基于潜在主题融合的跨媒体图像语义标注

刘 杰[1,2],杜军平[1]
(1.北京邮电大学计算机学院,北京 100876;2.中国电子科技集团公司第三十研究所,四川成都 610041)

摘 要: 图像语义标注是图像语义分析研究中的一个重要问题.在主题模型的基础上,本文提出一种新颖的跨媒体图像标注方法来进行图像间语义的传播.首先,对训练图像使用主题模型,抽取视觉模态和文本模态信息的潜在语义主题.然后,通过使用一个权重参数来融合两种模态信息的主题分布,从而学习到一种融合主题分布.最后,在融合主题分布的基础上训练一个标注模型来给目标图像赋予合适的语义信息.在标准的 MSRC 和 Corel5K 数据集上将提出的方法与最近著名的标注方法进行比较实验.标注性能的详细评价结果表明提出方法的有效性.

关键词: 图像语义标注;跨媒体;主题模型;加权融合
中图分类号: TP37;TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2014)05-0987-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.05.024

Latent Topic Fusion-Based Cross-Media Image Semantic Annotation

LIU Jie[1,2], DU Jun-ping[1]
(1. *School of Computer, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*;
2. *No.30 Institute of China Electronics Technology Group Corporation, Chengdu, Sichuan* 610041, *China*)

Abstract: Image semantic annotation is an important issue in image semantic analysis research. Based on the topic model, this paper proposes a novel cross-media image annotation approach for propagating the semantics among images. First, the topic model is used to capture the latent semantic topics from the visual and textual modal information in the training images. Then, a fused topic distribution is learned by merging the topic distribution of each modality using a weight parameter. Finally, an annotation model based on the fused topic distribution is trained to assign the target images using appropriate semantics. A comparison of the proposed approach with the recent state-of-the-art annotation approaches on the standard MSRC and Corel5K datasets is presented, and a detailed evaluation of the performance shows the validity of our approach.

Key words: image semantic annotation; cross media; topic model; topic-weighted fusion

1 引言

在图像语义分析研究中图像语义标注是一种重要的手段.其中,一个关键的问题[1]是通过建立视觉特征和语义关键词的某种关联关系来决定图像属于某个语义概念.因此,一个有效的语义标注模型,应该将目标语义空间与图像特征空间关联起来,在训练和测试数据之间有效地传播语义信息,帮助跨越"语义鸿沟"[2].

基于相关模型的方法[3]是目前图像语义标注领域的一个研究热点.此种方法挖掘图像视觉特征集合与语义标注间的关联关系.该领域的一些早期工作[4]包括翻译模型(Translation model, TM),跨媒体相关模型(Cross-media relevance model, CMRM)和连续空间相关模型(Continuous-space relevance model, CRM).后来,出现了著名的多伯努利相关模型(Multiple Bernoulli relevance model, MBRM)[5].近期,又出现了结合空间马尔科夫核的统一相关模型(Generalized relevance model with spatial Markov kernel, GRM-SMK)[6].以上工作逐步提高了图像语义标注的性能.

主题模型是从文档语义分析中衍生出来的一种流行的机器学习技术[7],并被广泛的用于图像标注领域[8],其中比较有代表性的工作是 PLSA-WORDS 模型[9].特别的,潜在狄利克莱分配模型(Latent Dirichlet allocation, LDA)[10]作为一种具有代表性的主题模型被成功的用于挖掘文本和图像数据中的潜在语义主题信息[11].

收稿日期:2013-01-07;修回日期:2013-06-17;责任编辑:马兰英
基金项目:国家 973 重点基础研究发展计划(No.2012CB821200, No.2012CB821206);国家自然科学基金(No.61320106006)

本文提出一种基于潜在主题加权融合的跨媒体图像标注模型(LDA-based Latent Topic-Weighted Fusion, LDA-LTWF).关键在于决定标注关键词和目标图像之间的内在联系,这种内在联系帮助决定了底层图像特征和高层语义之间的一个中间过渡层.本文所提出的跨媒体图像标注方法的示意图如图 1 所示.

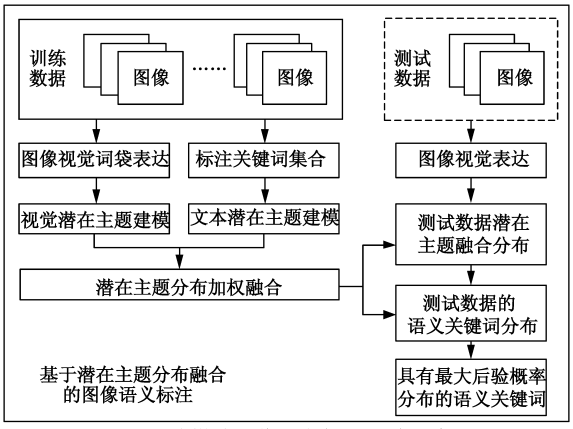


图1 跨媒体图像语义标注方法示意

2 潜在主题加权融合

训练数据的标注关键词通过文本词袋模型表示[7].通过融合图像的加速稳健特征[12]和多分辨率直方图矩特征[13]来生成一种复合底层特征[14],而后用视觉词袋模型对图像进行表示[15].利用基于 Gibbs 抽样的 LDA 模型计算文本模态和视觉模态数据的潜在主题分布[16].

2.1 潜在主题加权融合

在训练阶段,对每一幅图像学习融合潜在主题分布 $P(z|v)$.然后,对于每一个融合潜在主题 z 学习视觉词汇的后验分布 $P(x|z)$和文本标注词的后验分布 $P(w|z)$.在标注阶段,首先生成每幅图像的视觉词汇表达 $v_{new}=\{x_1,x_2,\cdots,x_N\}$.然后,利用在训练阶段得到的 $P(x|z)$为每幅图像计算融合潜在主题分布 $P(z|v_{new})$.最后,根据每个潜在融合主题 z 的文本标注词分布 $P(w|z)$计算测试图象的文本标注词分布 $P(w|v_{new})$.选择具有最大后验概率的文本词汇作为目标图像的语义标注词.潜在主题加权融合过程如图 2 所示.

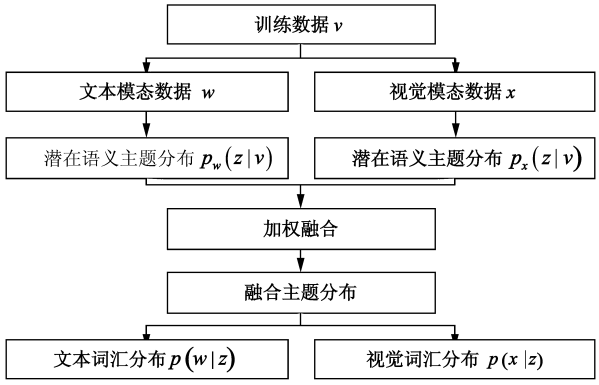


图2 潜在主题加权融合过程示意

假设视觉模态包含 k 个主题,文本模态包含 l 个主题.则融合主题空间就有 $t=k+l$ 个主题.视觉模态的主题分布是 $P_x(z|v)$,而文本模态的主题分布是 $P_w(z|v)$.对于图像 v_i,视觉模态的主题分布是 $P_x(z|v_i)$,而文本模态的主题分布是 $P_w(z|v_i)$.该图像在融合主题空间的主题分布可由下式得出:

$$p(z_t|v_i)=\begin{cases}\tau p_x(z_t|v_i), & t=1,2,\cdots,k\\ (1-\tau)p_w(z_{t-k}|v_i), & t=k+1,k+2,\cdots,k+l\end{cases} \tag{1}$$

在式(1)中,τ 表示图像 v_i 的融合主题分布中视觉模态部分的权重.τ 由下式计算得出:

$$\tau=\frac{\alpha-H(x(v_i))}{\alpha} \tag{2}$$

在式(2)中,$H(x(v_i))$是图像 v_i 的视觉词汇分布的信息熵.α 是 $H(x(v_i))$的上界,通过交叉验证得出.

2.2 图像语义标注

给定训练集合 $L=\{(v_1,c_1),(v_2,c_2),\cdots,(v_N,c_N)\}$,$V=\{v_1,v_2,\cdots,v_N\}$为图像集合,$C=\{c_1,c_2,\cdots,c_N\}$为标签集合.每个 c_i 包含若干关键词$\{w_i\}$,关键词集合为 $W=\{w_1,w_2,\cdots,w_N\}$.测试集合为 V_T,$V_T\cap V=\varnothing$.标注过程细节如下:

步骤 1 对于 $v_i\in V$,计算其视觉词汇表达 $v_i=\{x_1,x_2,\cdots,x_N\}$.对于标注信息 c_i,生成文本表达 $c_i=\{w_1,w_2,\cdots,w_M\}$.

步骤 2 利用 LDA 模型计算概率分布 $P(z_x|v)$、$P(x|z_x)$、$P(z_w|v)$和 $P(w|z_w)$.

步骤 3 使用式(2)计算权重参数 τ.使用式(1)融合概率分布 $P(z_x|v)$和 $P(z_w|v)$,生成融合分布 $P(z|v)$.

步骤 4 由步骤三中得出的 $P(z|v)$,利用 MCMC 算法计算对应的视觉词分布 $P(x|z)$和文本词分布 $P(w|z)$.

步骤 5 在标注阶段,对测试图像 $v_t\in V_T$,计算其视觉词汇表达 $v_t=\{x_1,x_2,\cdots,x_N\}$.

步骤 6 利用 MCMC 算法以及步骤四中得出的 $P(x|z)$来计算其融合主题分布 $P(z|v_t)$.

步骤 7 计算关键词集合 W 中每个关键词的后验概率,计算式如下:

$$p(w|v_t)=\sum_{n=1}^{N}p(w|z_n)p(z_n|v_t) \tag{3}$$

步骤 8 选择具有最大后验概率的关键词来标注测试图像 v_t.

3　实验

本文使用 MSRC 数据集[17]和 Corel5K 数据集[18].使用准确率、召回率、F 度量和召回率非零的关键词数量来评价标注方法的性能.

3.1　超变量和交叉验证

两个超变量分别是视觉词汇的数量和潜在主题的数量.同时,还要决定视觉词汇分布信息熵的上界.首先,使用传统的 K-means 算法来对图像特征进行聚类.聚类数依次为 100、200、400、600、700、800、900 和 1000,使用 F 度量值作为评价指标.对于 MSRC 数据集,使用 10 个主题学习文本模态,使用 50 个主题学习视觉模态;对于 Corel5K 数据集,使用 50 个潜在学习文本模态,使用 50 个主题学习视觉模态.表 1 显示了取自 10 折交叉验证平均值的比较结果.当 k 大于 800 时,两个数据集的 F 度量值有微弱的提高,但是计算花销却显著增加.因此,对于 MSRC 和 Corel5K 数据集,本文使用 k = 800 作为视觉词汇数的最佳取值.

表 1　不同视觉词汇数量在 10 折交叉验证中的平均 F 度量值

	MSRC	Corel5K
k = 100	0.29	0.11
k = 200	0.40	0.19
k = 400	0.48	0.24
k = 600	0.55	0.27
k = 700	0.58	0.29
k = 800	0.59	0.30
k = 900	0.59	0.30
k = 1000	0.59	0.30

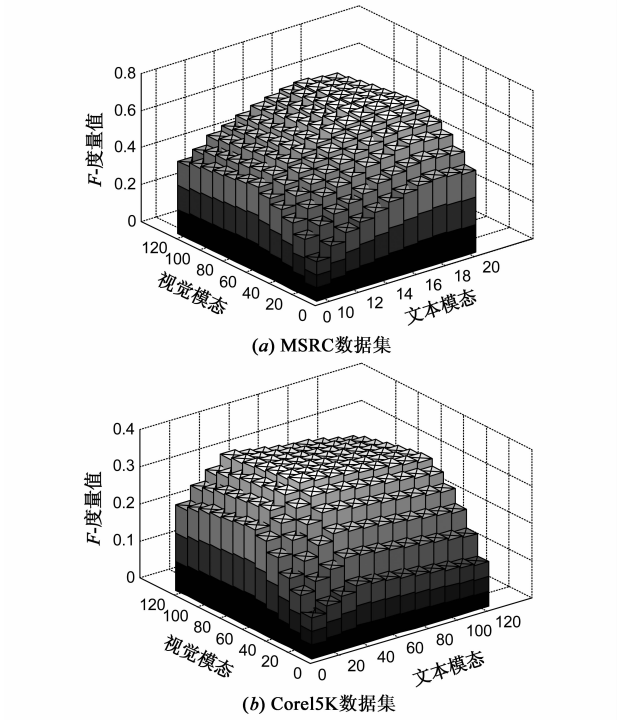
然后确定视觉词汇分布的信息熵的上界.视觉词汇分布的信息熵用 $H(x(v_i))$表示,满足 $0 \leqslant H(x(v_i)) \leqslant \ln k$[16].其中,$k$ 表示视觉词汇的数目.因此,式(2)中需要的视觉词汇分布信息熵的上界为 $\alpha = \ln 800$.

接着,对于 MSRC 和 Corel5K 数据集,估计最佳潜在主题的数目.其中,对于两个数据集,视觉模态的潜在主题数从 10 增加到 120,间隔为 10.对于 MSRC 数据集,文本模态的潜在主题数从 10 增加到 20,间隔为 1;对于 Corel5K 数据集,文本模态的潜在主题数从 10 增加到 120,间隔为 10.图 3 显示了在 MSRC 和 Corel5K 数据集上文本和视觉模态潜在主题数目的联合交叉验证中取得的 F 度量值比较结果.

从图 3(a)中可以看出,对于 MSRC 数据集,当文本模态的主题数为 18,且视觉模态的主题数为 60 时,F 度量值为最优.从图 3(b)中可以看出,对于 Corel5K 数据集,当文本模态的主题数为 40,且视觉模态的主题数为 60 时,F 度量值为最优.

3.2　标注性能

在 MSRC 数据集上,使用仅利用视觉模态成分的 LDA-LTWF 模型(L-VM)、MBRM 模型和 GRM-SMK 模型与 LDA-LTWF 模型进行比较.不同模型的标注结果取自 10 折交叉验证的平均值,如表 2 所示.



(a) MSRC数据集

(b) Corel5K数据集

图3　文本和视觉模态潜在主题数目的联合交叉验证中取得的F度量值

表 2　MSRC 数据集上的标注结果比较

	平均准确率	平均召回率	平均 F 度量值
L-VM	0.40	0.47	0.44
MBRM[5]	0.43	0.53	0.48
GRM-SMK[6]	0.61	0.62	0.62
LDA-LTWF	0.65	0.70	0.68

从表 2 中可以看出 LDA-LTWF 模型的标注性能大幅超过 L-VM 模型.这证明训练数据的视觉信息和文本信息的融合确实发挥了作用,并且能够较使用单一模态信息取得更好的标注性能.利用威尔考克森符号秩检验($P < 0.05$)对标注结果进行测试,LDA-LTWF 模型在平均准确率、平均召回率和 F 度量值等指标上比其他模型中的最优者依次高出 7%、13% 和 10%.同时,所有关键词的召回率均不为零.

在 Corel5K 数据集上使用 L-VM 模型、MBRM 模型、GRM-SMK 模型以及 PLSA-WORDS 模型[9]与 LDA-LTWF 模型进行比较,结果如表 3 所示.利用威尔考克森符号秩检验($P < 0.05$)对标注结果进行测试.LDA-LTWF 模型在平均准确率、平均召回率和 F 度量值等指标上比其他模型中的最优者依次高出 7%、12% 和 10%.同时,

PLSA-WORDS 模型、L-VM 模型、MBRM 模型、GRM-SMK 模型和 LDA-LTWF 模型的召回率非零的关键词数依次为 105、120、122、143 和 146. 图 4 显示了 Corel5K 数据集上图像原始标注与 LDA-LTWF 模型标注的比较结果. LDA-LTWF 模型能够为一些图像标注上原始标注中没有的关键词,并且这些关键词是合理的.

图像			
原始标注	马匹, 母马, 马驹, 场地	云朵, 树木, 天空, 山脉	汽车, 建筑物, 天空
LDA-LTWF 标注	树木, 马匹, 母马, 马驹, 场地	云朵, 树木, 天空, 山脉, 倒影	天空, 汽车, 城市, 建筑物, 云朵
图像			
原始标注	建筑物, 汽车, 地平线, 街道	喷气飞机, 树木, 天空,	轮船, 天空, 海水
LDA-LTWF 标注	建筑物, 汽车, 地平线, 街道, 铁轨	喷气飞机, 树木, 飞机, 天空, 跑道	轮船, 天空, 云朵, 海水, 桅杆

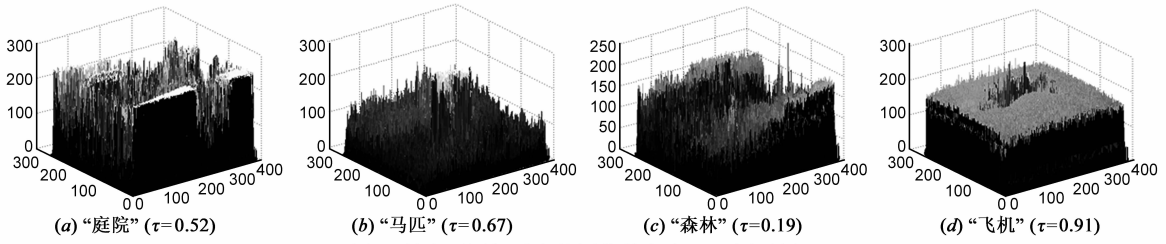
图4 Corel5K数据集上图像原始标注与LDA-LTWF模型标注的比较

表 3 Corel5K 数据集上的标注结果比较

	平均准确率	平均召回率	平均 F 度量值
PLSA-W[9]	0.14	0.20	0.17
L-VM	0.22	0.25	0.24
MBRM[5]	0.24	0.25	0.25
GRM-SMK[6]	0.30	0.33	0.31
LDA-LTWF	0.32	0.37	0.34

3.3 权重参数的讨论

用每幅图像的灰度 mesh 图来直观表示图像的视觉内容. 在 Corel5K 数据集上的大量实验表明利用式(2)计算出的权重取值训练标注模型,当视觉词汇分布的信息熵小于 2 时,图像的语义标注性能较好. 这表明潜在主题融合分布中的视觉模态成分在图像语义学习过程中发挥了主要作用. 如果信息熵大于 4,标注模型的性能仍然较好,则融合分布中视觉模态成分的权重较低,文本模态成分发挥较大作用. 当信息熵的取值在 2 到 4 之间时,标注性能不甚令人满意. 这表明视觉词汇分布的信息熵在 2 到 4 之间的图像有着很强的内容复杂性. 因此,很难通过简单的权重取值来确定每种模态数据的贡献,从而难以学习到每幅图像所包含的准确的语义. 本文通过测试 4 幅样例图像来表明权重参数 τ 对标注性能的影响,如图 5 所示.



(*a*) "庭院" (τ=0.52) (*b*) "马匹" (τ=0.67) (*c*) "森林" (τ=0.19) (*d*) "飞机" (τ=0.91)

图5 具有不同权重取值图像的灰度mesh图示例

在图 5(*a*)中,"庭院"类图像视觉词汇分布的信息熵 $H(x(v_i))$为 3.2,因此其 τ 值为 0.52,其标注平均准确率为 0.55. 在图 5(*b*)中,"马匹"类图像视觉词汇分布的信息熵 $H(x(v_i))$为 2.8,因此其 τ 值为 0.58,其标注平均准确率为 0.67. 所以由于上述两类图像视觉内容的复杂性,使得难以学习到合适的权重取值. 在图 5(*c*)中,"森林"类图像视觉词汇分布的信息熵 $H(x(v_i))$为 5.4,因此其 τ 值为 0.19,其标注平均准确率为 0.70. 这种情况中标注性能的改进是得益于文本模态数据的贡献. 在图 5(*d*)中,"飞机"类图像视觉词汇分布的信息熵 $H(x(v_i))$为 0.60,因此其 τ 值为 0.91,其标注平均准确率为 0.73. 这种情况中标注性能的改进是得益于视觉模态数据的贡献.

4 结论

本文提出基于潜在主题加权融合的跨媒体图像语义标注方法,该方法的关键是对文本和视觉模态的潜在主题分布进行加权融合. 其中,各模态信息的潜在主题分布由 LDA 主题模型抽取. 然后利用融合潜在主题分布构建跨媒体图像语义标注模型. 最后使用 MSRC 数据集和 Corel5K 数据集对该模型进行实验验证. 实验结果证明了所提标注方法的有效性.

参考文献

[1] A Tousch, S Herbin, J Audibert. Semantic hierarchies for image annotation: A survey[J]. Pattern Recognition, 2012, 45(1): 333 – 345.

[2] J Tang, Z Zha, D Tao. Semantic-gap-oriented active learning for multi-label image annotation[J]. IEEE Transactions on Image Processing, 2012, 21(4): 2354 – 2360.

[3] D S Zhang, Md M Islam, G J Lu. A review on automatic image annotation techniques[J]. Pattern Recognition, 2012, 45(1): 346 – 362.

[4] A Makadia, V Pavlovic, S Kumar. Baselines for image annota-

tion[J]. International Journal of Computer Vision, 2010, 90(1):88 – 105.

[5] S Feng, R Manmatha, V Lavrenko. Multiple Bernoulli relevance models for image and video annotation[A]. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition[C]. Washington, DC, USA: IEEE, 2004: 1002 – 1009.

[6] Z W Lu, H S I Horace. Automatic image annotation based on generalized relevance models[J]. Journal of Signal Processing Systems, 2011, 65(1):23 – 33.

[7] J Zhong, Q G Sun, X Li, L S Wen. A novel feature selection method based on probability latent semantic analysis for Chinese text classification[J]. Chinese Journal of Electronics, 2011, 20(2):228 – 232.

[8] X Ke, S Z Li, D L Cao. A two-level model for automatic image annotation[J]. Multimedia Tools and Applications, 2012, 61(1):195 – 212.

[9] F Monay, D G Perez. Modeling semantic aspects for cross-media image indexing[J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2007, 29(10):1802 – 1817.

[10] D M Blei. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4):77 – 84.

[11] D Putthividhy, H T Attias, S S Nagarajan. Topic regression multi-modal latent Dirichlet allocation for image annotation [A]. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition[C]. La Jolla, CA, USA: IEEE, 2010: 3408 – 3415.

[12] H Bay, A Eelaars, L V Gool. Speed-up robust features (SURF)[J]. Computer Vision and Image Understanding, 2008, 110(3):346 – 359.

[13] Y Jiang, R Wang, P Zhang. Texture description based on multiresolution moments of image histograms[J]. Optical Engineering, 2008, 47(3):037005.

[14] J Liu, J P Du, X R Wang. Research on the robust image representation scheme for natural scene categorization[J]. Chinese Journal of Electronics, 2013, 22(2):341 – 346.

[15] W J Wen, D Xu, Y J Tang, S Y Liu, S H Feng. Mutual information based codebooks construction for natural scene categorization[J]. Chinese Journal of Electronics, 2011, 20(3):419 – 424.

[16] M B Christopher. Pattern Recognition and Machine Learning [M]. New York, USA. Springer. 2006.

[17] J Shotton, J Winn, C Rother, A Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation[A]. In Proc 9th European Conf. on Computer Vision[C]. Graz, Austria: Elsevier, 2006: 1 – 15.

[18] P Duygulu, K Barnard, J F G de Freitas, D A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary[J]. Lecture Notes in Computer Science, 2006, 2353:349 – 354.

作者简介

刘　杰　男,1984 年出生,博士,工程师,中国电子科技集团公司第三十研究所,主要研究方向:智能信息处理、机器学习、网络通信.
E-mail: sleetext2@163.com

杜军平(通信作者)　女,1963 年出生,博士,教授/博士生导师,北京邮电大学计算机学院,主要研究方向:人工智能、智能信息系统.
E-mail: junpingdu@126.com