-of-the-art approaches on a standard Corel dataset. The experimental results show that this approach performs more effectively and accurately.

Key words: automatic image annotation; topic model; probabilistic latent semantic analysis; adaptive asymmetric learning; image retrieval

摘 要: 由于语义鸿沟的存在,图像自动标注已成为一个重要课题.在概率潜语义分析的基础上,提出了一种融合语义主题的方法以进行图像的标注和检索.首先,为了更准确地建模训练数据,将每幅图像的视觉特征表示为一个视觉“词袋”;然后设计一个概率模型分别从视觉模态和文本模态中捕获潜在语义主题,并提出一种自适应的不对称学习方法融合两种语义主题.对于每个图像文档,它在各个模态上的主题分布通过加权进行融合,而权值由该文档的视觉词分布的熵值来确定.于是,融合之后的概率模型适当地关联了视觉模态和文本模态的信息,因此能够很好地预测未知图像的语义标注.在一个通用的 Corel 图像数据集上,将提出的方法与几种前沿的图像标注方法进行了比较.实验结果表明,该方法具有更好的标注和检索性能.

* 基金项目: 国家自然科学基金(60933004, 60903141, 60805041); 国家重点基础研究发展计划(973)(2007CB311004)
收稿时间: 2009-05-13; 定稿时间: 2009-10-10

关键词: 图像自动标注;主题模型;概率潜语义分析;自适应不对称学习;图像检索
中图法分类号: TP391 文献标识码: A

随着数字成像、数据存储和互联网等技术的发展,对大规模图像库进行有效的组织、索引和检索成为近年来的重要课题.图像检索技术经过十几年的发展,形成了两种主流解决方案:基于文本的图像检索和基于内容的图像检索.基于文本的图像检索利用人工对图像进行标注,并在此基础上,利用传统的文本搜索引擎查询图像,这种查询方式比较直观而且能够取得较好的效果.但是,由于人工标注费时费力,使得这种检索方案不能推广到大的图像数据库.基于内容的图像检索采用特征提取和高维索引技术进行图像检索,它为每幅图像提取若干低层视觉特征,以高维向量的形式存入数据库,通过比较这些特征的相似度来获得检索结果.这种方案在某些特殊领域得到了很好的应用,但由于存在语义鸿沟[1],视觉特征相似的图像很可能在语义上是不相关的.为了获得语义相关的检索结果,同时避免大量的手工标注,图像自动标注成为目前关键的具有挑战性的课题[2,3].

主题模型(topic model)又称为层面模型(aspect model),最初主要应用于文本分类和信息检索等领域,近年来在计算机视觉领域也得到了广泛的应用.具有代表性的主题模型有概率潜语义分析(probabilistic latent semantic analysis,简称 PLSA)[4]模型和潜在狄里克雷分布(latent Dirichlet allocation,简称 LDA)[5]模型,它们不仅在场景分类、对象识别等领域取得了良好的效果,而且也成功地应用于图像的自动标注和检索[6−8].

本文提出一种基于 PLSA 的图像标注方法,它通过两个 PLSA 模型分别获取训练图像的视觉模态和文本模态的信息,将每幅图像表示为两种不同模态的主题分布,这两种主题分布通过一种自适应的不对称学习算法进行融合,从而形成新的潜在空间.这个潜在空间能够较好地关联图像的视觉特征和文本关键词,所以本方法能够较为准确地预测未知图像的语义信息.通过在一个包含 5 000 幅图像的 Corel 数据集上进行的测试和比较,实验结果表明,与若干前沿的图像标注方法相比,本方法具有更好的标注质量和检索性能.

1 相关工作

图像自动标注的主要目标是,确定图像从属于元数据给定的某个语义概念的概率,从而为图像的语义检索奠定基础.自动概念检测和语言索引等工作本质上都与此目标相关.当前,图像自动标注的方法大致可以分为两类:有监督分类的方法和关联建模的方法.

有监督分类的方法将各个语义类别(一个关键词或关键词集合)看作独立的概念,为每个语义类别建立各不相同的分类器.给定一个未知图像,可以通过比较在视觉层次的特征相似度,将相应的关键词传播给新图像.具有代表性的工作是 Li 等人[9]提出的图像自动语义索引系统(automatic linguistic indexing of pictures,简称 ALIP)和 Chang 等人[10]提出的基于内容的软标注系统(content-based soft annotation,简称 CBSA).ALIP 使用一个二维多分辨率隐马尔可夫模型捕获给定语义类别的图像特征之间及其内部的空间依赖关系,各个语义类别的模型是独立学习和分别存储的.标注方法是计算查询图像与各个语义类别之间的相似度,然后选择最相似的类别所包含的关键词进行标注.而 CBSA 首先选择一个训练图像集对全体分类器进行训练,其中,每幅图像具有一个标注(如森林、动物、天空等);然后,将全体分类器应用到一幅给定的图像上以获取图像的多个软标注.Carneiro 等人[11]采用基于最小错误率的优化准则和统计分类的思想,提出一种监督多类标注方法(supervised multiclass labeling,简称 SML).其基本思想是,将每一个语义概念定义为一个语义类别,引进一个随机变量 W,其取值范围为 $\{1,\ldots,T\}$,使得当且仅当样本 x 具有语义概念 w_i 时 $W=i$(这里,$i\in\{1,\ldots,T\}$).同时,引进条件概率密度 $P_{X|W}(x|i)$作为给定语义类别的低层特征分布,然后利用贝叶斯决策规则推导具有最小错误率的 W 的状态.SML 在训练分类器阶段为每幅图像提取一个特征集,利用多示例学习算法从多幅图像的特征集中学习语义概念,从而为每个语义概念建立概率模型.于是在标注阶段,SML 通过训练好的各个分类器的竞争标注机制来推导图像所具有的多个语义概念,同时,根据后验概率产生语义标注的自然排序,便于实现图像的语义检索.

从文本领域的研究得到启发,许多图像自动标注方法通过建立关联模型的方法来标注图像.这类方法利用现有的已标注好的图像数据集,试图在无监督的基础上学习图像的视觉特征和文本关键词之间的关联,然后将

这种关联应用于未标注的图像,通过统计推理的方法来预测图像的语义信息.一个较早的工作是 Duygulu 等人[12]提出的机器翻译模型(translation model,简称 TM),该方法将图像分割为任意形状的区域,这些区域大致对应于一个对象或对象的一部分,然后依据区域特征将图像区域聚类为 blob.随之而来的一个自然的假设是:图像的 blob 和某个关键词之间存在某种隐含的一一对应关系.借助机器翻译的概念,TM 将 blob 和关键词看作是两种对等的“语言”,于是,标注可以看作是一个将 blob 翻译为关键词的过程.随后,Barnard 等人[13]讨论了几个用来表示 blob 和关键词的联合分布的概率模型,包括分层聚类模型、翻译模型和多模态 LDA 混合模型,并考虑对整幅图像的标注问题和对图像区域的命名问题.一旦通过学习得到 blob 和关键词的联合概率分布,图像标注和区域命名问题就转化为图像、blob 和关键词的相关性问题.Blei 等人[6]使用更复杂的关联 LDA 模型对关键词和图像建模,该模型可以看作一个生成式过程:首先生成一系列隐藏变量(潜在主题)用以关联视觉模态和文本模态.于是,一幅图像可以分解为一系列潜在主题的混合;然后,在这些潜在主题中选择一个子集转换为若干基于 LDA 的混合模型,使用高斯分布为图像的区域特征建模,使用多项分布为标注关键词建模,从而在该混合模型的基础上产生图像的语义标注.Jeon 等人[14]提出的跨媒体相关模型(cross-media relevance model,简称 CMRM)也采用分割区域来表示图像,但与翻译模型不同,它并不认为图像的关键词和区域之间是一对一的对应关系,而是通过学习关键词和区域的联合概率分布为整幅图像标注若干关键词.Lavrenko 等人[15]随后提出类似的连续空间相关模型(continuous-space relevance model,简称 CRM).CRM 与 CMRM 有两个重要的区别:(1) CMRM 是一个离散模型,不能利用连续的特征,使用它进行标注需要对连续的特征进行量化得到离散的词汇表,而 CRM 可以对连续的特征建模;(2) CMRM 依赖于对特征向量的聚类,标注质量对聚类错误非常敏感,需要预先选择聚类粒度,而 CRM 不依赖于特征向量的聚类且不受聚类粒度问题的困扰.因此,CRM 获得了比 CMRM 高得多的标注和检索精度.随后,Feng 等人[16]在此基础上提出多贝努里相关模型(multiple Bernoulli relevance model,简称 MBRM),该模型使用多贝努里分布代替 CRM 中的多项分布来估计关键词概率,使用无参核密度函数估计图像区域特征的概率,能够获得更好的标注性能.Monay 等人[7]使用 PLSA 进行建模,提出 PLSA-WORDS 标注方法.该方法也将图像看作是一系列潜在主题的混合,并在潜在主题上分别生成视觉特征和文本关键词的概率分布.但该方法将图像和文本视为两种不对等的模态,采用不对称的学习算法,仅从文本模态的数据学习一个潜在空间,并保持与视觉模态的关联,获得了较好的标注和检索性能.综上所述,关联建模的基本思想是,引入随机变量 L 对客观世界的隐藏状态进行编码,变量 L 的各个状态定义了语义关键词和图像特征的联合分布.不同的标注方法对于隐藏变量的状态作了各不相同的定义:有些方法[14−16]将数据库中的图像与隐藏变量相联系,另一些方法[12,13]则将图像聚类与隐藏变量相联系,还有些方法[6−8]是将主题模型的潜在主题与隐藏变量相联系.

本文提出的标注方法称为 PLSA-FUSION.从概率结构上看,PLSA-FUSION 与 PLSA-WORDS[7]相似.然而,PLSA-FUSION 采用的学习算法与 PLSA-WORDS 非常不同:首先,PLSA-WORDS 仅从文本模态中学习潜在空间,而 PLSA-FUSION 使用两组潜在主题分别从文本模态和视觉模态中学习,然后再融合为一个潜在空间;更为重要的是,PLSA-WORDS 的学习过程相对而言是静态的,而 PLSA-FUSION 采用一种自适应的动态方法进行学习,该方法根据每幅图像各自的视觉词分布确定各不相同的权值对两组主题进行融合.PLSA-FUSION 与 PLSA-WORDS 的区别不仅是概念上的,在实验部分将看到,PLSA-FUSION 的标注和检索性能都超过了 PLSA-WORDS.为保证比较的公平性,这两种方法使用了相同的特征和数据集.值得注意的是,尽管这两种方法都使用离散的视觉特征建模,却能达到连续模型 CRM 的标注精度.

2 PLSA 模型

PLSA[4]和 LDA[5]是目前最常用的两个主题模型.LDA 能够有效地解决 PLSA 存在的过拟合问题,而且在某些文本分类的任务中具有更好的性能.本文选择 PLSA 进行建模有下面几个原因:首先,PLSA 可以使用相对精确的期望最大化(expectation maximization,简称 EM)算法进行参数估计,这使得对 PLSA 的学习过程进行改进不会影响模型的结果分析;其次,PLSA 在对图像进行识别和分类等任务中的表现并不比 LDA 差[17];最后,使用提早停止(early stopping)技术可以有效地控制 PLSA 的过拟合问题.

2.1 原 理

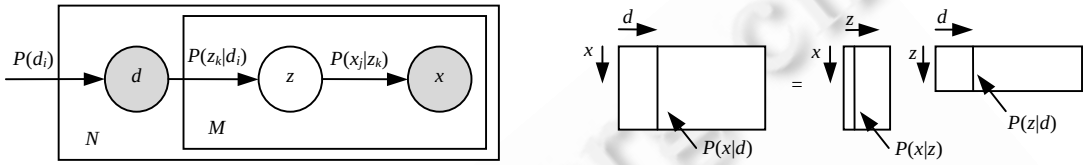
PLSA 在文档 $d_i(i\in1,\ldots,N)$生成其各个组成元素 $x_j(j\in1,\ldots,M)$的过程中引入一个隐含变量(潜在主题)z_k $(k\in1,\ldots,K)$.给定潜在主题 z_k,PLSA 假设每个事件 x_j 独立于其所属文档,于是,对应的联合概率可由下式表示:

$$P(d_i,z_k,x_j)=P(d_i)P(z_k|d_i)P(x_j|z_k) \tag{1}$$

则可观察变量的联合概率可以通过边缘化潜在主题 z_k 得到:

$$P(d_i,x_j)=P(d_i)\sum_{k=1}^{K}P(z_k\mid d_i)P(x_j\mid z_k) \tag{2}$$

PLSA 的图模型表示如图 1(a)所示.使用 PLSA 模型可以将一个文档表示为一个对应于主题分布的 K 维向量,这等价于图 1(b)中所示的矩阵分解.



(a) Graphical model representation of PLSA
(a) PLSA 的图模型表示

(b) Matrix decomposition of conditional distribution
(b) 条件分布的矩阵分解

Fig.1
图 1

PLSA 的模型参数是两个条件概率分布 $P(x|z)$和 $P(z|d)$,这两个参数都满足多项分布.对于一个给定的文档集,$P(x|z)$表示各个主题的特征,它对于文档集之外的文档仍然适用;而 $P(z|d)$仅表示某个特定文档的主题分布,不能给未知图像带来任何先验信息.

2.2 参数估计算法

PLSA 使用 EM 算法估计模型参数,EM 算法通过最大化下面的对数似然函数推导得到:

$$\mathcal{L}=\sum_{i=1}^{N}\sum_{j=1}^{M}n(d_i,x_j)\log P(d_i,x_j) \tag{3}$$

这里,$n(d_i,x_j)$表示文档 d_i 中元素 x_j 的个数.EM 算法的两个步骤表示如下:

E 步. 利用前面估计的模型参数,计算给定观察对(d_i,x_j)时潜在主题 z_k 的条件概率分布:

$$P(z_k\mid d_i,x_j)=\frac{P(z_k\mid d_i)P(x_j\mid z_k)}{\sum_{l=1}^{K}P(z_l\mid d_i)P(x_j\mid z_l)} \tag{4}$$

M 步. 利用新的期望值 $P(z|d,x)$更新参数 $P(x|z)$和 $P(z|d)$:

$$P(x_j\mid z_k)=\frac{\sum_{i=1}^{N}n(d_i,x_j)P(z_k\mid d_i,x_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N}n(d_i,x_m)P(z_k\mid d_i,x_m)} \tag{5}$$

$$P(z_k\mid d_i)=\frac{\sum_{j=1}^{M}n(d_i,x_j)P(z_k\mid d_i,x_j)}{\sum_{j=1}^{M}n(d_i,x_j)} \tag{6}$$

对于参数 $P(x|z)$和 $P(z|d)$,如果已知其中一个分布 $P(x|z)$(或 $P(z|d)$),则另一个分布 $P(z|d)$(或 $P(x|z)$)可以使用 folding-in 算法计算得到.folding-in 算法是 EM 算法的不完全版本,该算法在迭代过程中保持已知参数不变,不断更新未知参数,使得公式(3)中的似然函数最大.

3 数据建模与学习

本节设计的概率模型使用 PLSA 对训练数据进行建模并学习视觉特征和文本关键词之间的关联.经过学

习,给定一幅未知图像,可以利用该模型对它进行标注;给定一个图像集,就可以先对其中所有的图像进行标注,然后实现在该图像集上的语义检索.使用 PLSA 建模首先需要将图像表示为离散化数据,可使用“词袋(bag-of-words,简称BOW)”模型将图像表示为离散的视觉词,并集成多种视觉特征.然后,使用自适应的不对称学习算法可以更好地学习视觉模态和文本模态之间的关联,从而提高图像的标注和检索性能.

3.1 图像表示

为了分析图像的内容,现有的图像自动标注方法使用了各种各样的局部特征,大致可分为 3 类:区域特征[6,12−15]、方块特征[9−11,16]和局部不变特征[7].区域特征首先采用自动分割算法对图像进行分割,然后提取各个分割区域的颜色、纹理等特征;方块特征使用一个固定的网格将图像划分为固定大小的小方块,再分别提取这些小方块的特征;局部不变特征首先检测图像中的局部极值点,再计算这些极值点的描述子.有的方法对局部特征进行量化后再建模[7,12,14],而有的方法则直接使用连续模型建模局部特征[6,9,11,15,16].由于特征在聚类的过程中会损失信息,聚类粒度的选择对标注质量影响很大.所以,连续模型的性能普遍高于离散模型.但是,连续特征的数据量比离散特征大得多,因而实践中使用连续特征难以集成多种局部特征.

由于不同的特征提供了不同类别的信息,对于特定的类别各有其优势;另一方面,在很多情形下,分析图像内容需要结合不同种类的特征,所以集成不同特征对于提高图像标注性能是有益的.BOW 模型最初应用于文本处理领域,目前已广泛应用在场景分类等计算机视觉领域[18].BOW 的基本原理是,对图像特征进行聚类得到一个视觉词汇表(一个聚类表示一个视觉词),一幅图像就可以表示为若干视觉词的集合.假设某种特征经过聚类得到 N_v 个视觉词,则图像 d_i 可以表示为一个 N_v 维的直方图 $v(d_i)$:

$$v(d_i)=\{n(d_i,v_1),...,n(d_i,v_j),...,n(d_i,v_{N_v})\} \tag{7}$$

这里,$n(d_i,v_j)$表示图像 d_i 中含有视觉词 v_j 的个数.多种特征的直方图可以经过简单的连接成为新的 BOW 表示.可见,BOW 模型为不同的特征提供了一个一致的表示,而每种特征也都能通过 BOW 模型为图像的内容表示做出各自的贡献.使用 BOW 模型表示图像的处理过程如图 2 所示.PLSA-FUSION 使用 BOW 模型集成了若干视觉特征.首先,对每幅图像使用 DOG(difference-of-Gaussians)点检测器在不同尺度和位置进行采样,然后将检测到的极值区域表示为 SIFT(scale invariant feature transform)描述子[19],这个特征非常适用于识别图像中的物体;其次,使用固定网格将图像划分为固定大小的方块,并采用 HSV(hue-saturation-value)颜色直方图[20]和 LBP(local binary pattern)纹理描述子[21]来表示各个小方块的特征.这两类特征都采用 K-means 聚类算法进行量化以得到对应的 BOW 表示.将两类特征的视觉词直方图进行连接,则每幅图像就可表示为视觉词分布所对应的直方图.

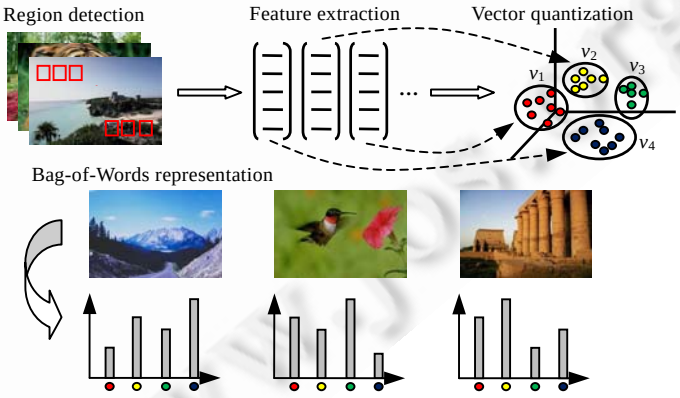


Fig.2 Process to compute the bag-of-words representation for images

图 2 用“词袋”模型表示图像的处理过程

另外,对文本标题的表示是比较直接的,因为关键词本身就是离散数据.图像集中的文本标题一般由给定的文本词汇表中的一些无序的关键词构成,假设关键词个数为 N_w,则图像 d_i 的文本标题可以表示为一个 N_w 维的

直方图 $w(d_i)$:

$$w(d_i)=\{n(d_i,w_1),...,n(d_i,w_j),...,n(d_i,w_{N_w})\} \tag{8}$$

这里,$n(d_i,w_j)$表示图像 d_i 的标题中含有 w_j 的个数(一般为 1).

经过以上预处理步骤,每幅图像都可表示为一个视觉词集合.特别地,对于已标注的训练图像,每幅图像都既可以表示为一个视觉词集合,同时也可以表示为一个文本词集合,这为使用 PLSA 建模图像数据提供了方便.

3.2 自适应不对称学习

使用 PLSA 进行图像自动标注的基本原理是:首先学习训练集中各个图像的主题分布 $P(z|d)$,然后依据这个参数进一步学习视觉词和文本词在给定各个语义主题下的分布 $P(v|z)$和 $P(w|z)$.由 PLSA 的条件独立假设,这两个分布独立于具体的训练图像,对于训练集之外的图像也是有效的.于是,给定一幅未知图像 d_{new},可以依据自动获取的视觉词表示 $v(d_{new})$和训练得到的参数 $P(v|z)$,使用 folding-in 算法计算该图像的主题分布 $P(z|d_{new})$,从而计算概率 $P(w|d_{new})$,并经过排序得到对应的标注关键词集合.

PLSA-WORDS 称其使用了不对称学习算法学习训练图像的主题分布 $P(z|d)$,而事实上,它仅从文本模态的数据中学习.也就是说,它利用文本模态的数据构造潜在空间,然后再将此空间链接到视觉模态.为了充分利用训练集的两种模态的数据,PLSA-FUSION 采用了两个 PLSA 模型分别建模视觉模态和文本模态的数据,然后再以自适应的方式不对称地融合两个 PLSA 模型,使得它们共享同样的潜在空间(即对于每幅训练图像具有相同的主题分布).融合的具体方式是,对于每幅图像,通过对从不同模态学习得到的主题分布进行加权,合并为一个新的主题分布.每种模态的融合权值由它们各自对图像内容的贡献来决定,而这个贡献由图像本身的视觉词分布的熵值来衡量.在一系列的实验中我们发现,如果视觉词直方图各个维的分布表现为稀疏和高峰态的特性,则从该图像的视觉模态学到的主题分布更加可靠;反之,若各个维的分布较平均,则从文本模态学到的主题分布更可靠.于是,通过为每幅图像计算其视觉词分布的熵,可以衡量该分布的稀疏度,从而确定该图像对应的视觉模态和文本模态的权值.综上所述,PLSA-FUSION 通过自适应地融合两个 PLSA 模型的语义主题来构造语义空间,其过程如图 3 所示.

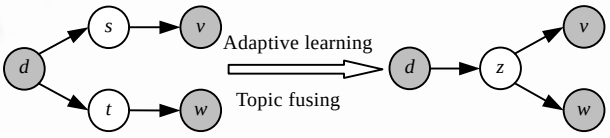


Fig.3 Fusing semantic topics with adaptive asymmetric learning algorithm

图 3 使用自适应不对称学习算法融合语义主题

假设视觉模态和文本模态对应的主题数分别为 m 和 n,那么融合后的模型有 $k=m+n$ 个主题.用 s 和 t 分别表示两个 PLSA 模型的主题,则视觉模态和文本模态对应的主题分布表示为 $P_v(s|d)$和 $P_w(t|d)$.通过拟合两个 PLSA 模型,得到每幅图像的两个主题分布 $P_v(s|d_i)$和 $P_w(t|d_i)$,则融合后的主题分布 $P(z|d_i)$由下式确定:

$$P(z_k\mid d_i)=\begin{cases}\alpha_{vi}P_v(s_k\mid d_i), & k=1,2,...,m\\ \alpha_{wi}P_w(t_{k-m}\mid d_i), & k=m+1,m+2,...,m+n\end{cases} \tag{9}$$

这里,α_{vi} 和 α_{wi} 分别表示图像文档 d_i 中视觉模态和文本模态的权值,可由下列经验公式计算得到:

$$\alpha_{vi}=\begin{cases}1, & H(v(d_i))\leqslant 3\\ \exp(3-H(v(d_i))), & H(v(d_i))>3\end{cases}$$
$$\alpha_{wi}=1-\alpha_{vi} \tag{10}$$

这里,$H(v(d_i))$表示图像文档 d_i 的视觉词分布 $v(d_i)$的熵.

实验结果表明,使用经验公式(10)进行学习,对于视觉词分布的熵值小于 4 或大于 6 的图像能够获得很好的标注效果,而对于熵值在 4~6 之间的图像的标注效果时好时坏.这是由于,图像大都具有较复杂的图像内容,仅仅依靠熵值和经验公式不能完全学习其复杂性,即无法确定视觉模态和文本模态所对应的最合理的权值.

自适应的不对称学习算法能够更好地控制各个模态对潜在空间的不同影响.一旦确定了主题分布 $P(z|d)$,则两种模态的元素在给定主题下的分布 $P(v|z)$和 $P(w|z)$就可以通过 folding-in 算法计算得到.由于主题分布 $P(z|d)$是从两种不同模态的语义信息中以自适应的方式学习得到,显然比从一种模态中学习得到的分布更为合理,故 PLSA-FUSION 比 PLSA-WORDS 具有更好的学习和泛化能力.

3.3 算法描述

下面给出基于 PLSA 建模的 3 种重要算法的描述,即训练算法、标注算法和检索算法.假设有一个包含图像和文本标题的训练集 $\mathbb{D}=\{(d_1,\boldsymbol{c}_1),\ldots,(d_N,\boldsymbol{c}_N)\}$,令 $\mathbb{T}_D=\{d_1,\ldots,d_N\}$表示训练图像集,$\mathbb{L}=\{w_1,\ldots,w_L\}$表示词汇表,则有图像 $d_i\in\mathbb{T}_D$ 和文本标题 $\boldsymbol{c}_i\subset\mathbb{L}(i\in 1,\ldots,N)$.此外,假设有测试图像集 $\mathbb{T}_T$,且有 $\mathbb{T}_T\cap\mathbb{T}_D=\varnothing$.

训练算法建模训练集 $\mathbb{D}$中的数据并学习图像和文本之间的关联,其算法描述如下:

(1) 对于每幅训练图像 $d_i\in\mathbb{T}_D$,提取其视觉特征并经过量化得到其视觉词表示 $v(d_i)$;处理与 d_i 连接的文本标题 $\boldsymbol{c}_d$ 并得到文本词表示 $w(d_i)$;

(2) 基于视觉词和文本词的表示 $v(d_i)$和 $w(d_i)$,分别拟合两个 PLSA 模型,可以得到两套模型参数:$P_v(v|s)$,$P_v(s|d)$和 $P_w(w|t)$,$P_w(t|d)$;

(3) 引入融合参数α_{vi} 和α_{wi} 来评估视觉模态和文本模态在图像 d_i 的参数估计中的重要性.使用公式(10)来计算融合参数,并使用公式(9)融合主题分布 $P_v(s|d_i)$和 $P_w(t|d_i)$,得到 $P(z|d_i)$;

(4) 根据融合后的主题分布 $P(z|d_i)$,使用 folding-in 算法计算最终的训练结果 $P(v|z)$和 $P(w|z)$,这个结果对于训练集外的图像仍保持有效.

标注算法用于处理训练集之外的新图像 $d_{new}\notin\mathbb{T}_D$,执行以下步骤:

(1) 对于每幅新图像 d_{new},执行训练算法的步骤(1);

(2) 给定 d_{new} 的视觉词表示 $v(d_{new})$和执行训练算法得到的参数 $P(v|z)$,可以使用 folding-in 算法推断图像的主题分布 $P(z|d_{new})$;

(3) 使用下式计算词汇表$\mathbb{L}$中各个关键词的后验概率:

$$P(w\mid d_{new})=\sum_{k=1}^{K}P(w\mid z_k)P(z_k\mid d_{new}) \tag{11}$$

(4) 选取具有最大后验概率的若干个关键词标注图像 d_{new}.

最后,检索算法将查询关键词 w_q 和测试集 $\mathbb{T}_T$ 作为输入,包含下列步骤:

(1) 对于测试集 $\mathbb{T}_T$ 中的每幅图像,执行标注算法中的步骤(1)~步骤(4);

(2) 对查询关键词 w_q,依据后验概率 $P(w_q|d)$对图像进行排序,并按降序输出若干幅图像作为检索结果.

通过以上 3 种算法,可以利用 PLSA 对图像的视觉信息、文本信息、上下文信息以及图像与文本之间的关联信息进行建模和集成,从而可以有效地完成图像的自动标注和语义检索的任务.

4 实验结果分析

为了检验 PLSA-FUSION 的性能和精度,我们开发了一个原型系统.该系统不仅实现了 PLSA-WORDS 和 PLSA-FUSION 的图像自动标注方法,也能进行图像的语义检索.模型的拟合过程和图像的自动标注采用离线方式执行,图像的语义检索采用在线方式执行.

4.1 数据集和实验设置

为了测试 PLSA-FUSION 的有效性并与其他前沿的图像自动标注方法进行比较,我们采用文献[12]使用的数据集(称为 Corel5k)进行实验.该数据集包含 5 000 幅图像,来自 50 个 Corel 库存图像 CD,每个 CD 包含同样语义内容的 100 幅图像,每幅图像标注有 1~5 个关键词.Corel5k 共有 371 个关键词,将至少标注了 8 幅图像的关键

词选入词汇表,合计 260 个关键词.整个数据集分为 3 部分:4 000 幅图像作为训练集,500 幅作为验证集,500 幅作为测试集.验证集用于确定系统参数,而当系统参数确定之后,4 000 幅图像的训练集和 500 幅图像的验证集就合并为一个新的 4 500 幅图像的训练集,这个训练集与文献[12]使用的训练集一致.

实现 PLSA-FUSION 方法主要有两个参数需要预先设置:一个是各种特征的视觉词个数,即使用 *K*-means 算法进行聚类的个数.这个参数对性能有较大影响,因为它决定了 BOW 模型的粒度,从而也决定了特性信息的损失程度.理论上说,视觉词个数越大,信息损失程度越低,从而可以获得更好的效果.但根据在验证集上的实验数据,视觉词个数在 1 000 左右达到最佳效果,即当视觉词个数超过 1 000 时,性能增长不明显,反而要付出较大的时间代价.在本实验中,SIFT 描述子和方块特征的视觉词个数都设置为 1 000,于是,每幅图像就可以表示为一个 2 000 维的直方图,每一维表示图像 d_i 中包含视觉词 v_j 的个数 $n(d_i,v_j)$;另一个参数是 PLSA 模型的潜在主题个数.这个参数决定了 PLSA 模型的容量——即 PLSA 模型的未知参数个数,从而在很大程度上确定训练时间和系统的效率.若主题个数过小,则得到的 PLSA 模型就不能充分表示数据的内在联系;若主题个数过大,则系统的效率会大为降低.并且随着 PLSA 模型未知参数的增加,其过拟合的可能性也会增大.通过实验验证,本系统对于 PLSA-WORDS 使用 200 个潜在主题;对于 PLSA-FUSION,使用 120 个潜在主题学习文本模态信息,使用 80 个潜在主题学习视觉模态信息,总共仍为 200 个潜在主题,这也保证了两种方法进行比较时的公平性.

PLSA 模型存在过拟合的问题,本系统使用提早停止技术控制过拟合.也就是说,不需要等到算法完全收敛时才停止迭代,而只要 hold-out 数据的性能不再提高就停止迭代过程.使用 folding-in 似然函数能够得到较好的性能而不需要使用缓和的 EM 算法[22],故本算法基于验证集的似然来决定迭代的终止条件.验证集的 folding-in 似然函数定义如下:

$$\mathcal{L}_{valid}=\sum_{i=1}^{N_{valid}}\sum_{j=1}^{M}n(d_i,x_j)\log\sum_{k=1}^{K}P(z_k\mid d_i)P(x_j\mid z_k) \tag{12}$$

4.2 自动标注结果比较

图像标注的性能通过比较测试集的自动标注与原始标注进行评估.类似于文献[14,15],PLSA-FUSION 只取前 5 个后验概率最大的关键词作为每幅图像的标注结果,并计算测试集中每个关键词的精度(也称为查准率)和召回率(也称为查全率).对于一个给定的语义关键词 w_q,精度 $precision=B/A$,召回率 $recall=B/C$.这里:A 表示所有自动标注了 w_q 的图像个数;B 表示正确标注 w_q 的图像个数,即这些图像的原始标注和自动标注都包含 w_q;C 表示原始标注中包含 w_q 的图像个数.于是,计算精度和召回率的平均值就能总结系统的标注性能.此外,我们也考虑了召回率大于 0 的关键词个数,这个数值表示系统能够有效学习的关键词的个数.

在本节中,使用平均精度和平均召回率比较若干图像自动标注方法的性能,包括翻译模型 TM[12]、跨媒体相关模型 CMRM[14]、连续空间相关模型 CRM[15]、PLSA-WORDS[7]和本文提出的方法 PLSA-FUSION.在两个关键词集合上报告标注结果:具有最佳性能的 49 个关键词构成的子集[12]和词汇表中全部 260 个关键词构成的集合.表 1 给出了在这两个集合上的标注性能,从表中数据可以看出,PLSA-FUSION 的性能不仅优于离散模型 TM,CMRM 和 PLSA-WORDS,也优于连续模型 CRM.我们认为,产生这样的结果的原因是,PLSA-FUSION 使用了 BOW 模型集成视觉特征及自适应的不对称学习算法.

表 2 给出了几个自动标注的结果实例,包括 PLSA-WORDS 和 PLSA-FUSION 的标注结果.从表中可见,对于某些视觉词分布较稀疏的图像(如第 3 幅图像),PLSA-FUSION 的标注结果要明显优于 PLSA-WORDS;而对于某些背景较复杂的图像则优势不明显.此外,对图像自动标注的关键词即使没有出现在原始标注中,这个词也能在某种意义上合理地标注该图像(如第 1 幅图像的标注 trees 和第 2 幅图像的标注 sand 等).

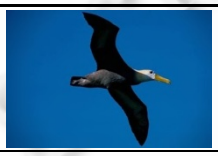
Table 1 Performance comparison on the task of automatic image annotation

表 1 图像自动标注的性能比较

Models	TM	CMRM	CRM	PLSA-WORDS	PLSA-FUSION
#words with recall>0	49	66	107	105	112
Results on 49 best words, as in Ref.[12,14,15]					
Mean per-word recall	0.34	0.48	0.70	0.71	0.76
Mean per-word precision	0.20	0.40	0.59	0.56	0.65
Results on all 260 words					
Mean per-word recall	0.04	0.09	0.19	0.20	0.22
Mean per-word precision	0.06	0.10	0.16	0.14	0.19

Table 2 Comparison of annotations made by PLSA-WORDS and PLSA-FUSION

表 2 PLSA-WORDS 和 PLSA-FUSION 的标注结果比较

Image				
Ground truth	Horses, mare, foal, field	Pyramids, stone, people, camels	Waved, albatross, flight, sky	Garden, flowers, landscape, trees
PLSA-WORDS annotation	Trees, garden, house, mare, foal	Stone, pyramids, mountain, columns, range	City, flight, ceremony, pond, swallow-tailed	Flowers, garden, farm, trees, bench
PLSA-FUSION annotation	Trees, field, mare, foal, horses	Stone, pyramids, sand, sky, antelope	Flight, bird, sky, waved, albatross	Flowers, garden, grass, trees, farm

4.3 语义检索结果比较

平均精度和平均召回率也能评估语义关键词检索的性能,但是它们不能体现检索的排位结果.于是,我们引入另一个称为 mAP(mean average precision)的度量标准来评估语义关键词的检索结果.mAP 在文本检索和视频检索中都已广泛使用并成为度量检索性能的标准,它能够为检索的排位提供有意义的评估.为了计算 mAP,首先需要定义每个查询 q 的平均精度(average precision,简称 AP).AP 定义为查询 q 在各个正确检索的相关图像的排位 i 的精度之和除以本次查询 q 的相关图像个数 $rel(q)$,即

$$AP(q)=\frac{\sum_{i\in relevant} precision(i)}{rel(q)} \tag{13}$$

可见,查询 q 的 AP 对于检索结果的排位顺序是敏感的.于是,mAP 定义为检索系统 N_q 次查询的 AP 的平均值.使用 mAP 一个值就能评估整个系统的检索性能,即

$$mAP=\frac{\sum_{q=1}^{N_q} AP(q)}{N_q} \tag{14}$$

本节主要比较 PLSA-WORDS 和 PLSA-FUSION 的性能,也给出 CMRM 和 CRM 相应的 mAP 值.而 TM 的 mAP 没有在文献[12]中报告,故无法比较.

对标注结果的评估并没有考虑检索结果的排位顺序.然而,用户总喜欢对检索的图像进行排位,并希望排位在前的图像是相关图像.现实中,对于一个查询结果,大部分用户都不愿意浏览多于十几幅的图像.所以,排位顺序对于图像检索而言是非常重要的.给定一个查询关键词,本系统按照该关键词的后验概率排序并输出检索图像,按照排位计算各个关键词的 AP 和 mAP 的值.表 3 给出了几种标注方法的 mAP,第 1 列是在词汇表中所有关键词集合上计算得到的 mAP,第 2 列是在召回率大于等于 0 的关键词集合(即在测试集的原始标注中出现过的关键词集合)上计算得到的 mAP.由表中的数据可见,PLSA-FUSION 的检索性能要高于其他方法.

图 4 给出了 10 个关键词在 PLSA-WORDS 和 PLSA-FUSION 方法下对应的 AP 值.由图 4 可见,无论是对学习效果较好的关键词还是学习效果不够充分的关键词,采用 PLSA-FUSION 获得的 AP 值大部分都高于采用 PLSA-WORDS 获得的 AP 值.

Table 3 Comparison of ranked retrieval results

表 3 排位检索结果比较

Mean average precision for Corel5k dataset		
Models	All 260 words	Words with recall≥0
CMRM	0.169 7	0.196 1
CRM	0.235 3	0.271 9
PLSA-WORDS	0.221 3	0.255 7
PLSA-FUSION	0.257 8	0.297 9

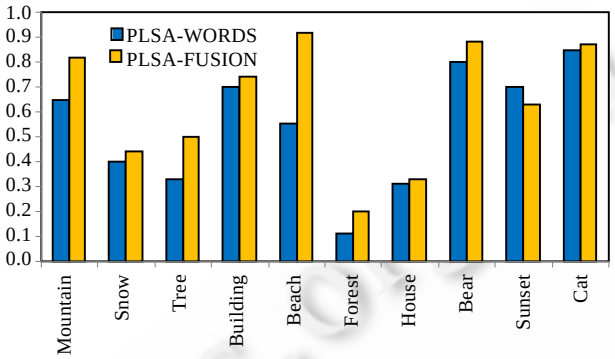


Fig.4 Average precision of selected 10 words when using PLSA-WORDS and PLSA-FUSION

图 4 使用 PLSA-WORDS 和 PLSA-FUSION 进行标注时给定的 10 个关键词的平均精度值

图 5 给出了两个关键词检索的实例,图中每行给出一个查询中排位最高的 5 幅匹配图像.第 1 行的查询关键词为 tiger,第 2 行为 street.返回图像的多样性表明,PLSA-FUSION 具有较好的学习和泛化能力.

Fig.5 Semantic retrieval results obtained by PLSA-FUSION

图 5 PLSA-FUSION 的语义检索实例

综上,所有实验结果表明,PLSA-FUSION 的自动标注和语义检索性能要高于其他几种前沿的图像标注方法,这证明了使用自适应的不对称学习算法对 PLSA 模型的语义主题进行融合是有效且可行的.然而,本方法也存在一定的局限性:首先,由于要对两种模态的数据建模并进行主题的融合,本方法需要花费更多的时间来学习模型参数,因而学习效率不高;其次,由于采用视觉词的离散形式表示图像,故不能完全消除聚类粒度的选择对标注性能的影响.

5 结论与展望

本文设计开发了一个基于 PLSA 的图像自动标注系统,它使用两个链接的 PLSA 模型分别学习视觉模态和文本模态的数据.此外,为了融合从两种模态数据中学习得到的语义主题,提出了一种自适应的不对称学习算

法.本系统在一个包含 5 000 幅图像的数据集中进行实验,对提出的标注方法进行评测.实验结果表明,我们的方法比其他几种前沿的图像标注方法具有更好的性能.

由于 PLSA 只能对离散量进行建模,所以本方法在对图像进行预处理时需要对视觉特征进行量化.尽管可以使用 BOW 模型对不同的特征进行集成,但特征数据在量化过程中仍会丢失重要信息,从而影响图像标注和检索的性能.下一步的工作准备对 PLSA 进行改进,使之能够直接对连续特征建模,从而避免特征量化的信息损失,以达到更高的标注精度和更好的检索效果.

References:

[1] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-Based image retrieval at the end of the early years. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(12):1349−1380. [doi: 10.1109/34.895972]

[2] Datta R, Joshi D, Li J, Wang JZ. Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys, 2008, 40(2):Article 5, 1−60. [doi: 10.1145/1348246.1348248]

[3] Li ZX, Shi ZP, Li ZQ, Shi ZZ. A survey of semantic mapping in image retrieval. Journal of Computer-Aided Design and Computer Graphics, 2008,20(8):1085−1096 (in Chinese with English abstract).

[4] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 2001,42(1-2):177−196. [doi: 10.1023/A:1007617005950]

[5] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003,3(1):993−1022. [doi: 10.1162/jmlr.2003.3.4-5.993]

[6] Blei DM, Jordan MI. Modeling annotated data. In: Proc. of the 26th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2003. 127−134. [doi: 10.1145/860435.860460]

[7] Monay F, Gatica-Perez D. Modeling semantic aspects for cross-media image indexing. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007,29(10):1802−1817. [doi: 10.1109/TPAMI.2007.1097]

[8] Li ZX, Liu X, Shi ZP, Shi ZZ. Learning image semantics with latent aspect model. In: Proc. of the IEEE Int'l Conf. on Multimedia and Expo. Los Alamitos: IEEE Computer Society Press, 2009. 366−369. [doi: 10.1109/ICME.2009.5202510]

[9] Li J, Wang JZ. Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003,25(9):1075−1088. [doi: 10.1109/TPAMI.2003.1227984]

[10] Chang E, Goh K, Sychay G, Wu G. CBSA: Content-Based soft annotation for multimodal image retrieval using Bayes point machines. IEEE Trans. on Circuits and Systems for Video Technology, 2003,13(1):26−38. [doi: 10.1109/TCSVT.2002.808079]

[11] Carneiro G, Chan AB, Moreno PJ, Vasconcelos N. Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007,29(3):394−410. [doi: 10.1109/TPAMI.2007.61]

[12] Duygulu P, Barnard K, de Freitas N, Forsyth D. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden A, Sparr G, Nielsen M, Johansen P, eds. Lecture Notes in Computer Science 2353. Berlin: Springer-Verlag, 2002. 97−112. [doi: 10.1007/3-540-47979-1_7]

[13] Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei DM, Jordan MI. Matching words and pictures. Journal of Machine Learning Research, 2003,3(2):1107−1135. [doi: 10.1162/153244303322533214]

[14] Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models. In: Proc. of the 26th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2003. 119−126. [doi: 10.1145/860435.860459]

[15] Lavrenko V, Manmatha R, Jeon J. A model for learning the semantics of pictures. In: Thrun S, Saul LK, Scholkopf B, eds. Advances in Neural Information Processing Systems 16. Cambridge: MIT Press, 2004. 553−560.

[16] Feng SL, Manmatha R, Lavrenko V. Multiple Bernoulli relevance models for image and video annotation. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2004. 1002−1009. [doi: 10.1109/CVPR.2004.1315274]

[17] Sivic J, Russell BC, Efros A, Zisserman A, Freeman WT. Discovering objects and their location in images. In: Proc. of the 10th IEEE Int'l Conf. on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2005. 370−377. [doi: 10.1109/ICCV.2005.77]

[18] Bosch A, Munoz X, Marti R. Which is the best way to organize/classify images by content? Image and Vision Computing, 2007, 25(6):778−791.

[19] Lowe DG. Distinctive image features from scale-invariant keypoints. Int'l Journal of Computer Vision, 2004,60(2):91−110. [doi: 10.1023/B:VISI.0000029664.99615.94]

[20] Swain MJ, Ballard DH. Color indexing. Int'l Journal of Computer Vision, 1991,7(1):11−32. [doi: 10.1007/BF00130487]

[21] Shi ZP, Hu H, Li QY, Shi ZZ, Duan CL. Texture spectrum descriptor based image retrieval. Journal of Software, 2005,16(6): 1039−1045 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/1039.htm [doi: 10.1360/jos161039]

[22] Brants T. Test data likelihood for PLSA models. Information Retrieval, 2005,8(2):181−196. [doi: 10.1007/s10791-005-5658-8]

附中文参考文献:

[3] 李志欣,施智平,李志清,史忠植.图像检索中语义映射方法综述.计算机辅助设计与图形学学报,2008,20(8):1085−1096.

[21] 施智平,胡宏,李清勇,史忠植,段禅伦.基于纹理谱描述子的图像检索.软件学报,2005,16(6):1039−1045. http://www.jos.org.cn/1000-9825/16/1039.htm [doi: 10.1360/jos161039]

李志欣(1971－),男,广西桂林人,博士,副教授,主要研究领域为图像理解,机器学习,基于内容的视觉信息检索.

李志清(1975－),男,博士,副教授,主要研究领域为图像理解,机器学习,视觉信息挖掘.

施智平(1974－),男,博士,助理研究员,主要研究领域为图像理解,机器学习,基于内容的视觉信息检索.

史忠植(1941－),男,研究员,博士生导师,CCF 高级会员,主要研究领域为人工智能,机器学习,神经计算,认知科学.