A Review On Image Tagging/Annotation

Author: Zhu Xinge

April 18, 2016

Agenda

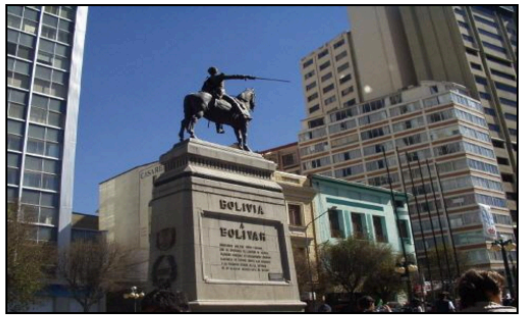
- Image Annotation
  - what?
  - why?
- Image Annotation Techniques
  - types
  - details
- Dataset & Result
- Future Work
- References

What is Image Annotation?

Definition

1.Automatic image annotation is a multi-label classification problem that aims at associating a set of textual with an image that describe its semantics. 2.Another view of the annotation is the task of matching images and sentences

{fight, grass, game, anime, man}



{building, base, horse, statue, man}

{fence, mountain, range, airplane, sky}

{bear, reflection, water, black, river}

{field, horses, mare, foals, tree}

{green, phone, woman, hair, suit}

Why Image Annotation?

Usage

Huge amount of digital images are produced every day, so there is need of automatic annotation.

1) The manual-annotation is high-priced and time-consuming.
2) Image annotation is one of best ways for image searching and retrieval.
3) Many other potential applications can be extended from the results of the annotation.

首先省时省力，其次他还是各种其他应用的基础，可以由此扩展出不同的应用，searching, retriveal, caption

Types of Annotation Techniques

Image Annotation

- Probabilistic Model
- Approaches Based on Nearest-Neighbors
- Mixture Model
- Recent Efforts

下面介绍一下图片标注的方法都有哪些，我这里简单的分成了三大类加近期的工作进展，
这三大类分别是概率模型，基于NN的模型，混合模型

Probabilistic Model

Representatives

- Latent Dirichlet Allocation(LDA)*
- Multiple Bernoulli Relevance Models(MBRM)*
- Probabilistic Latent Semantic Analysis(PLSA)[1]
- MRF* and ...

在概率模型中主要有一下几类代表方法，LDA，MBRM，PLSA，MRF，
下面主要针对PLSA做一个介绍，论文是CIKM2007的一个工作，该论文是将图片标注问题看做一个另类的翻译的问题，把文本中的topic的提取应用到了图片标注上。

Introduce the PLSA[1] in details

First of all, let's consider the model of word-to-topic, how to generate the topic given the document?

利用贝叶斯公式可以得到，右边的式子

$$P(topic|w,d) = \frac{P(w|topic)P(topic|d)}{P(w|d)}$$

[1] Wang Y. Translating Topics to Words for Image Annotation. CIKM, 2007.

PLSA

Con't

将右边的P(topicId)利用贝叶斯拆成如下形式，最后回带到原始公式中就得到了，P(topicIw,d)的表示

$$P(topic|w,d) = \frac{P(w|topic)P(topic|d)}{P(w|d)}$$

where w are words, d are documents. Due to

$$P(topic|d) = \frac{P(d|topic)P(topic)}{P(d)} \propto P(d|potic)P(topic)$$

Hence,

$$P(topic|w,d) \propto P(w|topic)P(d|topic)P(topic)$$

Employ the MLE to $P(topic|w,d)$ 针对这个求他的极大似然函数

$$L = \prod_{i=1}^{N}\prod_{j=1}^{M} P(topic|w_j,d_i)^{n(w_j,d_i)}$$

PLSA

Con't

Log the MLE

对上式求log，得到如下公式
对于该公式可以利用EM算法来
求解

$$log\ L = \sum_{i=1}^{N}\sum_{j=1}^{M} n(w_j, d_i) logP(topic|w_j, d_i)$$

Next employ the **EM alogorithm** to solve the $log\ L$

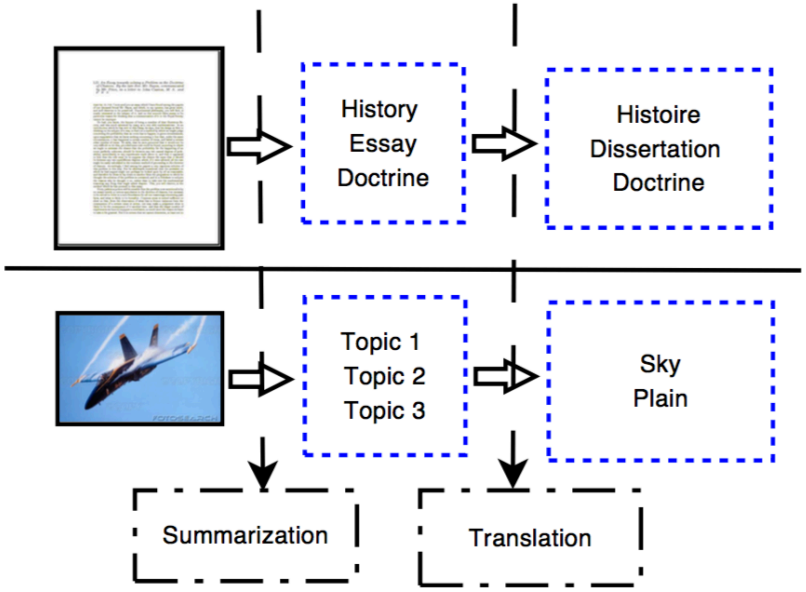
Ok, this is the word-to-model as the aforementioned.

How to project the word-to-topic to visual-to-annotation ?

word $\Longrightarrow$ visual word

documents $\Longrightarrow$ images

PLSA



The paper employs a slide window to fetch the image patch, cluster the patches according to the SIFT and Color to generate the visual word vocabulary, that's closely similar to the bag-of-words.

Approaches Based Nearest-Neighbors

NN for image annotation

- Two Pass KNN(2PkNN)*
- Joint Equal Contribution(JEC)*
- Visual-Clustering[2]
- .....

Introduce the Visual-Clustering in details

[2] **Tsai David. Large-scale image annotation using visual synset. ICCV, 2011.**

Visual-Clustering

Pipeline of Visual-Clustering

1. Visual Synsets 首先视觉同义词合集的提取，这里的Synset就是根据你不同的特征进行聚类得到的
首先根据不同的特征构建一个相似度矩阵，然后利用AP聚类，得到该视觉同义词合集
   - generate the similarity matrix according to the color, shape, local features, etc,(In practice, the strategy is more sophisticated.)
   - employ the affinity propagation(AP) for clustering.
   - visual synsets contain similar images and corresponding labels.
2. Assign the weights to labels according to the Score

$$Score = TF * IDF$$

Visual-Clustering

Con't

3. Train a linear SVM for each visual synset
4. Prediction by Voting
   - calculate feature **x** and pass it to all linear SVM
   - if the response of SVM is above a threshold T, then accept.
   - voting by aggregating the label Scores(TF*IDF)

$$L = \sum_{i=1}^{n} \sum_{j=1}^{m_i} Score_{i,j}$$

Mixture Model

Mixture Model

1. CCA-KNN*
2. KCCA-2PKNN*
3. NMF-KNN[3]
4. .....

第三类是一个混合模型，把之前效果不错的一些方法，进行一个mix。

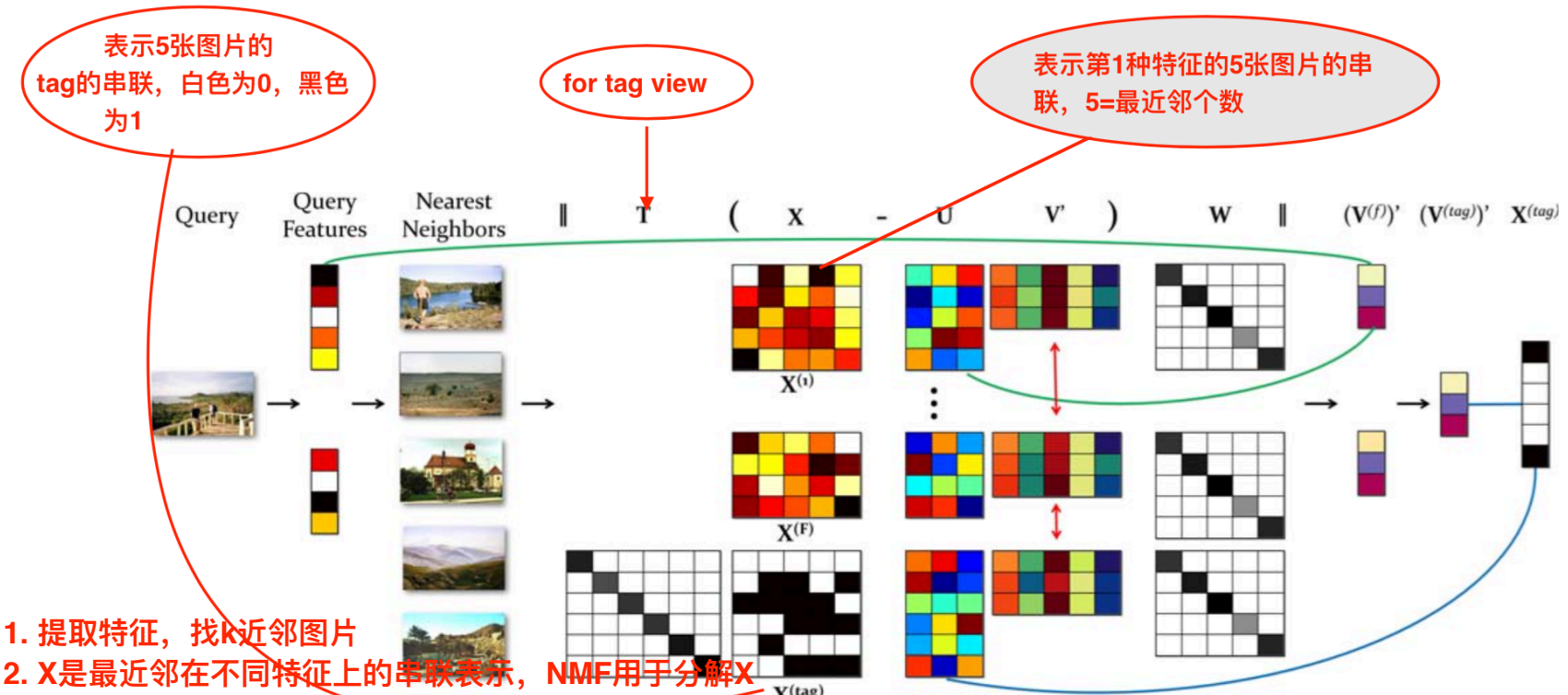
Introduce the NMF-KNN[3] in details

NMF:Non-negative Matrix Factorization

其中NMF就是非负矩阵分解，主要用在人脸和多媒体上

[3] Kalayeh M. NMF-KNN: Image annotation using weighted, multi-view non-negative matrix factorization, CVPR, 2014.

NMF-KNN



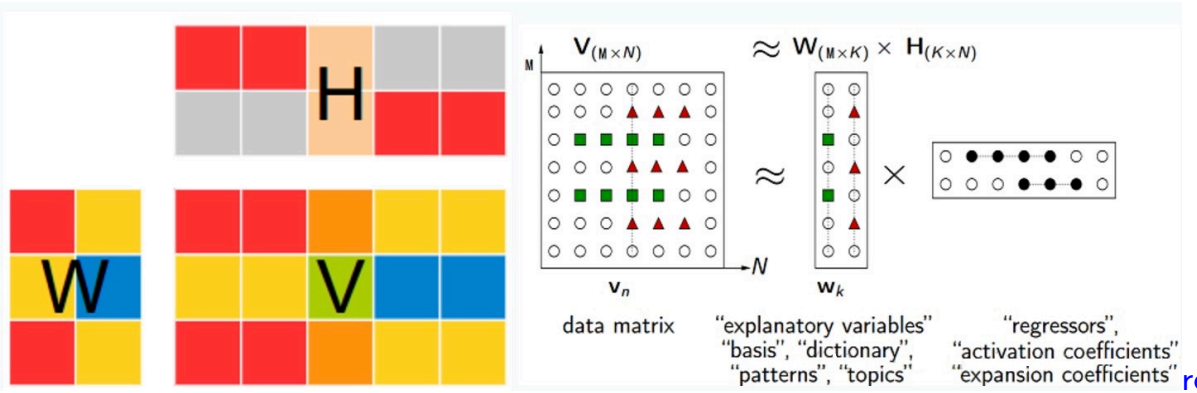
1. 提取特征，找k近邻图片
2. X是最近邻在不同特征上的串联表示，NMF用于分解X
3. U，V是通过NMF分解得到，其中V在不同特征之间保持一致性，如何做到就是下式子中的第二部分
4. 由于V的一致性，所以U可以很好的表征X的特征变换，所以我们通过query特征和U相乘，得到query在近邻图片的及特征下的表示，Vtag。
5. Xtag表示的5长图片的tag的串联，通过Utag和Vtag相乘，得到query的tag分布情况
6. T和W是为了解决tag的分布不均匀的情况
7. 下面式子中，前面表示的X与X分解的UV之间的误差，后面的是V进行归一化，保持一致性的误差，目标就是使得L最小，分解过程就是一个迭代的过程。

Where U is basis matrix and $X^{(tag)}$ is tag matrix, and X^{tag} is computed from NMF, $X^{tag} = U^{(tag)} V^{(tag)'}$

$$L = \sum_{f=1}^{F+1} ||T(X^{(f)} - U^{(f)} V^{(f)'}) W||_F^2 + \sum_{f=1}^{F+1} \lambda_f ||W (V^{(f)'} Q^{(f)} - V^*)||_F^2$$

NMF-KNN

NMF



ref: http://diamondxg.xyz/

Result of datasets of ESP Game

Figure 3. Example images from ESP Game dataset and the corresponding top 5 tags predicted using NMF-KNN are shown in this figure. Predicted tags in green appear in the ground truth while red ones do not. In many cases, even though the proposed method has predicted

Recent Efforts

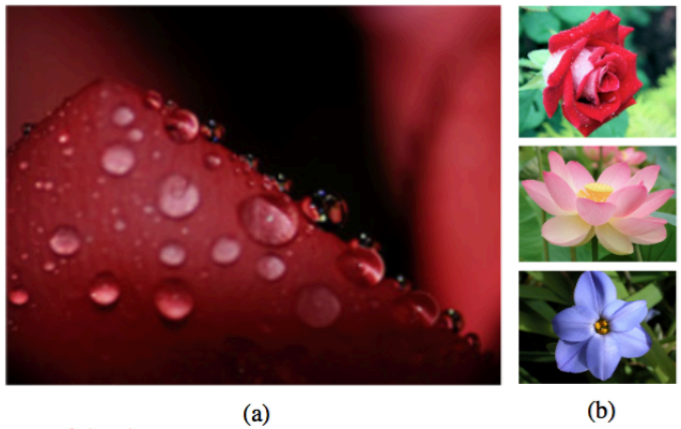
1. Deep Learning Representations[4]

- Feature Extraction
  1. extract the fc-layer feature from **VGG19** pretrained on the ILSVRC-2012
  2. obtain Word Embedding from a skip-gram text embedding architecture
  3. employ the **CNN Feature** and **Word Embeddings** into KCCA-KNN
- CNN Regression Model
  1. CNN provides the mapping function which regresses the fixed size of image to a word embedding vector
  2. use L2 instead of Softmax Loss

[4] **Murthy V N. Automatic Image Annotation using Deep Learning , Representations. ICMR 2015.**

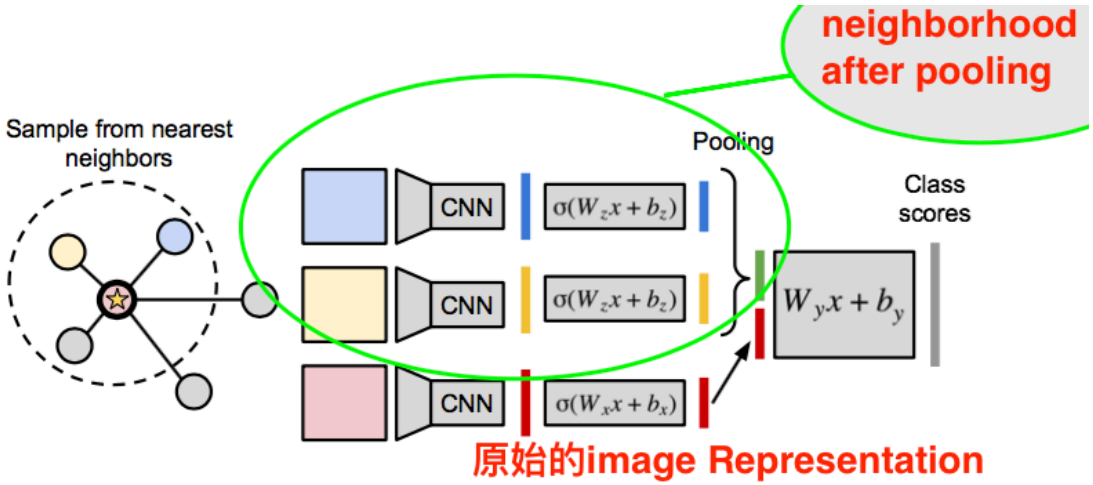
Recent Efforts

2. Incorporate Metadata[5]



Types of Metadata
are user-tags, image photo-sets, image groups Image photo-sets are galleries of image collected by same user. Image group are from same event or somethings.

1. 图片解释
2. metadata解释
3. 模型解释



[5] Justin J. Image Annotation by Exploiting Image Metadata, ICCV 2015

Recent Efforts

2. Incorporate Metadata

Con't

图片本身特征 $v_x = \delta(W_x\phi(x) + b_x)$

图片的近邻的特征，取max $v_z = \max_{i=1,\ldots,m}(\delta(W_z\phi(z_i) + b_z))$ (1)

变换操作，求class score $f(x, w; z) = W_y[v_x, v_z] + b_y$

details

The learnable parameters are W_x, b_x, W_z, b_z, W_y, b_y.

Apply dropout, RMSProp,and L_2 regularization

The loss function is a sum of one-vs-all logistic classifiers.

Recent Efforts

3. Misc

1. Introduce the Multiple Instance Learning to Image annotation[6] 类似于多目标检测的一种
2. Present a CNN Model with WARP loss[7]
3. RNN Fisher Vectors[8]

把annotation看做一种排序问题
给定图片，将tag根据相关性进行排序
warploss是用来排序的loss function

使用RNN处理tag，得到tag特征，然后利用FV来进行一个映射操作

[6] Wu J. Deep multiple instance learning for image classification and auto-annotation, CVPR 2015.

[7] Jia Y. Deep convolutional ranking for multilabel image annotation, ICMR 2013.

[8] Lev G. RNN Fisher Vectors for Action Recognition and Image Annotation. arXiv.

Dataset & Result

Ref: http://lijiancheng0614.github.io/

Dataset

Dataset	Corel 5K	ESP Game	IAPR TC-12	NUS-WIDE
No. of images	5000	20770	19627	269648 (209347 annotated)
No. of labels	260	268	291	81

Result

介绍一下常用的数据集和最新的结果
为什么效果不好？一个是跟数据库有关，

		Corel-5K				ESP Game				IAPRTC-12			
Method	**Year**	**P**	**R**	**F1**	**N+**	**P**	**R**	**F1**	**N+**	**P**	**R**	**F1**	**N+**
NMF-KNN	2014	38	**56**	45.3	150	33	26	29.1	238				
CCA-KNN	2015	42	52	46.5	**201**	46	**36**	**40.4**	**260**	45	38	41.2	278
context-RM-B	2015					**61**	24	34.4	242	**61**	20	30.1	234
SLED	2015	35	51	41.5						49.82	**47.36**	**48.6**	

Future Work

- transfor learning 把其他的模型迁移到这个问题上
- promising improvement in NLP 在图片特征的提取上，陷入了瓶颈，考虑在tag的处理上的新方法
- weakly supervised learning, semi-supervised learning
- . . . . . 由于问题本身数据量偏大，人工标注不是长久之计，所以需要弱监督，甚至是无监督的学习。

References

- K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. JMLR, 3(2), 2003.
- F. Monay and D. Gatica-Perez. pLSA-based image auto-annotation: constraining the latent space. In ACM MM, 2004.
- R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using un-aligned text corpora. In CVPR, 2010.J. Van De Weijer and C. Schmid. Col
- S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metaxas. Automatic image annotation using group sparsity.In CVPR, 2010.

References
Con't

- M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In ICCV, 2009.
- A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In ECCV, 2008.
- Y. Verma and C. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In ECCV. 2012.