Automatic Image Captioning[*]

Jia-Yu Pan[†], Hyung-Jeong Yang[†], Pinar Duygulu[‡] and Christos Faloutsos[†]
[†]*Computer Science Department, Carnegie Mellon University, Pittsburgh, U.S.A.*
[‡]*Department of Computer Engineering, Bilkent University, Ankara, Turkey*
[†]*{jypan, hjyang, christos}@cs.cmu.edu,* [‡]*duygulu@cs.bilkent.edu.tr*

Abstract

In this paper, we examine the problem of automatic image captioning. Given a training set of captioned images, we want to discover correlations between image features and keywords, so that we can automatically find good keywords for a new image. We experiment thoroughly with multiple design alternatives on large datasets of various content styles, and our proposed methods achieve up to a 45% relative improvement on captioning accuracy over the state of the art.

1. Introduction and related work

"Given a large image database, find images that have tigers. Given an unseen image, find terms which best describe its content." These are some of the problems that many image/video indexing and retrieval systems deal with (see [4][5][10] for recent surveys). Content based image retrieval systems, matching images based on visual similarities, have some limitations due to the missing semantic information. Manually annotated words could provide semantic information, however, it is time consuming and error-prone. Several automatic image annotation (captioning) methods have been proposed for better indexing and retrieval of large image databases [1][2][3][6][7].

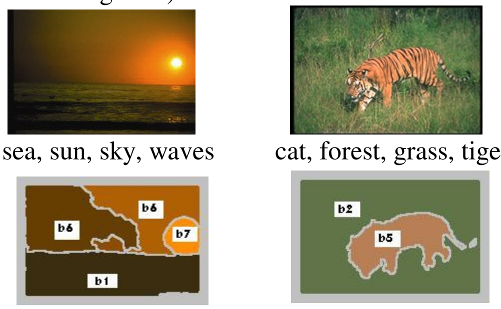
We are interested in the following problem: "*Given a set of images, where each image is captioned with a set of terms describing the image content, find the association between the image features and the terms*". Furthermore, "*with the association found, caption an unseen image*". Previous works caption an image by captioning its constituting regions, by a mapping from image regions to terms. Mori *et al.* [10] use co-occurrence statistics of image grids and words for modeling the association. Duygulu *et al.* [3] view the mapping as a translation of image regions to words, and learn the mapping between region groups and words by an EM algorithm. Recently, probabilistic models such as cross-media relevance model [6] and latent semantic analysis (LSA) based models [11] are also proposed for captioning.

In this study, we experiment thoroughly with multiple design alternatives (better clustering decision; weighting image features and keywords; dimensionality reduction for noise suppression) for better association model. The proposed methods achieve a 45% relative improvement on captioning accuracy over the result of [3], on large datasets of various content styles.

The paper is organized as follows: Section 2 describes the data set used in the study. Section 3 describes an adaptive method for obtaining image region groups. The proposed *uniqueness* weighting scheme *and correlation-based image captioning* methods are given in Section 4 and 5. Section 6 presents the experimental results and Section 7 concludes the paper.

2. Input representation

We learn the association between image regions and words from manually annotated images (examples are shown in Figure 1).



sea, sun, sky, waves cat, forest, grass, tiger

w6 w7 w8 w1 w2 w9 w10 w11

Figure 1. ***Top:*** *annotated images with their captions,* ***bottom:*** *corresponding blob-tokens and word tokens.*

[*] This material is based upon work supported by the National Science Foundation under Grants No. IRI-9817496, IIS-9988876, IIS-0113089, IIS-0209107, IIS-0205224, INT-0318547, SENSOR-0329549, EF-0331657, by the Pennsylvania Infrastructure Technology Alliance (PITA) Grant No. 22-901-0001, and by the Defense Advanced Research Projects Agency under Contract No. N66001-00-1-8936. Additional funding was provided by donations from Intel, and by a gift from Northrop-Grumman Corporation.

An image region is represented by a vector of features regarding its color, texture, shape, size and position. These feature vectors are clustered into *B* clusters and each region is assigned the label of the closest cluster center as in [3]. These labels are called **blob-tokens**.

Formally, let $\mathrm{I}=\{\mathrm{I}_1,\ldots,\mathrm{I}_N\}$ be a set of annotated images where each image I_i is annotated with a set of terms $\mathrm{W}_i=\{\mathrm{w}_{i,1},\ldots,\mathrm{w}_{i,Li}\}$ and a set of blob tokens $\mathrm{B}_i=\{\mathrm{b}_{i,1},\ldots,\mathrm{b}_{i,Mi}\}$, where L_i is the number of words, and M_i is the number of regions in image I_i. The goal is to construct a model that captures the association between terms and the blob-tokens, given W_i's and B_i's.

3. Blob-token generation

The quality of blob-tokens affects the accuracy of image captioning. In [3], the blob-tokens are generated using the K-means algorithm on feature vectors of all image regions in the image collection, with the number of blob-tokens, *B*, set at 500. However, the choice of *B*=500 is by no means optimal.

In this study, we determine the number of blob-tokens *B* adaptively using the G-means algorithm [12]. G-means clusters the data set starting from small number of clusters, *B*, and increases *B* iteratively if some of the current clusters fail the Gaussianity test (e.g., Kolmogorov-Smirov test). In our work, the blob-tokens are the labels of the clusters adaptively found by G-means. The numbers of blob-tokens generated for the 10 training set are all less than 500, ranging from 339 to 495, mostly around 400.

4. Weighting by uniqueness

If there are *W* possible terms and *B* possible blob-tokens, the entire annotated image set of *N* images can be represented by a data matrix $\mathbf{D}_{[\mathbf{N\text{-}by\text{-}(W+B)}]}$. We now define two matrices: one is *unweighted*, the other is *uniqueness weighted* as initial data representation.

Definition 1 *(Unweighted data matrix)* Given an annotated image set $\mathrm{I}=\{\mathrm{I}_1,\ldots,\mathrm{I}_N\}$ with a set of terms W and a set of blob-tokens B, the unweighted data matrix $\mathbf{D_0=[D_{W0}|D_{B0}]}$ is a *N-by-(W+B)* matrix, where the *(i,j)*-element of the *N-by-W* matrix $\mathbf{D_{W0}}$ is the count of term w_j in image I_i, and the *(i,j)*-element of the *N-by-B* matrix $\mathbf{D_{B0}}$ is the count of blob-token b_j in image I_i.

We weighted the counts in the data matrix **D** according to the "uniqueness" of each term/blob-token. If a term appears only once in the image set, say with image I_1, then we will use that term for captioning only when we see the blob-tokens of I_1 again, which is a small set of blob-tokens. The more common a term is, the more blob-tokens it has association with, and the uncertainty of finding the correct term-and-blob-token association goes up. The idea is to give higher weight to terms/blob-tokens which are more "unique" in the training set, and low weights to noisy, common terms/blob-tokens.

Definition 2 *(Uniqueness weighted data matrix)* Given an unweighted data matrix $\mathbf{D_0=[D_{W0}|D_{B0}]}$. Let z_j (y_j) be the number of images which contain the term w_j (the blob-token b_j). The weighted data matrix $\mathbf{D=[D_W|D_B]}$ is constructed from $\mathbf{D_0}$, where the *(i,j)*-element of $\mathbf{D_W(D_B)}$, $d_{W(i,j)}$ $(d_{B(i,j)})$, is

$$d_{W(i,j)}=d_{W0(i,j)}\times\log(\frac{N}{z_j}),d_{B(i,j)}=d_{B0(i,j)}\times\log(\frac{N}{y_j}), \qquad (3)$$

where *N* is the total number of images in the set.

In the following, whenever we mention the data matrix **D**, it will be always the weighted data matrix.

5. Proposed methods for image captioning

We proposed 4 methods (**Corr**, **Cos**, **SvdCorr**, **SvdCos**) to estimate a *translation table* **T**, whose *(i,j)*-element can be viewed as $p(w_i|b_j)$, the probability we caption the term w_i, given we see a blob-token b_j.

Definition 3 *(Method **Corr**)* Let $\mathbf{T_{corr,0}=D_W^T D_B}$. The correlation-based translation table $\mathbf{T_{corr}}$ is defined by normalizing each column of $\mathbf{T_{corr,0}}$ such that each column sum up to 1. Note that the *(i,j)*-element of $\mathbf{T}_{\mathbf{cor}r}$ can be viewed as an estimate of $p(w_i|b_j)$.

$\mathbf{T_{corr}}$ measures the association between a term and a blob-token by the co-occurrence counts. Another possible measure could be to see how similar the overall occurrence pattern (over the training images) of a term and a blob-token is. Such occurrence patterns are in fact the columns of $\mathbf{D_W}$ or $\mathbf{D_B}$, and the similarity can be taken as the cosine value between pairs of column vectors.

Definition 4 *(Method **Cos**)* Let the *i*-th column of the matrix $\mathbf{D_W}$ $\mathbf{(D_B)}$ be $d_{Wi}(d_{Bi})$. Let $cos_{i,j}$ be the cosine value of the angle column vectors d_{Wi} and d_{Bj}, and let $\mathbf{T_{cos,0}}$ be a *W-by-B* matrix whose *(i,j)*-element $\mathbf{T_{cos,0}}(\mathrm{i,j})=cos_{i,j}$. Normalize the columns of $\mathbf{T_{cos,0}}$ such that each column sums up to 1, and we get the cosine-similarity translation table $\mathbf{T_{cos}}$.

Singular Value Decomposition (SVD) decomposes a given matrix $\mathbf{X}_{[\mathbf{nxm}]}$ into a product of three matrices $\mathbf{U}$, $\mathbf{\Lambda}$, $\mathbf{V^T}$. That is, $\mathrm{X}=\mathbf{U\Lambda V^T}$, where $\mathbf{U=[u_1,\ldots,u_n]}$, and $\mathbf{V=[v_1,\ldots,v_m]}$ are orthonormal, and $\mathbf{\Lambda}$ is a diagonal matrix. Note that $\mathbf{u_i}(\mathbf{v_i})$ are columns of the matrix $\mathbf{U}(\mathbf{V})$. Let $\mathbf{\Lambda}=\mathrm{diag}(\sigma_1,\ldots,\sigma_{\min(n,m)})$, then $\sigma_j > 0$, for $j \leq \mathrm{rank(X)}$, $\sigma_j=0$, for $j > \mathrm{rank(X)}$.

Previous works [14] show that by setting small σ_j to zero, yielding an optimal low rank representation $\hat{X}$, SVD could be used to clean up noise and reveal informative structure in the given matrix **X**, and achieve better performance in information retrieval applications. We propose to use SVD to suppress the noise in the data matrix before learning the association. Following the general rule-of-thumb, we keep the first r σ_j's which preserve the 90% variance of the distribution, and set others to zero. In the following, we denote the data matrix after SVD as $\mathbf{D}_{\mathbf{svd}}=[\mathbf{D}_{\mathbf{W,svd}}|\mathbf{D}_{\mathbf{B,svd}}]$.

Definition 5 *(Method **SvdCorr** and **SvdCos**)* Method **SvdCorr** and **SvdCos** generates the correlation-based translation table $\mathbf{T}_{\mathbf{corr,svd}}$ and $\mathbf{T}_{\mathbf{cos,svd}}$ following the procedure outlined in Definition 3 and 4, but instead of starting with the weighted data matrix **D**, here the matrix $\mathbf{D}_{\mathbf{svd}}$ is used.

Algorithm 1 ***(Captioning)*** Given a translation table $\mathbf{T}_{[\mathbf{WxB}]}$ (W: total number of terms; B: total number of blob-tokens), and the number of captioning terms m for an image. An image with l blob-tokens B' = {b'$_1$, …, b'$_l$}, can be captioned by: First, form a query vector $\mathbf{q}=[q_1, \ldots, q_B]$, where q_i is the count of the blob-token b_i in the set B'. Then, compute the **term-likelihood vector** $\mathbf{p}=\mathbf{Tq}$, where $\mathbf{p}=[p_1, \ldots, p_W]^T$, and p_i is the predicted likelihood of the term w_i. Finally, we select the m captioning terms corresponding to the highest m p_i's in the **p** vector.

6. Experimental results

The experiments are performed on 10 Corel image data sets. Each data set contains about *5200* training images and *1750* testing images. The sets cover a variety of themes ranging from urban scenes to natural scenes, and from artificial objects like jet/plane to animals. Each image has in average 3 captioning terms and 9 blobs.

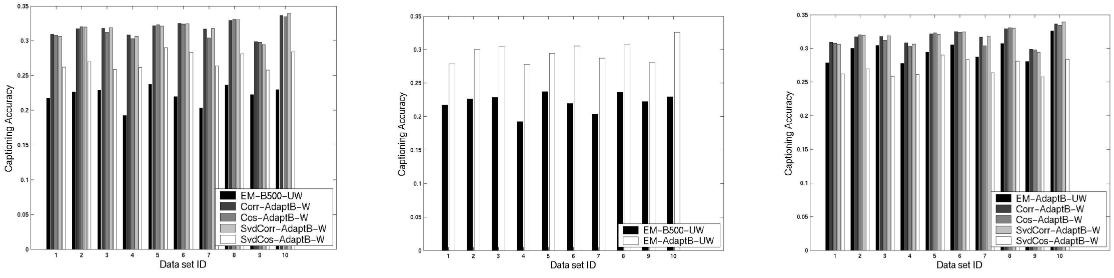
We apply G-means and uniqueness weighting to show the effects of clustering and weighting. We compare our proposed methods, namely **Corr, Cos, SvdCorr** and **SvdCos,** with the state-of-the-art machine translation approach [3] as the comparison baseline. For each method, a translation table, an estimate for the conditional probability of a term w_i given a blob-token b_j ($p(w_i|b_j)$), is constructed. These translation tables are then used in **Algorithm 1**.

The captioning accuracy on a test image is measured as the percentage of correctly captioned words [1]. The captioning accuracy is defined as $S = m_{correct}/m$, where $m_{correct}$ (m) is the number of the correctly (truth) captioned terms. The overall performance is expressed by the average accuracy over all images in a (test) set.

Figure 2(a) compares the proposed methods with the baseline algorithm [3] which is denoted as **EM-B500-UW** (which means **EM** is applied to an unweighted matrix, denoted **UW**, in which the number of blob tokens is 500, denoted as **B500**). For the proposed methods, blob-tokens are generated adaptively (denoted **AdaptB**) and uniqueness weighting (denoted **W**) is applied. The proposed methods achieve an improvement around 12% absolute accuracy (45% relative improvement) over the baseline.

The proposed adaptive blob-token generation could also improve the baseline **EM** method. Figure 2(b) shows that the adaptively generated blob-tokens improve the captioning accuracy of the EM algorithm. The improvement is around 7.5% absolute accuracy (34.1% relative improvement) over the baseline method (whose accuracy is about 22%). In fact, we found that the improvement is not only on the **EM** method, but also on our proposed methods. When the number of blob-tokens is set at 500, proposed methods are 9% less accurate (detail figures not shown). This suggests that the correct size of blob-token set is not 500, since all methods perform worse when the size is set at 500.

Before applying the "uniqueness" weighting, the 4 proposed methods perform similar to the baseline EM method (accuracy difference is less than 3%). The uniqueness weighting improves the performance of all proposed methods except **Cos** method, which stays put. We also observed that weighting does not affect the result of **EM**. Due to the lack of space, we do not show detail figures here.



(a) (b) (c)

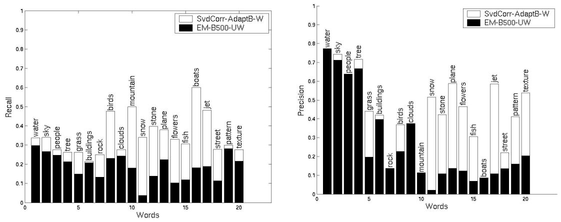
Figure 2*. Captioning accuracy improvement **(a)** proposed methods vs. the baseline* **EM-B500-UW**, ***(b) a****daptive blob-token generation on* **EM** *vs. the baseline,* ***(c)*** *proposed methods vs.* **EM** *when the adaptively generated blob-tokens are used.*

Another measurement of the performance of a method is the recall and precision values for each word (Figure 3). Given a word w, let the set R_w contains *r* test images captioned with the word w by the method we are evaluating. Let *r** be the actual number of test images that have the word w (set R^*_w), and *r'* be size of the intersection of R_w and R^*_w. Then, the precision of word w is *r'/r*, and the recall is *r'/r**.

Note that some words have zero precision and recall, if they are never used or are always used for the wrong image (un-"*predictable"* words). We prefer a method that has fewer unpredictable words, since it could generalize better to unseen images. Table 1 shows that the proposed methods have two to three times more predictable words on average than **EM** does. **EM** captions frequent words with high precision and recall, but misses many other words. That is, **EM** is biased to the training set.

Table 1. *Average recall and precision values and the number of predictable words.*

	EM	**Corr**	**Cos**	**SvdCorr**	**SvdCos**
# predicted	36	57	72	56	132
Avg recall	0.0425	0.1718	0.1820	0.1567	0.2128
Avg prec.	0.0411	0.1131	0.1445	0.1197	0.2079



Recall Precision

Figure 3. *Recall and precision of the top 20 frequent words in the test set.* **SvdCorr** *(white bars) gives more general performance than* **EM** *(black bars).*

As an example of how well the captioning is, for the image in Figure 1(a), **EM-B500-UW** and **SvdCorr-AdaptB-W** both give "sky", "cloud", "sun" and "water". As for the image in Figure 1(b), **EM-B500-UW** gives "grass", "rocks", "sky" and "snow", while **SvdCorr-AdaptB-W** gives "grass", "cat", "tiger", and "water". Although the captions do not match the truth (in the figure) perfectly, they describe the content quite well. This indicates that the "truth" caption may be just one of the many ways to describe the image.

7. Conclusion

In this paper, we studied the problem of automatic image captioning and proposed new methods (**Corr**, **Cos**, **SvdCorr** and **SvdCos**) that consistently outperform the state of the art **EM** (45% relative improvement) in captioning accuracy. Specifically,

- We do thorough experiments on 10 large datasets of different image content styles, and examine all possible combinations of the proposed techniques for improving captioning accuracy.
- The proposed "uniqueness" weighting scheme on terms and blob-tokens boosts the captioning accuracy.
- Our improved, "adaptive" blob-tokens generation consistently leads to performance gains.
- The proposed methods are less biased to the training set and more generalized in terms of retrieval precision and recall.

The proposed methods can be applied to other areas, such as building an image glossary of different cell types from figures in medical journals [13].

References

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, M. Jordan, "Matching words and pictures", Journal of Machine Learning Research, 3:1107:1135, 2003.

[2] D. Blei, M. Jordan, "Modeling annotated data", 26th Annual Int. ACM SIGIR Conf., Toronto, Canada, 2003.

[3] P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth, "Object recognition as machine translation: learning a lexicon for a fixed image vocabulary", In Seventh European Conference on Computer Vision (ECCV), Vol. 4, pp. 97-112, 2002.

[4] D. A. Forsyth and J. Ponce, "Computer Vision: a modern approach", Prentice-Hall, 2001.

[5] A. Goodrum, "Image information retrieval: An overview of current research", Informing Science, 3(2), 2000.

[6] J. Jeon, V. Lavrenko, R. Manmatha, "Automatic Image Annotation and Retrieval using Cross-Media Relevance Models", 26th Annual Int. ACM SIGIR Conference, Toronto, Canada, 2003.

[7] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach", IEEE Trans. on Pattern Analysis and Machine Intelligence, 25(10), 2003.

[8] M. Markkula, E. Sormunen, "End-user searching challenges indexing practices in the digital newspaper photo archive", Information retrieval, vol.1, 2000.

[9] Y. Rui, T. S. Huang, S.-F.Chang, "Image Retrieval: Past, Present, and Future", Journal of Visual Communication and Image Representation, 10:1-23, 1999.

[10] Y. Mori, H. Takahashi, R. Oka, "Image to word transformation based on dividing and vector quantizing images with words", First Int. Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999.

[11] F. Monay, D. Gatica-Perez, "On Image Auto-Annotation with Latent Space Models", Proc. ACM Int. Conf. on Multimedia (ACM MM), Berkeley, 2003.

[12] Greg Hamerly and Charles Elkan, "Learning the k in k-means", Proc. of the NIPS 2003.

[13] Velliste, M. and R.F. Murphy, 2002. "Automated Determination of Protein Subcellular Locations from 3D Fluorescence Microscope Images," Proc 2002 IEEE Intl Symp. Biomed Imaging (ISBI 2002), pp. 867-870.

[14] G. W. Furmas, 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," Proc. of the 11th ACM SIGIR conf., pp. 465-480, 1998.