Effective Automatic Image Annotation Via A Coherent Language Model and Active Learning

Rong Jin
Department of Computer
Science and Engineering
Michigan State University
1-517-353-7284
rongjin@cse.msu.edu

Joyce Y. Chai
Department of Computer
Science and Engineering
Michigan State University
1-517-432-9239
jchai@cse.msu.edu

Luo Si
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
1-412-268-3951
lsi@cs.cmu.edu

ABSTRACT

Image annotations allow users to access a large image database with textual queries. There have been several studies on automatic image annotation utilizing machine learning techniques, which automatically learn statistical models from annotated images and apply them to generate annotations for unseen images. One common problem shared by most previous learning approaches for automatic image annotation is that each annotated word is predicated for an image independently from other annotated words. In this paper, we proposed a coherent language model for automatic image annotation that takes into account the word-to-word correlation by estimating a coherent language model for an image. This new approach has two important advantages: 1) it is able to automatically determine the annotation length to improve the accuracy of retrieval results, and 2) it can be used with active learning to significantly reduce the required number of annotated image examples. Empirical studies with Corel dataset are presented to show the effectiveness of the coherent language model for automatic image annotation.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis-Object Recognition

General Terms

Algorithms, Management, Experiment

Keywords

Image annotation, Image retrieval, Expectation-Maximization algorithm, and statistical models

1. INTRODUCTION

Efficient access to multimedia database requires the ability to search and organize multimedia information. In traditional image retrieval, users have to provide examples of images that they are looking for. Similar images are found based on the match of image features. Even though there have been many studies on image retrieval, empirical studies have shown that using image features to find similar images is usually insufficient [20]. One solution is the region-based image retrieval [4], which allows users to specify the regions of interest and the similarity is computed only based on the specified regions. Another alternative is to allow users to pose textual queries against image databases. To support such a capability, automatic image annotation is a critical technique that bridges the gap between image features and textual words.

Previously, there have been many studies on automatic image annotation [2, 3, 5, 6, 8-14]. Many of them applied machine-learning techniques to first learn the correlation between image features and textual words from the examples of annotated images and then apply the learned correlation to predicate words for unseen images. One common problem shared by most approaches for automatic image annotation is that each annotated word for an image is predicated independently from other annotated words for the same image. Correlations between annotated words are not taken into consideration. The word-to-word correlation is important particularly when image features are insufficient in determining an appropriate word annotation. Consider word 'sky' and 'ocean'. Since both words are related to regions with blue color, it is difficult to make a decision between the word 'sky' and 'ocean' based on the color distribution of a region. However, since certain words such as 'grass' and 'horse' are more likely to be correlated with 'sky' than 'ocean', we can use the word-to-word correlation to further determine the annotation word between 'sky' and 'ocean'. For example, if 'grass' is very likely to be an annotated word, 'sky' will usually be preferred over 'ocean' because of the word-to-word correlation.

In this paper, we propose a coherent language model for automatic image annotation that takes into account word-to-word correlation. Instead of predicating each annotated word independently for a given image, this new approach estimates a coherent language model for the image. Compared to previous approaches for automatic image annotation, this new approach has two important advantages:

1) It is able to automatically determine the annotation length for a given image. This is in contrast to most previous approaches that use a fixed annotation length. With automatically determined annotation length, the generated annotation provides a more accurate description for the content of an image without irrelevant words. As a result,

MM'04, October 10–16, 2004, New York, New York, USA.

automatically determined annotation length will significantly improve the accuracy of image retrieval for textual queries.

2) It can be naturally used for active learning to significantly reduce the required number of annotated image examples. Due to the large variance in image features, automatic image annotation usually requires a large number of annotated images as training examples. Active learning for automatic image annotation can significantly reduce the number of training examples by selectively sampling annotated images. In contrast to previous studies which used a fixed set of annotated images for training, our approach supports active learning to significantly reduce training effort and achieve comparable performance.

The rest of this paper is organized as follows: Section 2 discusses the related work on image annotation; Section 3 describes the coherent language model and how to use it for automatically determining annotation length and active learning; Section 4 presents the experimental results; Conclusions are drawn in Section 5.

2. RELATED WORK

A variety of learning methods have been applied to automatic image annotation, including machine translation model [8], co-occurrence model [14], latent space approaches [2, 13], graphic models [3], classification approaches [5, 6, 11], and relevance language models [9, 10]. The co-occurrence model [14] collects the co-occurrence counts between words and image features and uses them to predicate annotated words for images. Duygulu et al. [8] improved the co-occurrence model by utilizing machine translation models. It views image annotation as a process of translation from 'visual language' to texts and collects the co-occurrence information by the estimation of translation probabilities. Another way of capturing co-occurrence information is to introduce latent variables to link image features with words. Standard latent semantic analysis (LSA) and probabilistic latent semantic analysis (PLSA) are applied to automatic image annotation [13]. Barnard et al. [2] introduced a hierarchical aspect model for image annotation in order to account for the fact that some words are more general than others. More sophisticated graphical models, such as Gaussian Mixture Model (GMM), Latent Dirichlet Allocator (LDA), and correspondence LDA, have also applied to the image annotation problem recently [3]. The classification approaches for automatic image annotation treat each annotated word as an independent class and create a different image classification model for every word. Work such as linguistic indexing of pictures [11], image annotation using SVM [6] and Bayes point machine [5] fall into this category.

Recently, relevance language models [9, 10] have been successfully applied to automatic image annotation. The essential idea is to first find annotated images that are similar to a test image and then use the words shared by the annotations of the similar images to annotate the test image. Empirical studies have shown that relevance language models are able to significantly outperform several other approaches for automatic image annotation [9, 10]. Since this model is most related to our approach, we will provide more details below.

Let the collection of annotated images denoted by *T*, and the size of the collection denoted by $|T|$. Each annotated image $J_i \in T$ is represented by its image regions and annotated words. Following the paper [8], we group image regions from all training examples into 500 clusters. These 500 clusters or blobs (as called in [8]) compose the vocabulary for images. As a result, each image J_i is represented by the combination of blobs and words, i.e., $J_i = \{b_{i,1}, b_{i,2,}..., b_{i,m}; w_{i,1}, w_{i,2}, ..., w_{i,n}\}$: m and n are the number of blobs and words, respectively; $b_{i,j}$ is the number of j-th blob that appears in the i-th image; $w_{i,j}$ is a binary variable indicating whether or not the j-th word appears in the i-th image. Given an image $I = \{b_1, b_2, ..., b_m\}$, the key to automatic image annotation is to estimate the likelihood for any word to be annotated for *I*, or $p(w_j = 1 | I)$. According to the relevance model [9], this likelihood is expanded as a sum over all annotated images, i.e.,

$$\begin{aligned} p(w_k = 1 | I) \propto p(w_k = 1, I) &= \sum_{i=1}^{|T|} p(w_k = 1, I, J_i) \\ &= \sum_{i=1}^{|T|} p(J_i) p(I | J_i) p(w_k = 1 | J_i) \\ &\approx \frac{1}{|T|} \sum_{i=1}^{|T|} p(w_k = 1 | J_i) \prod_{j=1}^{m} p(b_j | J_i) \end{aligned} \quad (1)$$

where $p(J_i)$ is assigned with a uniform distribution.. Both $p(w_j = 1 | J_i)$ and $p(b_j | J_i)$ are assumed to be multinomial distributions and are computed as follows:

$$\begin{aligned} p(w_j = 1 | J_i) &= (1-\alpha) \frac{w_{i,j}}{\sum_{k=1}^{n} w_{i,k}} + \alpha \frac{\sum_{l=1}^{|T|} w_{l,j}}{\sum_{l=1}^{|T|} \sum_{k=1}^{n} w_{l,k}} \\ p(b_j | J_i) &= (1-\beta) \frac{b_{i,j}}{\sum_{k=1}^{m} b_{i,k}} + \beta \frac{\sum_{l=1}^{|T|} b_{l,j}}{\sum_{l=1}^{|T|} \sum_{k=1}^{m} b_{l,k}} \end{aligned} \quad (2)$$

where α and β are both smoothing constants. Finally, words with the largest probabilities $p(w_j = 1 | I)$ are used to annotate image *I*.

Note that what is described above is the "blob version" relevance language model, which uses blobs (centroids of clusters) instead of image features to represent image regions. Although using extracted image features can enhance the accuracy of automatic image annotation [10], it will significantly increase the computational cost. Thus, in this paper, we only consider the blob representation of image regions instead of extracted image features.

Finally, there has prior work on multimedia indexing that captures the correlation between different concepts [16]. The main difference is that this paper uses language modeling approaches while the prior work uses a graphical modeling approach.

3. COHERENT LANGUAGE MODEL FOR IMGE ANNOTATION

In this section, we will first describe the coherent language model for image annotation and then apply it to automatically determine the annotation length and incorporate active learning through selective sampling.

3.1 Basic Algorithm

As aforementioned, one important problem with previous approaches for automatic image annotation is that they assume the predication of annotation words for an image is independent from one word to another. More specifically, for the relevance language model described in Section 2, words are predicated based on probability $p(w_j = 1 | I)$ that only captures the correlation between words and image features and the word-to-word correlation is not considered in this model directly.

To incorporate the word-to-word correlation, we need to estimate the probability of annotating image *I* with a set of word $\{w\}$, i.e., $p(\{w\} | I)$. This is contrast to the relevance language model for automatic image annotation, where annotation probability for each word, i.e., $p(w_j = 1 | I)$, is estimated separately. However, directly estimating $p(\{w\} | I)$ is computationally prohibitive because the number of different set of words $\{w\}$ is exponential with respect to the size of the vocabulary. To avoid the computation difficulty and yet still utilize word-to-word correlation information in predicating annotation words, we consider estimating the probability for using a language model θ_w to generate annotation words for image *I*, i.e., $p(\theta_w | I)$. A language model $\theta_w = \{p_1(\theta_w), p_2(\theta_w), ..., p_n(\theta_w)\}$ consists of a set of word probabilities. Each probability $p_j(\theta_w) = p(w_j = 1 | \theta_w)$ determines how likely the j-th word will be used for annotation. Relaxing the estimation of probability for a set of words $\{w\}$ to the estimation of probability for a language model θ_w can simplify the computation dramatically because a word set $\{w\}$ is a set of binary variables while the word probability $p_j(\theta_w)$ in the language model θ_w is a bounded continuous variable. This is analogous to the relaxation of integer programming problems to linear programming problems.

Using the Bayes rule, i.e., $p(\theta_w | I) \propto p(I | \theta_w) p(\theta_w)$, we convert the estimation of $p(\theta_w | I)$ to the estimation of $p(I | \theta_w)$ and $p(\theta_w)$. A Dirichlet prior is assumed for $p(\theta_w)$, i.e., $p(\theta_w) \propto \prod_{j=1}^{n} \left[p_j(\theta_w) \right]^{\alpha_j}$, where $\alpha_j = \delta \frac{\sum_{l=1}^{|T|} w_{l,j}}{\sum_{l=1}^{|T|} \sum_{k=1}^{n} w_{l,k}}$ and δ is a smoothing constant. Following the idea of relevance language model, we expand probability $p(I | \theta_w)$ as a sum over all annotated images:

$$\begin{aligned} p(I | \theta_w) &= \sum_{i=1}^{|T|} p(I, J_i | \theta_w) = \sum_{i=1}^{|T|} p(I | J_i) p(J_i | \theta_w) \\ &= \sum_{i=1}^{|T|} \prod_{j=1}^{m} p(b_j | J_i) \prod_{j=1}^{n} \left[p_j(\theta_w) \right]^{w_{i,j}} \end{aligned} \quad (3)$$

where $p(b_j | J_i)$ is already defined in Equation (2). Putting expressions for $p(I | \theta_w)$ and $p(\theta_w)$ together, we have

$$p(\theta_w | I) \propto \sum_{i=1}^{|T|} \prod_{j=1}^{m} p(b_j | J_i) \prod_{j=1}^{n} \left[p_j(\theta_w) \right]^{w_{i,j} + \alpha_j} \quad (4)$$

With the expression for $p(\theta_w | I)$, our goal is to find an optimal language model θ_w^* that maximizes $p(\theta_w | I)$, i.e., $\theta_w^* = \arg\max_{\theta \in \Theta} p(\theta | I)$. Directly searching for the optimal language model θ_w^* is computationally expensive due to the summation of a large number of products in Equation (4). Instead, we can apply the Expectation-Maximization algorithm [7] to find the optimal solution iteratively. More specifically, in the E- step, the posterior probability $p(J_i | I)$ is estimated for each annotated image J_i as

$$p(J_i | I) = \frac{1}{Z_I} \prod_{j=1}^{m} p(b_j | J_i) \prod_{j=1}^{n} \left[p_j(\theta_w) \right]^{w_{i,j} + \alpha_j} \quad (5)$$

where Z_I is a normalization constant that ensures $\sum_{i=1}^{|T|} p(J_i | I) = 1$. In the M-step, the language model θ_w is re-estimated using the updated posterior probability $p(J_i | I)$:

$$p_j(\theta_w) = \frac{1}{Z_w} \left(\alpha_j + \sum_{i=1}^{|T|} p(J_i | I) w_{i,j} \right) \quad (6)$$

where Z_w is a normalization constant that ensures $\sum_{j=1}^{n} p_j(\theta_w) = 1$. By applying E-step (i.e., Equation (5)) and M-step (i.e., Equation (6)) alternatively, a local optimal solution for language model θ_w is guaranteed to be found. Equation (6) also indicates the impact of the prior on the final language model. The larger the prior δ is, the more similar the resulting $p_j(\theta_w)$ will be to the global language model α_j / δ, thus avoiding the overfitting problem with the EM algorithm. Furthermore, a careful choice of parameter δ will make a good tradeoff between the prediction of common words and the prediction of rare words. In our experiment, δ is determined using cross evaluation.

To further illustrate why this approach is able to incorporate the word-to-word correlation, we can substitute the expression for $p(J_i | I)$ (in Equation (5)) in Equation (6), which results in the following expression for $p_k(\theta_w)$:

$$p_k(\theta_w) = \frac{\alpha_j + \sum_{i=1}^{|T|} \frac{w_{i,j}}{Z_J} \prod_{j=1}^{m} p(b_j | J_i) \prod_{j=1}^{n} \left[p_j(\theta_w) \right]^{w_{i,j} + \alpha_j}}{Z_w}$$

According to the above expression, the estimation of word probability $p_k(\theta_w)$ depends on the estimation of other word probabilities $p_j(\theta_w)$. As a result, the predication of annotation words is no longer independent from one word to another.

In summary, we proposed a new framework for automatic image annotation that estimates the probability for a language model to be used for annotating an image. The word-to-word correlation is explicitly taken into account through the EM algorithm for finding optimal language model for the given image. For late reference, we call this method for automatic image annotation ‘coherent language model’, or **CLM**.

3.2 Determining Annotation Length

Most previous studies for automatic image annotation assume a fixed annotation length for any image. The described CLM can easily accommodate the fixed length annotation. For example, given the fixed annotation length *k*, words with top-*k* largest probability $p_j(\theta_w)$ are selected as annotation. However, the fixed length annotation can result in either insufficient annotations or overly long annotations. When the length is small (e.g., 2), it is likely that some image content is not captured by the annotation. When the length is long (e.g., 6), it is likely that generated annotations contain words that are irrelevant to the content of images. Both of these cases are undesirable to support efficiently access and browse, the image databases. For example, the irrelevant words as a result of over annotation can severely degrade the accuracy of image retrieval. This will be demonstrated in the experiment section. Next we describe how our CLM approach can be augmented to support annotation with a flexible length.

Recall that in the original relevance language model described in Section 2, both blobs and words are modeled by multinomial distributions (see Equation (2)). Since multinomial distribution is used to describe a random variable whose value is integer, it is appropriate to use it for image blobs since each image can have multiple copies of the same objects/regions. However, it is inappropriate to use a multinomial distribution for annotation words because each word is annotated at most once for an image. Because of the binary nature, it would be more appropriate to describe annotation words with Bernoulli distributions than multinomial distributions. As a result, the probability $p(I \mid \theta_w)$ in Equation (3) is modified as follows:

$$p(I \mid \theta_w) = \sum_{i=1}^{|T|}\prod_{j=1}^{m} p(b_j \mid J_i)\prod_{j=1}^{n}\left[p_j(\theta_w)\right]^{w_{i,j}}\left[1-p_j(\theta_w)\right]^{1-w_{i,j}} \qquad (3’)$$

Meanwhile, a Beta distribution is used for prior $p(\theta_w)$, i.e., $p(\theta_w) \propto \prod_{j=1}^{n}\left[p_j(\theta_w)\right]^{\eta_j}\left[1-p_j(\theta_w)\right]^{\mu_j}$ where η_j and μ_j are computed as $\eta_j = \tau\frac{1}{|T|}\sum_{l=1}^{|T|} w_{l,j}$, $\mu_j = \tau - \eta_j$. τ is a smoothing constant and is determined empirically. Then, the expression for $p(\theta_w \mid I)$ is changed accordingly as follows:

$$\begin{aligned} p(\theta_w \mid I) &\propto p(I \mid \theta_w)p(\theta_w) \\ &\propto \sum_{i=1}^{|T|}\prod_{j=1}^{m} p(b_j \mid J_i)\prod_{j=1}^{n}\left[p_j(\theta_w)\right]^{w_{i,j}+\eta_j}\left[1-p_j(\theta_w)\right]^{1-w_{i,j}+\mu_j} \end{aligned} \qquad (4’)$$

Finally, the EM algorithm for finding optimal language model is changed to the following equations:

$$p(J_i \mid I) = \frac{1}{Z_I}\prod_{j=1}^{m} p(b_j \mid J_i)\prod_{j=1}^{n}\left[p_j(\theta_w)\right]^{w_{i,j}+\eta_j}\left[1-p_j(\theta_w)\right]^{1-w_{i,j}+\mu_j} \qquad (5’)$$

$$p_j(\theta_w) = \frac{\eta_j + \sum_{i=1}^{|T|} p(J_i \mid I) w_{i,j}}{\tau + 1} \qquad (6’)$$

Note that different from Equation (6) where $\sum_{j=1}^{n} p_j(\theta_w)$ is enforced to be 1, $\sum_{j=1}^{n} p_j(\theta_w)$ for Equation (6’) is no longer a constant. This indicates that the number of annotation words may vary from one image to another. Similar to the smoothing parameter δ , the prior τ helps avoid the overfitting problem with EM algorithm and balances the tradeoff between the prediction of common words and the prediction of rare words. In our experiment, it is determined by cross validation.

Since the estimated language model is based on Bernoulli distributions, we can use the natural threshold 0.5 for determining if a word should be used for annotation. More precisely, a word is used for annotation if and only if the corresponding probability $p_j(\theta_w) > 0.5$. This is different from our original CLM model (described in Section 3.1), where word probability $p_j(\theta_w)$ only provides a relative measurement about which word is more likely to be used for annotation. For later reference, we call this model ‘coherent language model with flexible length’, or **CLMFL**.

3.3 Active Learning for Automatic Image Annotation

Due to the large variance in image features for same or similar objects, automatic image annotation usually requires a large number of training examples of annotated images. In machine learning, active learning has been shown very effective in reducing the required number of labeled examples. The basic idea of active learning is to selectively sample examples for labeling so that the uncertainty in determining the right model is reduced most significantly. Most supervised learning techniques assume that examples are randomly chosen for users to label. However, random sampling can be very inefficient for learning when many of sampled examples are similar or even identical. The selective sampling strategy used in active learning avoids this problem by choosing examples that are most informative to a statistical model. As a result, the number of training examples can be significantly reduced while a statistical model of good quality is still achieved. Previous studies have shown that active learning is effective for video scene classification [15].

Active learning is usually conducted in an iterative fashion. At each iteration, an active learning method examines the uncertainty in determining appropriate statistical models and chooses the examples for soliciting labeling information that can most effectively reduce its model uncertainty. There have been many studies on active learning. They differ in the strategies for selective sampling. Some active learning methods select the example for which the model is most uncertain about their labels [1, 18]. Others look at the distribution of predication errors for test data [17, 19]. Empirical studies have shown that predication-error based active learning approaches appear to be more effective than the uncertainty-based approaches [17].

Similar to active learning methods used in other applications, the key of active learning to automatic image annotation is how to determine the image example that is most informative to the statistical model. One type of candidates is the image for which the annotation model is most uncertain about its annotation. This uncertainty can be measured by the averaged word probability $p(w_j = 1 | I)$ for all annotation words. More specifically, for each un-annotated image, we apply the CLMFL model to determine its annotation words and compute its averaged word probability $p(w_j = 1 | I)$. The un-annotated image with the least averaged word probability is chosen for users to annotate. In practice, there will be multiple images that the CLMFL model cannot generate any annotation since all the word probabilities are no more than 0.5. In this case, we will randomly choose one of such images for users to annotate.

We want to emphasize that the proposed active learning algorithm is only useful for the CLMFL model and cannot be applied to CLM model or other relevance language models. This is because both the CLM model and relevance language models are unable to automatically determine the annotation length and therefore the averaged word probability for annotation will not accurately reflect the quality of annotations. In the case of fixed length, we can certainly use the average word annotation probabilities for the top *k* words as the indicator for the uncertainty of model. But, the tradeoff is that we can severely overestimate the quality of auto-annotation when *k* is set too small and underestimate the quality of auto-annotation when *k* is set too large.

The above active learning strategy can be refined by further examining the pool of test images. The above strategy selects images for effectively improving the quality of the annotation model in general but does not specifically target on the characteristics of test images. As a result, even though the overall quality of the annotation model is improved with the annotations for selected images, the improvement may not be reflected on the test data. Therefore, it is better to select the images that not only are poorly annotated by the current model but also are similar to test images. In practice, instead of randomly selecting images from the set of images that the current annotation model cannot produce any annotations, this refined strategy will choose the ones that are most similar to the test images. The similarity of any image $J = \{b_1, b_2, ..., b_m\}$ to all test images $\{I_i\}$ is estimated using probability $p(J | \{I_i\})$, which is approximated as follows:

$$p(J | \{I_i\}) \approx \frac{1}{|\{I_i\}|} \sum_{I \in \{I_i\}} p(J | I) = \frac{1}{|\{I_i\}|} \sum_{I \in \{I_i\}} \prod_{j=1}^{m} p(b_j | I)$$

In summary, we propose an active learning strategy for automatic image annotation using the CLMFL model. It computes the average word probability for auto-annotations and selects the one with the least averaged word probability for a user to annotate. Furthermore, the refined strategy finds the image examples that are not only difficult for the current model to generate annotations but also similar to test images.

4. EXPERIMENTS

The effectiveness of the proposed approach is tested in the following aspects:

1) *How effective is the coherent language model for automatic image annotation* (CLM)? In the experiment, we will compare the proposed model to the relevance language model (described in Section 2) for automatic image annotation in terms of their accuracy in generated annotations. We will not compare CLM to other models because it has been shown empirically that relevance language model performs significantly better than the machine translation model in [9].
2) *How effective is the coherent language model with flexible length for automatic image annotation* (CLMFL)? In the experiment, we will evaluate the effectiveness of the proposed method by comparing the CLMFL with CLM using fixed length annotation.
3) *How effective is the proposed active learning method for automatic image annotation*? In the experiment, we compare the proposed strategy for selective sampling to a random sampling approach and investiage whether or not the proposed active learning method is more efficient in improving the quality of automatic image annotation.

For the rest of this section, we first describe the design of the experiment and then devote a separate subsection to each of the three issues.

4.1 Experiment Design

Since the focus of this paper is on a statistical model for image annotation and not the image features for effective annotation, we use the dataset provides in [8]. The dataset consists of 5,000 images from 50 Corel Stock Photo CDs. Normalized cut is used to segment images and the largest 10 regions are kept for each image. The K-means algorithm is used to cluster all image regions into 500 different blobs. Each image is annotated with 1 to 5 words and totally 371 distinct words have been used in annotations. Following [8], 4500 images out of 5000 are used as training examples and the rest 500 images are used for testing.

Similar to the previous studies on automatic image annotation, the quality of automatic image annotation is measured by the performance of retrieving auto-annotated images regarding to single-word queries. For each single-word query, **precision** and **recall** are computed using the retrieved lists that are based on the true annotations and the auto-annotations. Let I_j be a test image, t_j be its true annotation, and g_j be its auto-annotation. For a given query word *w*, precision and recall are defined respectively as:

$$\text{precision}(w) = \frac{|\{I_j | w \in t_j \wedge w \in g_j\}|}{|\{I_j | w \in g_j\}|}$$

$$\text{recall}(w) = \frac{|\{I_j | w \in t_j \wedge w \in g_j\}|}{|\{I_j | w \in t_j\}|}$$

The precision(*w*) measures the accuracy in annotating images with word *w* and the recall(w) measures the completeness in annotating images with word *w*. The average precision and recall over different single-word queries are used to measure the overall quality of automatically generated annotations for images. The third metric is the number of single-word queries for which at least one relevant image can be retrieved using the auto-annotations, or **#Ret_Query**. It is defined as:

$$\text{\#Ret\_Query} = |\{w | \text{precision}(w) > 0 \wedge \text{recall}(w) > 0\}|$$

Table 1: Evaluation results for single-word queries. 'RLM' refers to the relevance language model and 'CLM' refers to the coherent language model for automatic image annotation. 'CLMFL' refers to the method that applies the coherent language model to automatically determine the annotation length. '#Ret_Query' refers to the number of single-word queries that are retrieved with relevant images

	Avg. Prec.		Avg. Recall		# Ret_Query	
Fixed Len	RLM	CLM	RLM.	CLM	RLM	CLM
3	15.0%	16.3%	12.4%	12.8%	61	64
4	18.2%	19.0%	17.3%	18.2%	73	77
5	17.1%	18.4%	19.5%	21.4%	76	79
CLMFL	18.8%		16.2%		75	

Note that this metric compensates the metrics of average precision and average recall by providing information about how wide is the range of words that contribute to the average precision and recall. This metric is important because a biased model can achieve high precision and recall value by only performing extremely well on a small number of queries with common words.

Finally, there are totally 263 distinct words in the annotations of test images. However, according to the zipf's law [13], many words only appear in a few training images. In fact, there are 123 words that are used for less than 20 training images, which is less than 0.5% of training examples. Since it is usually difficult to handle the rare words, in this paper, we will only focus on 140 words that appear at least 20 times in the training dataset. Note that this is different from [9], in which metrics are computed based on the union of words retrieved by different methods in comparison. Our method has the advantage in that the set of words for evaluation is static and thus the metric values do not change from one set of comparison to another.

4.2 Experiment I: Coherent Language Model vs. Relevance Language Model

We train both the relevance language model and the proposed coherent language model over the same set of 4500 images, and test them against the same set of 500 images. A fixed annotation length is used for both models and it is varied from 3, 4, to 5. Table 1 summarizes the results for both models using the three metrics.

First, according to Table 1, for both models, the averaged recall improves significantly when the fixed annotation length is increased from 3 to 5. This is because, when more words are generated for each image, the chance for any one of them to be a true annotation word will increase. As a result, a longer annotation length leads to a higher average recall. In contrast, the average precision reaches its maximal when the annotation length is set as 4. This is because even though a longer annotation length results in more matched words, the number of unmatched words is also increased. Since precision is a ratio of the number of matched words to the total number of generated words, it appears that annotations with length four make the best tradeoff between these two factors.

Table 2: Evaluation results for two-word queries. CLM model is applied to generate annotations of fixed length varying from 3 to 5. CLMFL model is used to generate annotations with variable length.

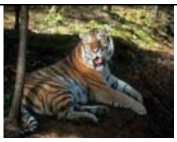
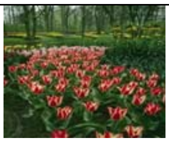
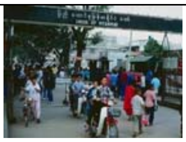
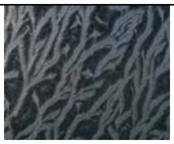
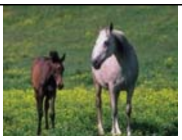
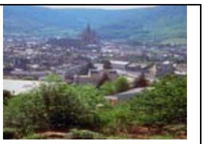
Fixed Length	Avg. Prec.	Avg. Recall
3	16.8%	13.2%
4	21.6%	20.3%
5	19.4%	23.8%
CLMFL	27.3%	18.9%

Second, for three different annotation lengths, the proposed coherent language model for automatic image annotation performs consistently better than or as well as the original relevance language model in all three metrics. Particularly, the advantage of using coherent language model for automatic image annotation is more noticeable when the annotation length is 5. In that case, the CLM model achieves precision as 18.4% and recall as 21.4% compared to precision 17.1% and recall 19.5% for the original RLM model. This is because word-to-word correlation has little impact on the very top-ranked words that have been determined by the image features with high confidence. It is much more influential to the words that are not ranked at the very top. For those words, the word-to-word correlation is used to promote the words that are more consistent with the very top-ranked words. Based on the above observation, we conclude that the proposed coherent language model is effective for automatic image annotation.

4.3 Experiment II: Generating Annotations with Automatically Determined Length

We test the effectiveness of the coherent language model with flexible length (CLMFL) for automatic image annotation. The

Table 3: Examples of annotations generated by both the relevance language model (i.e., RLM) and the coherent language model with flexible length (i.e., CLMFL). The manual annotations are included in the last row.

						
RLM	tree people forest cat tiger	tree grass house stone water	tree people street market food	ice glass frost frozen water	field horses mare foal water	tree buildings town church sky
CLMFL	tree forest cat tiger bengal	tree grass	people street market food	ice glass frost frozen	field horses mare foals	tree buildings
Manual	forest cat tiger Bengal	tree flowers garden tulip	people street sign	ice frost frozen crystal	field horses mare foals	tree buildings town church

evaluation results for CLMFL model using single-word query is listed Table 1. The average length for the generated annotations is about 3 words for each annotation.

According to Table 1, the CLMFL model performs significantly better than the CLM models when the fixed length is 3. But it is outperformed by CLM when the number of annotated words is increases to 4 and 5, particularly in terms of recall. This is because the average annotation length for CLMFL is only 3 and a longer annotation length usually leads to a better recall value. To demonstrate the benefits of CLMFL over CLM model, we also compare their performance over two-word queries. We select the 100 most frequent combinations of two words from the annotations of test images and use them as two-word queries. Similar to the single-word queries, we compute the precision and recall for each query and use the average precision and recall as the evaluation metrics. Table 2 summarizes the results for two-word queries for both the CLMFL model and the CLM model that use fixed annotation lengths. The most noticeable difference between the two models is on their average precision. For the 100 two-word queries, the precision for the CLMFL model is 27.3%, which is significantly better than the CLM model using fixed length. As the tradeoff, the CLMFL model achieves a lower recall than the CLM model when the fixed annotation length is four or five. In summary, the advantage of the CLMFL model is that it can automatically determine the annotation length and therefore the generated annotations are able to reflect the content of images more accurately than the CLM model that uses a fixed annotation length. Therefore, CLMFL model is particularly desirable for retrieval applications that prefer high precision than recall.

To further illustrate the effect of the CLMFL model for automatic image annotation, examples of auto-annotations by the CLMFL model are listed in Table 3, together with the manual annotations and the annotations generated by the relevance language model. First, according to Table 3, we clearly see that for the CLMFL model, the number of annotated words varies from one image to another. For example, five annotation words are generated for the first image and only two annotation words are created for the second image. Second, examples in Table 3 clearly indicate that the proposed model does benefit significantly from the word-to-word correlation. Take the fifth image as an example. Word 'water' appears in the auto-annotation generated by the relevance language model. This is because word 'water' is the most popular one in the training data and therefore has substantially more chances to be used as an annotation word than any other words. In contrast, word 'water' does not appear in the annotation that is generated by the CLMFL. This is because, even though word 'water' is used most frequently in training data, it is not strongly correlated with other words 'field', 'horse', 'mare', and 'foals'. Utilizing the word-to-word correlation information, the CLMFL model is able to substantially decrease the word probability $p(w_j = 1 | I)$ for 'water' and as a result remove it from the annotation.

Table 4: Results of number of single-word queries returned with nonzero precision and recall for the proposed active learning method and the baseline method using random selection.

Iteration	1	2	3	4	5
Active Learning	61	65	69	70	70
Random Selection	61	60	60	60	60

4.4 Experiment III: Active Learning for Automatic Image Annotation

In this section, we test the active learning method for automatic image annotation described in Section 3.3. First, 1000 annotated images are randomly selected from the training set and used as the initial training examples. Then, the system will iteratively acquire annotations for selected images. For each iteration, at most 20 images from the training pool (i.e., 4500 images) can be selected for manual annotation. Totally, four iterations of active learning are conducted and at most 80 additional annotated images are acquired. To evaluate the effectiveness of the proposed active learning method for automatic image annotation, at each iteration, we use the 1000 initially annotated images together with the acquired annotated images to generate annotations for the 500 testing images. The aforementioned three evaluation metrics are used to evaluate the quality of the current annotation model. A baseline method that randomly selects 20 images for each iteration from the training pool is also evaluated in the same way. Results of precision and recall for both the active learning method and the baseline method are displayed in Figure 1 and 2. The number of single-word queries that retrieve relevant images as a result of active learning and random selection are listed in Table 4.

First, according to Figure 1 and 2, the precision and recall for

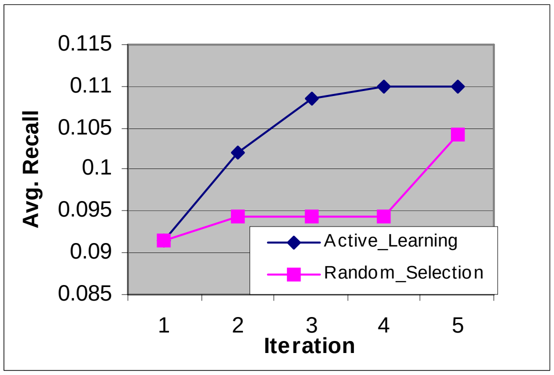


Figure 1: Average precision for the proposed active learning method and the baseline model using random selection.

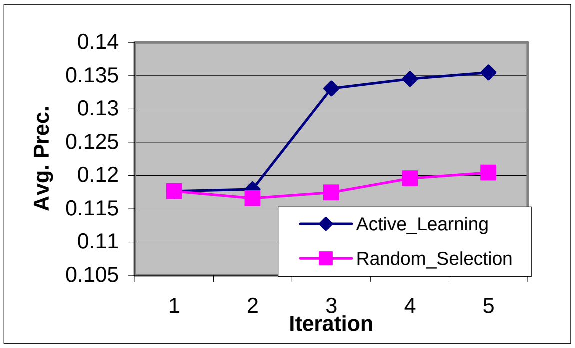


Figure 2: Average recall for the proposed active learning method and the baseline model using random selection.

both methods are improved through iterations. This is because for most supervised learning problems, more training examples usually lead to better performance. Second, compared to the baseline model, the active learning method is substantially more efficient in improving both the precision and recall of the annotation model. According to Table 4, the difference between these two methods is even more dramatic when we measure the number of single-word queries that retrieve at least one relevant image. For the baseline method, after four iterations of soliciting annotated images, the number of single-word queries that can be returned with relevant images is almost unchanged. Whereas for the active learning method, the number of queries with returned relevant images has been increased from 61 to 70. This is because images selected by the active learning method are the ones that the current annotation model cannot produce any annotations. As a result, the active learning method provides more chance for the annotation model to learn new objects with new words.

5. CONCLUSION

In this paper, we developed a coherent language model (i.e., CLM) for automatic image annotation. Compared to other approaches for automatic image annotation, the proposed model takes an advantage of word-to-word correlation. The word-to-word correlation is useful in predicating annotation words when the image features do not provide sufficient clues to distinguish between different words. Empirical studies have shown that the CLM model is noticeably more accurate than the relevance language model for automatic image annotation. More important, the coherent language model provides effective solutions to two important issues with automatic image annotation:

1) The coherent language model can be applied to automatically determine the annotation length. A variant of the coherent language model, called coherent language model with flexible length (i.e. CLMFL), is developed in this paper. Empirical studies show that the CLMFL model provides more accurate annotations than the original CLM model.

2) An active learning method for automatic image annotation based on the CLM model is developed to effectively reduce the required number of annotated images. Empirical studies have shown that it is substantially more effective than a simple random sampling approach.

Current model uses 0.5 as threshold for determining the length of annotations. In our future work, we plan to improve it by learning the threshold values from training examples. In particular, instead of having a fixed threshold, we can have thresholds that depend the properties of annotation words. We also plan to improve the proposed active learning methods for automatic annotation using different measurements of uncertainty, for example, the prediction-error based approaches for active learning.

6. REFERENCE

1. Abe, N. and H. Mamitsuka, *Query Learning Strategies Using Boosting and Bagging.* Proceedings of 15th International Conference on Machine Learning, 1998.
2. Barnard, K., P. Duygulu, and D. Forsyth. *Clustering Art*. in Proceedings of the 2001 IEEE Computer Society Conference on Pattern Recognition. 2001.
3. Blei, D. and M. Jordan. *Modeling Annotated Data*. in Proceedings of 26th International Conference on Research and Development in Information Retrieval (SIGIR). 2003.
4. Carson, C., et al. *Blobworld: A System for Region-Based Image Indexing and Retrieval*. in Proceedings of theThird International Conference on Visual Information Systems. 1999.
5. Chang, E., et al., *Cbsa: Content-Based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines.* CirSysVideo, 2003. **13**(1): p. 26-38.
6. Cusano, C., G. Ciocca, and R. Schettini. *Image Annotation Using Svm*. in Proceedings of Internet imaging IV, Vol. SPIE 5304. 2004.
7. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum Likelihood from Incomplete Data Via the Em Algorithm.* Joural of Royal Statistical Society, 1977. **39**(1): p. 1-38.
8. Duygulu, P., et al. *Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary*. in Proceedings of 7th European Conference on Computer Vision. 2002.
9. Jeon, J., V. Lavrenko, and R. 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 information retrieval. 2003.
10. Lavrenko, V., R. Manmatha, and J. Jeon. *A Model for Learning the Semantics of Pictures*. in Proceedings of Advance in Neutral Information Processing. 2003.
11. Li, J. and J.Z. Wang, *Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach,.* IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003. **25**(19): p. 1075-1088.
12. Maron, O. *Learning from Ambiguity*. MIT, 1998
13. Monay, F. and D. Gatica-Perez. *On Image Auto-Annotation with Latent Space Models*. in Proceedings of ACM International Conference on Multimedia. 2003.
14. Mori, Y., H. TAKAHASHI, and R. Oka. *Image-to-Word Transformation Based on Dividing and Vector Quantizing Images with Words*. in MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management. 1999.
15. Naphade, M.R., et al. *Learning to Annotate Video Databases*. in Proceedings of SPIE. 2001.
16. Naphade, M.R., I.V. Kozintsev, and T.S. Huang, *A Factor Graph Framework for Semantic Video Inexing.* IEEE Trans. on Circuits and Systems for Video Technology, 2002. **12**(1).
17. Roy, N. and A. McCallum. *Toward Optimal Active Learning through Sampling Estimation of Error Reduction*. in Proceedings of the 18th International Conference on Machine Learning. 2001.
18. Seung, H.S., M. Opper, and H. Sompolinsky, *Query by Committee.* Computatinal Learning Theory, 1992(287-294).
19. Tong, S. and D. Koller. *Active Learning for Parameter Estimation in Bayesian Networks*. in Advances in Neural Information Processing Systems. 2000.
20. Westerveld, T. and A.P.d. Vries. *Experimental Result Analysis for a Generative Probabilistic Image Retrieval Model*. in Proceedings of the 26th ACM SIGIR. 2003.