Image Auto-annotation using ‘Easy’ and ‘More Challenging’ Training Sets

Jiayu Tang and Paul H. Lewis

Intelligence, Agents, Multimedia Group, School of Electronics and Computer Science,
University of Southampton, Southampton, SO17 1BJ, United Kingdom
e-mail: {jt04r, phl}@ecs.soton.ac.uk

Abstract The Corel Image set [1] is widely used for image annotation performance evaluation although it has been claimed [2] that the set is easy to annotate. The aim of this paper is to demonstrate some of the disadvantages of sets like the Corel set for effective auto-annotation evaluation. We first compare the performanace of several annoatation algorithms using the Corel set and find that simple near neighbour propagation techniques perform almost as well as the best of the more sophisticated algorithms. We then build a new image collection using the Yahoo Image Search engine[1] and query-by-single-word searches to create a more challenging annotated set automatically. Then, using two very different image annotation methods, we demonstrate some of the problems of annotation using the Corel set compared with the Yahoo based training set. In both cases the training sets are used to create a set of annotations for the Corel test set. Finally we show how self-annotation can be used to improve the original annotations of our Yahoo set.

1 Introduction

Image auto-annotation has been drawing more and more attention in recent years, not only because it turns the traditional content-based image retrieval problem into a standard text retrieval problem by attaching annotations to images, but also it is a form of pattern recognition (either region based or image based) since it predicts words that describe objects in the images.

Semantic propagation and statistical inference are two image auto-annotation approaches. Propagation is a supervised learning technique that compares image similarity at a purely visual level and then annotates images by propagating keywords over the most similar images [3,4]. Statistical inference is an unsupervised learning method that tries to capture the association between visual features and keywords by estimating their joint probability distribution [1,5–8].

The Corel Image set [1] has been widely used for image auto-annotation evaluation. However, as people [2] have argued, the Corel set is easy to annotate. We built a new image collection that is more difficult, by obtaining images from the Yahoo Image Search. Two auto-annotation methods, a propagation method and a statistical correlation method, are applied to the two image sets. In addition, a new way of experimentation on image auto-annotation, using the Yahoo set itself as the training set to improve its own annotations, is proposed.

2 Two Image Collections

2.1 The Corel Set

We use the Corel Image set provided by [1] which is already separated into a training set with 4500 images and a test set with 500 images. Most of the images have 4 word annotations, while a few have 1, 2, 3 or 5. The vocabulary size of the whole set is 374 and that of the test set is 263. It will be shown in Section 5 that the simple CSD-based propagation method (detailed in Section 3) achieves very good result, compared with the state-of-the-art methods, on this image set. This is probably because the Corel images are easy to annotate. Many of the images are very close to each other in terms of both low-level features (such as color) and semantics and thus have the same combination of keywords as their annotations. A query image can be annotated correctly if there exists a training image that is very similar (both at the low-level and semantically) and meanwhile is chosen for propagation.

2.2 The Yahoo Set

We created an image collection of 5260 images by querying the Yahoo Image Search engine using each of the the

[1] http://images.yahoo.com

Table 1 Illustration of propagation for the CSD-Prop method

Image Index	Captions		
1	a	b	c
2	b	d	e
3	a	b	d

263 keywords from the Corel test set [1], such as 'water', 'sky' and 'people'. For each keyword, the top 20 images returned by Yahoo are adopted and annotated with the single query keyword used to retrieve it. In some cases these annotations were not particularly appropriate becasuse of the text based nature of the Yahoo image search. All images are JPG color images, with a resolution of 120x80 on average. It is also a more challenging set because, unlike the Corel set, the collection is less likely to contain groups of images with very similar content. The implication is that training with the Yahoo set will be more difficult than for Corel.

3 Two Auto-annotation Methods

We used two very different ways of image auto-annotation for the main comparisons. The first is a propagation method based on global feature vectors and the second a more complex region based method using correlation statistics.

3.1 The CSD-Prop Method

Propagation methods [3,4] work by propagating annotations from the most similar images in the training set. In this work, the MPEG-7 Colour Structure Descriptor (CSD) [9] is used as the feature descriptor to rank the training images. The similarity between images are measured by the CSD distance (squared euclidean). For each test image, propagation starts from the top training image and goes on until a desired number of different annotations are found. Because the number of predicted words for a test image is fixed, sometimes only a portion of the annotations of a training image can be used. When it is the case, the choice is made randomly. For example, if the top 3 training images in the ranked list for a test image have the captions as showed in Table 1 and 4 words need to be predicted, they are either 'a', 'b', 'c', 'd', or 'a', 'b', 'c', 'e'.

3.2 The SvdCos Method

The region based SvdCos Method is proposed by Pan *et al.* [8] and uses the blob representation proposed by [1]. Follow [8]'s derivation, the SvdCos method works as follows. Suppose there are N_W words in the vocabulary and N_B blobs in the visual vocabulary, the whole training set $I = \{I_1, ..., I_{N_I}\}$ can be represented by a matrix $D_{[N_I - by - (N_W + N_B)]}$, where $D = [D_W | D_B]$. The (i, j)-element of D_W is the count of word w_j in image I_i, and the (i, j)-element of D_B is the count of blob b_j in image I_i. This method captures the association between words and blobs through their pattern of occurrence over the whole image set, which is represented by each column of D_W and D_B. A translation table $T_{[N_W - by - N_B]}$ is created. T_{ij} is the cosine value of the ith column vector of D_W and jth column vector of D_B. Each column of T is normalized to be added up to 1. Thus, T_{ij} can be treated as the probability of translation between word w_i and blob b_j.

Singular Value Decomposition (SVD) decomposes a matrix $X_{[n-by-m]}$ into a product of three matrices U, Λ and V^T, where U and V are orthonormal, and Λ is diagonal. Previous works [10] show that by eliminating small diagonal values of Λ, "SVD could be used to clean up noise and reveal informative structure" ([8]) in X. Therefore, SVD is applied to D before constructing the translation table. Given a test image, which is represented by $q = \{q_1, ..., q_{N_B}\}$ (where q_i is the count of blob b_i), it can be annotated by choosing the words that have the highest values in p, where $p = Tq$.

Details of this method can be found in [8].

4 Evaluation Metrics

The *Mean Per-word Precision and Recall* and *Keyword Number with Recall>0*, as used by previous researchers [1,6,7,11], are adopted for evaluating annotation effectiveness. Per-word precision is defined as the number of images correctly predicted with a given word, divided by the total number of images predicted with this word. Per-word recall is defined as the number of images correctly predicted with a given word, diveded by the total number of images having this word in its ground-truth or manual annotations. Per-word precision and recall values are averaged over the set of test words to generate the mean per-word precision and recall. A keyword has recall>0 if it is predicted correctly once or more, otherwise not.

We also introduced *Mean Per-image Precision and Recall* and *Cumulative Correct Annotations* for evaluation. Per-image precision is the number of correctly predicted words of a given image divided by the number of total words predicted for that image, and per-image recall is the number of correct words divided by the number of manual annotations for that image. Per-image precision and recall are averaged over the whole test images to get the mean per-image precision and recall. Cumulative Correct Annotations is the total number of correct annotations.

5 Results and Discussion

5.1 Comparison with state-of-the-art methods

We applied the two methods to the Corel set. Table 2 compares the CSD-Prop and SvdCos methods with some state-of-the-art methods when the Corel training set is trained to annotate the Corel test set; specifically the Translation model [1], the CRM model [6], the MBRM model [7], and the Mix-Hier model [11]. It is interesting to note that the simple CSD-Prop method achieves a result almost as good as the best results from the more advanced methods.

5.2 Comparison between the two methods when different training sets are used.

For each of the two methods, we used the Corel training set and the Yahoo set for training respectively, to annotate the Corel test set. However, for fair comparison, only one random word out of the complete set of captions (normally 4) is used for each Corel training image, since each Yahoo image has only one caption. Table 3 compares the two methods using the two different image sets for training.

It can be seen that the CSD-Prop method performs better than the SvdCos method when it is trained on the Corel training set, but worse than the SvdCos method when trained on the Yahoo set. In other words, the CSD-Prop method degrades more rapidly when it moves from an easy training set to a more difficult one. Moreover, even though only about 1/4 of the annotations of the Corel training set are used, both methods still achieve relatively good results when compared with the methods refered to in Table 2, where all the annotations are used. We conclude that it is relatively easy to annotate the Corel test set using the Corel training set, and that the CSD-Prop method does not transfer as well as the SvdCos method to the more challenging Yahoo dataset. It could be argued that a good auto-annotation approach should perform at least as well as, if not better than, propagation-based approaches. Finally we conclude that simple sets like the Corel set should be used with caution for effective annotation evaluation.

5.3 Self re-annotation to improve the Yahoo training set.

Since we used query-by-single-word searches to obtain Yahoo images, there is only one annotation available for each Yahoo image. Besides, becasuse of the text based nature of the Yahoo image search, some annotations are not particularly appropriate. It is meaningful to improve the annotations. Under the assumption that most of the images are correctly annotated, we believe the annotations can be improved by utilizing the Yahoo set itself. In other words, the Yahoo set is trained to re-annotate itself, in order to correct wrong annotations, and also find un-annotated objects by predicting more words. Some initial experiments, in which the SvdCos method is adopted, show this is a promising approach. As shown in Figure 1, the initial image labels are significantly improved if the most probable word (in bold) is considered. For images at the bottom, more objects are annotated correctly. This new way of improving image annotations is also worth adopting for testing the robustness of auto-annotation methods against missing information (only one word for images with multiple objects) and noise (images with wrong annotations).

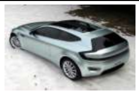
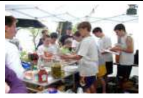
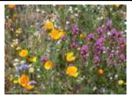
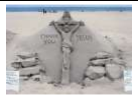
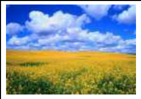
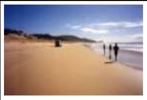
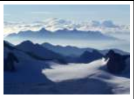
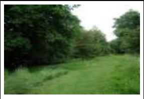
Correcting Wrong Annotations				
Images				
Label	jet	water	mountain	sand
Re-annotations	**cars** boats train pool house	**people** statue horses sculpture street	**flowers** plants birds people train	**stone** sculpture sand wall leaf
Finding Hidden Objects				
Images				
Label	sky	beach	mountain	grass
Re-annotations	sky flowers clouds mountain hills	sky clouds desert beach mountain	clouds sky snow hills mountain	forest leaf field scotland grass

Fig. 1 Label is the keyword through which the image is obtained from Yahoo. Re-annotations (5 for each image) are predicted by the SvdCos method. Using just the most probable re-annotation word (*in bold*) shows improvement over the original annotation.

6 Conclusion and Future Works

Two image auto-annotation methods, a propagation method and a correlation method, are applied to annotate the Corel test set, by training on two different training sets, the Corel training set and the Yahoo set. The Yahoo set is constructed by obtaining images from Yahoo Image Search through 263 query words. The results show that the Corel set is easy to annotate, especially for the simple propagation method which achieves a result as good as the best results found elsewhere. As the Corel set has been popular for experiments on image auto-annotation, we recommend that researchers be aware of the disadvantages of data sets like the Corel Image set for effective annotation evaluation. The results also show that the propagation method does not transfer as well as the correlation method over different data sets.

In addition, self re-annotation is used to improve the captions of the Yahoo images. We tried to correct wrong

Table 2 Comparison between the CSD-Prop method, the SvdCos method and some other state-of-the-art methods using the Corel images

Models	Translation	CRM	MBRM	Mix-Hier	CSD-Prop	SvdCos
words with recall>0	49	107	122	137	130	102
Results on 49 best words						
Mean Per-word Recall	0.34	0.70	0.78	–	0.80	0.59
Mean Per-word Precision	0.20	0.59	0.74	–	0.58	0.51
Results on all 263 words						
Mean Per-word Recall	0.04	0.19	0.25	0.29	0.27	0.15
Mean Per-word Precision	0.06	0.16	0.24	0.23	0.20	0.15

Table 3 Comparison between the two methods on different training sets

Training Set	Corel(4500)		Yahoo(5260)	
Test Set	Corel(500)			
Models	CSD-Prop	SvdCos	CSD-Prop	SvdCos
words with recall>0	107	100	46	58
Results on all 263 words				
Mean Per-word Recall	0.19	0.15	0.053	0.057
Mean Per-word Precision	0.14	0.11	0.038	0.040
Results on all 500 test images				
Cumulative Correct Annotations	577	349	102	123
Mean Per-image Recall	0.327	0.196	0.058	0.069
Mean Per-image Precision	0.231	0.140	0.040	0.049

annotations and predict non-annotated objects for the Yahoo set, by utilizing the informations exist in itself. Initial attempt shows promises. This new way of experimentation is worth adopting for testing the robustness of auto-annotation methods against the low quality of data-sets like the Yahoo set.

Statistical measures for the experiments that use the Yahoo set itself to improve the annotations need to be addressed for comparison purpose in future work. Most of the current auto-annotation methods, such as the SvdCos method [8] used in this work, tend not to consider information missing and noise in the data set. The problem of fitting them to data sets like the Yahoo set need to be solved.

References

1. P. Duygulu, K. Barnard, J.F.G de Freitas and D.A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," *The Seventh European Conference on Computer Vision*, Copenhagen, Denmark, 2002.
2. T. Westerveld and A.P.de Vries, "Experimental Evaluation of a Generative Probabilistic Image Retrieval Model on 'Easy' Data," *Proceedings of SIGIR Multimedia Information Retrieval Workshop 2003*, Aug, 2003.
3. F. Monay and D. Gatica-Perez, "On Image Auto-Annotation with Latent Space Models," *Proceedings of the 7th ACM international conference on Multimedia*, 2003.
4. J. S. Hare and P. H. Lewis, "Saliency-based Models of Image Content and their Application to Auto-Annotation by Semantic Propagation," *Proceedings of Multimedia and the Semantic Web / European Semantic Web Conference 2005*, 2005.
5. K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei and M. I. Jordan, "Matching Words and Pictures," *Journal of Machine Learning Research*, vol. 3. pp. 1107-1135, 2003.
6. J. Jeon, V. Lavrenko and R. Manmatha, "Automatic Image Annotation and Retrieval using Cross-Media Relevance Models," *SIGIR '03 Conference*, pp. 119-126, 2003.
7. S. L. Feng, R. Manmatha and V. Lavrenko, "Multiple Bernoulli Relevance Models for Image and Video Annotation," *Proceedings of IEEE CVPR*, vol. 2, 2004.
8. Jia-Yu Pan, Hyung-Jeong Yang, P. Duygulu and C. Faloutsos, "Automatic Image Captioning," *Proceedings of the 2004 IEEE ICME*, 2004.
9. J. M. Martinez, *MPEG-7 Overview*, N6828 ISO/IEC JTC1/SC29/WG11, Oct, 2004.
10. G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter and K. E. Lochbaum, "Information retrieval using a singular value decomposition model of latent semantic structure," *SIGIR '88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 465-480, ACM Press, Grenoble, France, New York, USA, 1988.
11. C. Carneiro and N. Vasconcelos, "Formulating Semantic Image Annotation as a Supervised Learning Problem," *CVPR (2)*, pp. 163-168, 2005.