SCENE-BASED AUTOMATIC IMAGE ANNOTATION

Amara Tariq , Hassan Foroosh

School of Electrical Engineering and Computer Science,
University of Central Florida, USA

ABSTRACT

Image search and retrieval systems depend heavily on availability of descriptive textual annotations with images, to match them with textual queries of users. In most cases, such systems have to rely on users to provide tags or keywords with images. Users may add insufficient or noisy tags. A system to automatically generate descriptive tags for images can be extremely helpful for search and retrieval systems. Automatic image annotation has been explored widely in both image and text processing research communities. In this paper, we present a novel approach to tackle this problem by incorporating contextual information provided by scene analysis of image. Image can be represented by features which indicate type of scene shown in the image, instead of representing individual objects or local characteristics of that image. We have used such features to provide context in the process of predicting tags for images.

Index Terms— Image annotation, Information extraction, Scene understanding, Context-based annotation

1. INTRODUCTION

Automatic generation of relevant and descriptive textual annotations for images can play an important role in the performance of image search, retrieval or archival systems. A popular approach to solve this problem, adapted from relevance models from machine translation, is to assume that image-description pairs are generated by independently chosen visual features for images and words for descriptions. Joint probability of visual features and words is estimated by expectation procedure over training data. This approach is computationally very efficient and flexible. The problem is the 'semantic gap' between information presented by low-level visual features and information described by words. Therefore, this approach has only been moderately successful. We argue that given information about context of visual content, there is a better chance of bridging the 'semantic gap' and predicting correct words for description. 'Scene' information provides a broad understanding of visual content while individual objects constitute details of the image and their appearance in the image is dependent on the type of scene presented in it. Studies have shown that humans can identify the type of scene shown in an image such as 'inside city' , 'open country scene' etc., without recognizing the individual objects [1, 2]. We argue that once the scene information is available, presence of certain objects in that image is more likely than presence of other objects. Thus certain words are more likely to be part of descriptions for images presenting certain scenes. For example, 'grass' is heavily likely to be a tag for 'open country scene'; 'cars' for 'inside city'. We present a novel image annotation system. This system models an image-description generative process which incorporates broad understanding of image content as context for generating the details of the image and its description. Our system achieves impressive performance without compromising on computational efficiency.

2. RELATED WORK

Various approaches have been explored for automatic generation of image annotations. Relevance models from machine translation have been adapted to develop systems for automatic image annotation [3, 4, 5, 6]. These models rely on estimation of joint probability of visual features and keywords. These models are computationally very efficient. Blei et al. presented the idea of modeling topics present in a text document as a hidden layer of variables [7]. This idea has also been used in automatic image annotation systems, assuming that document representation includes both visual and textual content [8, 9, 10]. Graph based or nearest-neighbor based automatic image annotation systems have shown considerable improvement in performance over relevance model based systems at the cost of increased computational complexity [11, 12, 13, 14, 15]. Iterative optimization algorithms are used to pick the best set of nearest neighbors of an image. Annotations are generated through transfer of tags form the nearest neighbors to the image itself. Social media websites have become a huge source of data. Research has been carried out to incorporate these sources in automatic image annotation systems [16, 17]. Another approach is to rely on recognition systems developed for various computer vision applications to learn about objects, attributes of objects and actions presented in images. These pieces of information are put together as nouns, adjectives and verbs respectively to form description of images [18, 19, 20]. This approach has been tested over very large dataset, but is restricted by the

availability of object and action recognition systems which are practically very limited in number.

We have worked to develop a system which combines computational efficiency and flexibility of relevance models with holistic image understanding to achieve better performance in the task of automatic image annotation.

3. MATHEMATICAL MODELING

We assume that each image is made up of A number of visual units i.e. $\mathbf{r} = \{r_1, r_2, ..., r_A\}$. We divide an image through a fixed size grid with A sections and calculate color and texture qualities for each section of the grid. Vector containing color and texture properties of one section of the grid is the smallest visual unit that we assume the image is made up of. On the other hand, image has some type of textual description associated with it. This description is made up of B number of words $\mathbf{w} = \{w_1, w_2, ..., w_B\}$. We also assume that there is certain set of scene-types i.e. $\mathbf{C} = \{C_1, C_2, ..., C_K\}$ available and the selection of visual units and words is dependent on the scene-type selected during generation of image-description pair. That is why we claim that the scene understanding provides context for creating details of the image and its description. We assume that the image-description pair being generated does not strictly belong to one scene-type and one scene-type only, rather it may have visual and textual characteristics of more than one scene types. $P(\mathbf{C}/\theta)$ represents association of an image, given a value for variable θ, with all scene-types. We need to estimate the joint probability of visual units and words that has generated the dataset and we will use following expectation process over training data weighted by $P(\mathbf{C}/\theta)$ for this purpose.

1. pick a scene-type $C_k \in \mathbf{C}$ with probability conditioned over variable θ i.e. $P(C_k/\theta)$
2. pick image J from training set $\mathbf{J}$ with probability $P(J/C_k)$
3. for a = 1,2,....,A
   (a) pick a visual unit r_a from conditional probability $P_R(./J)$
4. for b = 1,2,....,B
   (a) pick a word w_b from conditional probability $P_{V_{C_k}}(./J)$

Thus, joint probability of $\mathbf{r}$ and $\mathbf{w}$ conditioned over θ is given by following equation.

$$P(\mathbf{w}, \mathbf{r}/\theta) = \sum_{C_k \in \mathbf{C}} P(C_k/\theta) \sum_{J \in \mathbf{J}} P(J/C_k) \prod_{b \in B} P_{V_{C_k}}(w_b/J) \prod_{a \in A} P_R(r_a/J) \quad (1)$$

We will discuss how availability of the image in automatic image annotation problem enables us to calculate value of variable θ.

3.1. Estimation

Given an image i.e. its visual units, our goal is to select words w with highest $P(\mathbf{r}, \mathbf{w}/\theta)$ where θ has a specific value. We need to estimate $P_{V_{C_k}}(w_b/J)$, $P_R(r_a/J)$, $P(C_k/\theta)$ and $P(J/C_k)$ for all $C_k \in \mathbf{C}$ and $J \in \mathbf{J}$.

Oliva et al. presented holistic visual feature vector called GIST, to classify images based on the type of scene they present [21]. This work is based on assumption that humans can identify scene-type without recognizing individual objects in the image. Therefore, scene classification system does not need to focus on local characteristics or identification of individual objects in the image; rather it should work with holistic representation of the image.

We compute GIST features for all of our training images and cluster them to form K sets of images, each represented by $\mathbf{J_k}$ with size N_k, for a certain scene type C_k. For training image $J \in \mathbf{J}$

$$P(J/C_k) = \begin{cases} 1/N_k, & \text{if } J \in \mathbf{J_k}. \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

K can be selected using a threshold on within cluster entropy. We have observed that performance of our system remains stable for a wide range of K and have reported the best results. For all test images, i.e. images for which annotations need to be predicted, we also compute their GIST feature representation and this representation is the variable θ for the selected image. We estimate the probability of all C_k by using a non-parametric Gaussian kernel, such as

$$P(C_k/\theta) = \frac{\exp((G_k - \theta)^T \Gamma^{-1} (G_k - \theta))}{\sqrt{2\pi|\Gamma|}} \quad (3)$$

G_k is GIST representation of the member of cluster $\mathbf{J_k}$ of scene-type C_k which is the closest match to θ of the test image at hand. Γ is covariance matrix, assumed to be of form γI where I is identity matrix and γ can be selected empirically over held-out data. We also add a 'general' type to set of scene-types $\mathbf{C}$ and assume that the cluster for this type consists of all training images. Idea is that many words will be specific to the type of scene being presented but some words are generic and appear in descriptions of images presenting varying types of scenes. Processing for the type 'general' will provide evidence for those words. $P(C_k/\theta)$ for test images, where C_k represents 'general' category, is assigned a fixed weight for all images. Overall $P(\mathbf{C}/\theta)$ is then renormalized. Fixed weight can be determined over a held out portion of data.

We have used multiple Bernoulli distribution for modeling image description. Thus, $P_{V_{C_k}}(w_b/J)$ is w_b*th* component of multiple Bernoulli distribution over words in the set V_{C_k} which has generated description for sample J of training data. V_{C_k} is the vocabulary for samples of scene-type C_k. Bayes estimation for this distribution, given beta prior is given by

following formula.

$$P_{V_{C_k}}(w_b/J) = \frac{\mu\delta_{w_b} + N_{w_b k}}{\mu + N_k} \quad (4)$$

where δ_{w_b} is 1 if caption of J contains word w_b, μ is an empirically selected constant, $N_{w_b k}$ is the number of training samples for k*th* scene-type containing word w_b and N_k is the size of cluster representing scene-type C_k.
$P_R(r_a/J)$ is the density estimate for generating visual unit r_a given a training image J. We have assumed a non-parametric kernel based density estimate. Assuming that the training image J consists of A visual units $\{j_1, j_2, ..., j_A\}$ where j_a represents the visual unit at position corresponding to r_a of the test image

$$P_R(r_a/J) = \frac{\exp((r_a - j_a)^T \Sigma^{-1}(r_a - j_a))}{\sqrt{2\pi|\Sigma|}} \quad (5)$$

This equation uses Gaussian density kernel with covariance matrix Σ which can be taken as βI for convenience where I is the identity matrix. β determines smoothness around point j_a and can be empirically selected on held-out set of data. Notice that this kernel signifies importance of spatial coherence while quantifying similarity between two images by comparing visual units at corresponding grid positions only. If index 'a' of visual unit presents some other information e.g. type of visual feature, this kernel would be able to correctly quantify similarity by comparing units of similar type with each other.

4. DATASETS AND EVALUATION

4.1. Datasets

We have evaluated our system on two popular image annotation datasets i.e. IAPR-TC 12 [1] and ESP [2]. IAPR dataset has $19,846$ images, each described carefully in a few sentences. Frequently occurring nouns are picked to form vocabulary set after tokenizing and part-of-speech tagging these sentences. ESP game dataset consists of images labeled by players of ESP game. A smaller subset of size $21,844$ has been popularly used to test different image annotation systems. We have used the same split of data in training and test sets (90% for training, 10% for test) for both datasets as used by other image annotation systems. IAPR and ESP datasets have been generally tested over vocabulary sets of 291 and 269 most frequently occurring words respectively by various image annotation systems. In our system, the vocabulary varies from collection of samples of one scene-type to the other, instead of being fixed to a specific number for all data. But we have made sure that approximately the same number of unique words appear in the final output i.e. annotations predicted for test images, by adjusting parameters of our system. We have reported results over these unique words to keep them comparable to those of other systems. We used approximately 50 scene-type clusters for both datatsets after dropping too small clusters.

[1] http://www.imageclef.org/photodata
[2] www.espgame.org

4.2. Visual Features

We have used 5×6 grid to divide images and have assigned each grid section a feature vector of length 46, representing its color and texture characteristics, similar to many other image annotation systems[4, 5]. Feature vector contains 18 color features (mean and std. deviation of each channel of RGB, LUV and LAB color-spaces), 12 texture features (Gabor energy computed over 3 scales and 4 orientations), 4 bin HoG and discrete cosine transform coefficients. We have observed that increasing the grid size beyond 5×6 does not improve performance. Guillaumin et al. showed performance improvement based on a combination of holistic and local visual features [11]. More recently, Chen et al. and Verma et al. used the same features in their systems [13, 14]. We also tested our system using these features and observed performance improvement.

4.3. Results

Image annotation systems are used to produce as many annotations per image as is the average number of words per image in training data. Mean values of precision and recall per word and number of words with positive recall (N^+) are reported as performance evaluation metrics. **Scene-AIA** represents our system employing grid-based visual features described in section 4.2. **Scene-AIA-B** represents our system making use of visual features devised by Guillaumin et al. [11]. Tables 1 and 2 show performance comparison of our system against many others over IAPR-TC 12 and ESP datasets respectively.

4.4. Observations

Tables 1 and 2 show that our system beats other generative probability estimation based methods such as CRM, MBRM, BS-CRM and many recently proposed methods such as JEC, Lasso, HGDM, AP. MBRM-G represents MBRM[5] using visual features described by Guillaumin et al.[11]. These features include GIST description. Our system beats MBRM-G, indicating that our approach of using scene understanding as contextual information works better than using it to replace low-level visual features while estimating generative probability. As discussed earlier, generative probability estimation based systems are computationally efficient and scalable to larger datasets. Nearest-neighbor based approaches such as TagProp and 2PKNN-ML employ iterative optimization algorithms, rendering them computationally expansive and not particularly scalable to larger datasets. Chen et al. proposed FastTag to reduce computational complexity and presented a detailed complexity analysis of different annotation systems

[13]. Proposed system is similar to generative probability estimation methods in terms of computational complexity and uses one-pass over training data to predict annotations. Clustering based on scene-type is only required for training sample and can be pre-computed using efficient clustering algorithms which practically have less time complexity because of better termination condition than that of iterative optimization algorithms used by TagProp or 2PKNN-ML. The system still beats TagProp, FastTag and 2KPNN-ML in terms of %-age mean precision and is comparable in terms of %-age mean recall against FastTag and TagProp-ML.

	%age mean Precision	%age mean Recall	N^+
CRM[4]	21	15	214
MBRM[5]	21	14	186
MBRM-G[11]	24	23	223
BS-CRM[6]	22	24	250
JEC[22]	25	16	196
Lasso[22]	26	16	199
HGDM [23]	29	18	–
AP[24]	28	26	–
TagProp-ML[11]	48	25	227
TagProp[11]	46	35	266
FastTag[13]	47	26	280
2PKNN-ML[14]	54	37	278
Scene-AIA	55	20	254
Scene-AIA-B	56	25	230

Table 1. Performance evaluation for IAPR-TC-12 dataset

	%age mean Precision	%age mean Recall	N^+
CRM[4]	29	19	227
MBRM[5]	21	17	218
MBRM-G[11]	18	19	209
JEC[22]	23	19	227
Lasso[22]	22	18	225
AP[24]	24	24	–
TagProp-ML[11]	49	20	213
TagProp[11]	39	27	239
FastTag[13]	46	22	247
2PKNN-ML[14]	53	27	252
Scene-AIA	45	19	246
Scene-AIA-B	60	20	234

Table 2. Performance evaluation for ESP-game dataset

4.5. Cluster Expansion for Large Datasets

To prove the scalability of our system, we tested it over complete ESP dataset, referred to as ESP-large in this paper. This dataset has 67796 image-description pairs (90% dataset for training, 10 % for testing). We used grid based visual features and reported results over set of 1400 unique words. We generated roughly 200 clusters based on scene-type using efficient implementation of K-means clustering[3], assuming that larger dataset contains greater variety of scene types. We compared the performance of our system against MBRM which is a member of the set of methods employing generative probability estimation and is computationally very efficient. Table 3 shows that our system beats MBRM for ESP-large dataset; proving that our system is scalable for larger datasets with vast vocabulary.

	%-age mean Precision	%-age mean Recall	N^+
MBRM	34	15	770
Scene-AIA	47	24	979
Scene-AIA-exp	44	23	901

Table 3. Performance evaluation for ESP-large dataset

We tried another variant of our system, named **Scene-AIA-exp** to reduce computational complexity even further. We split the training data in two halves and used the clustering algorithm on one half of training data. Then we expanded the clusters by adding each image from the other half of training data to the cluster containing its closest match based on GIST features. Thus, even the computationally efficient clustering algorithm needs to be run over only half of training data. Only slight reduction in performance is observed. This also indicates that our system is flexible enough to make use of additional training data as it becomes available without having to start training process from scratch.

5. CONCLUSION AND FUTURE WORK

We have proposed a novel approach for automatic image annotation which takes into account similarity between holistic representation of the image i.e. the type of scene presented in the image, to predict annotations for that image. We have shown through detailed experiments that the proposed system is scalable and computationally efficient. It achieves impressive performance even in comparison to more expansive iterative optimization based methods. Scene-type basically induces a measure of background knowledge or context in the process of identifying details of an image. In future, we intend to incorporate other sources of such information e.g. meaningful album names such as 'graduation', 'birthday photos' provided by user while uploading images on social media websites or keywords and meta-data assigned to news pieces with images. We intend to properly formalize use of such information in automatic image annotation system.

[3] http://www.vlfeat.org/

6. REFERENCES

[1] Ronald A Rensink, J Kevin O'Regan, and James J Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psychological Science*, 1997.

[2] Aude Oliva and Philippe G Schyns, "Diagnostic colors mediate scene recognition," *Cognitive psychology*, vol. 41, no. 2, 2000.

[3] Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*.

[4] Victor Lavrenko, R Manmatha, and Jiwoon Jeon, "A model for learning the semantics of pictures," in *Advances in neural information processing systems*.

[5] SL Feng, Raghavan Manmatha, and Victor Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[6] Sean Moran and Victor Lavrenko, "Optimal tag sets for automatic image annotation," in *Proceedings of the British Machine Vision Conference*, 2011.

[7] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, 2003.

[8] Duangmanee Putthividhy, Hagai Thomas Attias, and Srikantan S Nagarajan, "Topic regression multi-modal latent dirichlet allocation for image annotation," in *CVPR*, 2010.

[9] Stephen Roller and Sabine Schulte im Walde, "A multimodal lda model integrating textual, cognitive and visual modalities," 2013.

[10] Oksana Yakhnenko and Vasant Honavar, "Annotating images and image objects using a hierarchical dirichlet process model," in *Proceedings of the 9th International Workshop on Multimedia Data Mining: held in conjunction with the ACM SIGKDD 2008*.

[11] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *IEEE 12th International Conference on Computer Vision*, 2009.

[12] Xirong Li, Cees GM Snoek, and Marcel Worring, "Learning social tag relevance by neighbor voting," *IEEE Transactions on Multimedia*, vol. 11, no. 7, 2009.

[13] Minmin Chen, Alice Zheng, and Kilian Q Weinberger, "Fast image tagging," 2013.

[14] Yashaswi Verma and CV Jawahar, "Image annotation using metric learning in semantic neighbourhoods," in *Computer Vision–ECCV 2012*. 2012.

[15] Chaoran Cui, Jun Ma, Tao Lian, Xiaofang Wang, and Zhaochun Ren, "Ranking-oriented nearest-neighbor based method for automatic image annotation," in *ACM SIGIR conference on Research and development in information retrieval*, 2013.

[16] Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li, "Flickr distance: a relationship measure for visual concepts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[17] Adrian Ulges, Marcel Worring, and Thomas Breuel, "Learning visual contexts for image annotation from flickr groups," *IEEE Transactions on Multimedia*, 2011.

[18] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daum III, "Midge: Generating image descriptions from computer vision detections," in *EACL*, 2012.

[19] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg, "Im2text: Describing images using 1 million captioned photographs," in *Advances in Neural Information Processing Systems*.

[20] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg, "Baby talk: Understanding and generating simple image descriptions," in *CVPR*, 2011.

[21] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, 2001.

[22] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar, "A new baseline for image annotation," in *Computer Vision–ECCV 2008*. 2008.

[23] Zhixin Li, Zhongzhi Shi, Weizhong Zhao, Zhiqing Li, and Zhenjun Tang, "Learning semantic concepts from image database with hybrid generative/discriminative approach," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 9, 2013.

[24] Michael Rubinstein, Ce Liu, and William T Freeman, "Annotation propagation in large image databases via dense image correspondence," in *Computer Vision–ECCV 2012*. 2012.