Topic Models for Image Annotation and Text Illustration

Yansong Feng and **Mirella Lapata**
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
Y.Feng-4@sms.ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

Image annotation, the task of automatically generating description words for a picture, is a key component in various image search and retrieval applications. Creating image databases for model development is, however, costly and time consuming, since the keywords must be hand-coded and the process repeated for new collections. In this work we exploit the vast resource of images and documents available on the web for developing image annotation models without any human involvement. We describe a probabilistic model based on the assumption that images and their co-occurring textual data are generated by mixtures of latent topics. We show that this model outperforms previously proposed approaches when applied to image annotation and the related task of text illustration despite the noisy nature of our dataset.

1 Introduction

Recent years have witnessed the rapid growth of image collections available for searching and browsing over the Internet. Although image search engines are still in their infancy, initial research suggests that the deployed algorithms are not very accurate (Hawking et al., 1999). Given a query, search engines retrieve relevant pictures by analyzing the image caption (if it exists), textual descriptions found adjacent to the image, and other text-related factors such as the file name of the image. However, since they do not analyze the actual content of the images, search engines cannot be used to retrieve pictures from unannotated collections. The ability to perform the annotation task *automatically* would be of significant practical import for many image-based applications. Besides search and retrieval, other examples include browsing support (e.g., by clustering images into groups that are visually and semantically coherent) and story picturing (i.e., automatically suggesting images to illustrate text).

Automatic image annotation is a popular task in computer vision; a large number of approaches have been proposed in the literature under many distinct learning paradigms. These range from supervised classification (Smeulders et al., 2000; Vailaya et al., 2001) to instantiations of the noisy-channel model (Duygulu et al., 2002), to clustering (Barnard et al., 2002), and methods inspired by information retrieval (Feng et al., 2004; Lavrenko et al., 2003). Despite differences in application and formulation, all these methods essentially attempt to learn the correlation between image features and words from examples of annotated images. The Corel database has been extensively used as a testbed for the development and evaluation of image annotation models. It is a collection of stock photographs, divided into themes (e.g., *tigers*, *sunsets*) each of which are associated with keywords (e.g., *sun*, *sea*) that are considered appropriate descriptors for all images belonging to the same theme.

Unfortunately, the Corel database is not representative of the size or content of real-world image collections. It has a small number of themes with many closely related images which in turn share keyword descriptions. It is therefore relatively easy to learn the associations between images and keywords and do well on annotation and retrieval tasks (Tang and Lewis, 2007; Westerveld and de Vries, 2003). An appealing alternative is the use of resources where images and their annotations co-occur naturally. Examples include images found in news documents, consumer photo collections, Wikipedia articles, illustrated stories and so on. The key idea here is to treat the words in the surrounding text as annotations for the image. These annotations are undoubt-

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 831–839,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

edly noisy, but plenty and cost-free. Moreover, the collateral text is often longer and more informative in comparison to the few keywords reserved for each image in Corel.

In this paper we propose a probabilistic image annotation model that learns to automatically label images under such noisy conditions. We use the database created in Feng and Lapata (2008) which consists of news articles, images, and their captions. Our model exploits the redundancy inherent in this multimodal dataset by assuming that images and their surrounding text are generated by a shared set of latent variables or topics. Specifically, we describe documents and images by a common multimodal vocabulary consisting of textual words and *visual terms* (visiterms). Due to polysemy and synonymy many words in this vocabulary will refer to the same underlying concept. Using Latent Dirichlet Allocation (LDA, Blei and Jordan 2003), a probabilistic model of text generation, we represent visual and textual meaning *jointly* as a probability distribution over a set of topics. Our annotation model takes these topic distributions into account while finding the most likely keywords for an image and its associated document. We also show how the model can be straightforwardly modified to perform automatic text illustration.[1] The task is routinely performed by news writers who often have to search large image collections in order to find suitable pictures for their text. Here, the model takes a document as input and suggests images that match its content. Experimental results on both tasks bring improvements over competitive models.

2 Related Work

A variety of learning methods have been applied to the image annotation task. These generally fall under two broad categories. Supervised methods define annotation as a classification task, e.g., by assuming a one-to-one correspondence between vocabulary words and classes or by grouping several words into a single class (see Chai and Hung 2008 for an overview). Unsupervised approaches attempt to discover the underlying connections between visual features and words, typically by introducing latent variables. Standard latent semantic analysis (LSA) and its probabilistic variant (PLSA) have been applied to this task (Hofmann, 2001; Monay and Gatica-Perez, 2007; Pan et al., 2004). More sophisticated models estimate the joint distribution of words and regional image features, whilst treating annotation as a problem of statistical inference in a graphical model (Barnard et al., 2002; Blei and Jordan, 2003; Wang et al., 2009).

Irrespectively of the underlying model or task at hand, much work has focused how to best represent the visual and textual modalities in order to exploit their synergy. Several approaches attempt to render images more word-like, by reducing the dimensionality of the image feature space (Bosch et al., 2008; Fei-Fei and Perona, 2005) or by learning a single representation for both visual and textual features (Monay and Gatica-Perez, 2007; Zhao and Grosky, 2003).

Our own work approaches the image annotation (and related story picturing) task from a slightly different angle. We train and test our model on images that contain implicit (and thus noisy) annotations that have not been specifically created for our task. On account of this, our model has access to knowledge sources other than the image and its keywords (i.e., the news article containing the image we wish to annotate). In Feng and Lapata (2008) we addressed this problem with a modified version of the continuous relevance annotation model (Lavrenko et al., 2003). Unlike other unsupervised approaches where a set of latent variables is introduced, each defining a joint distribution on the space of keywords and image features, the relevance model captures the joint probability of images and annotated words *directly*, without requiring an intermediate clustering stage (i.e., each annotated image in the training set is treated as a latent variable). We modified this model so as to exploit the information present in the document in two ways. First, in estimating the conditional probability of a keyword given an image, we also considered its likelihood in the collateral document. Secondly, we used an LDA model (trained on the document collection) to prune from the model's output words that are not representative of the document's topics.

The proposed approach differs from Feng and Lapata (2008) in three important respects: (a) document-based information is an integral part of our model as we predict caption words given the image *and* its accompanying document (b) LDA is no longer a post-processing step; our model relies on LDA to infer meaningful topics that capture the co-occurrence of visual features and words; (c) beyond image annotation, we show how the same framework can be applied to story picturing (Joshi et al., 2006), a task which has received less attention in the literature.

In terms of model structure, Blei and Jordan

[1] We use the terms "text illustration" and "story picturing" interchangeably throughout the paper.

(2003) and Monay and Gatica-Perez (2007) are closest to our work. The first model, known as correspondence LDA (CorrLDA), has been successfully employed for modeling annotated images in the Corel domain. CorrLDA also uses the notion of topic to model the generation of images and their captions. In this model, the visual modality drives the definition of the latent space to which the textual modality is linked. The second model is based on PLSA and learns a representation similar to ours consisting of textual and visual features. It is also trained using captioned images from the Corel database. We work with noisier and larger datasets. Our model exploits the captions accompanying the images as well as their surrounding documents. As a result, we obtain a similar number of textual and visual words; these are often imbalanced in the Corel database, where visual words are nearly 50 times more than textual words. The different nature of our data dictates the use of a model where the visual and textual modality are given equal importance in defining the latent space.

3 Problem Formulation

In this section we give a brief description of the image database we employ and also define the image annotation and story picturing tasks we are attempting here. As mentioned earlier, we use the dataset created in Feng and Lapata (2008).[2] It contains 3,361 articles that have been downloaded from the BBC News website[3]. Each article contains a news image which in turn is associated with a caption. The images are usually 203 pixels wide and 152 pixels high. The average caption length is 5.35 tokens, and the average document length 133.85 tokens. The captions vocabulary is 2,167 words and the document vocabulary is 6,253. The vocabulary shared between captions and documents is 2,056 words. In contrast to the Corel database, this dataset contains more complex images (with many objects) and has a larger vocabulary (Corel's vocabulary is approximately 300 words). An example of an abridged database entry is shown in Figure 1.

Due to the non-standard nature of the database we assume that the caption and news article describe the content of the image either directly or indirectly. It also follows that we may not be able to name all objects depicted in the image. Now, given these constraints our goal is twofold. Firstly, we will perform image annotation. Our model is trained on document-image-caption tuples like the one shown in Table 1. During testing, we must infer the caption for an image. Secondly, we use the same dataset to perform automatic text illustration. During training, the model has access to the same collection of image-caption-document tuples. During testing, we are given a document and must find the images that best illustrate it.

A woman from East Sussex who bought an emu egg sold as a novelty food item on a farm on the Isle of Wight has managed to hatch it into a chick. Gillian Stone, from Bexhill, who breeds chickens, brought home three large green emu eggs from a holiday and put them in an incubator in her kitchen. Two turned out to be infertile, but after 52 days little Osborne hatched

Osborne the emu will grow to over 6ft tall

Table 1: Each entry in the BBC News database contains a document, an image, and its caption (shown in boldface).

4 Image and Document Representation

Words and images represent distinct modalities, images live in a continuous feature space, whereas words are discrete. Yet, both modalities on some level capture the same underlying concepts as they are used to describe the same objects. A common first step to all previous methods is the segmentation of the image into regions, using either a fixed-grid layout or an image segmentation algorithm. Regions are usually described by a standard set of features including color, texture, and shape which are treated as continuous vectors (e.g., Barnard et al. 2002; Blei and Jordan 2003) or in quantized form (e.g., Duygulu et al. 2002). Through this process, the low-level image features are made to resemble word-like units.

Here, we go one step further and represent each image by a bag of visual words, thereby converting visual features from a continuous onto a discrete space. In order to do this we use the Scale Invariant Feature Transform (SIFT) algorithm (Lowe, 1999). The general idea behind the algorithm is to first sample an image with the difference-of-Gaussians point detector at different scales and locations. Importantly, this detector is, to some extent, invariant to translation, scale, rotation and illumination changes. Each detected region is represented with a SIFT descriptor which is a histogram of edge directions at

[2] Available from `http://homepages.inf.ed.ac.uk/s0677528/data.html`.

[3] http://news.bbc.co.uk/

different locations. SIFT descriptors are quantized using the K-means clustering algorithm to obtain a discrete set of visual terms (visiterms) which form our visual vocabulary Voc_v. Each entry in this vocabulary stands for a group of image regions which are similar in content or appearance and assumed to originate from similar objects. More formally, each image I is expressed in a bag-of-words format vector, $[w_{v1}, w_{v2}, ..., w_{v_L}]$, where $w_{v_i} = n$ only if I has n regions labeled with v_i.

Since visual and textual modalities have now the same status—they are both represented as bags-of-words—we can also represent any image-caption-document tuple *jointly* as a mixed document d_{Mix}. The underlying assumption is that the two modalities express the same meaning which, as we explain below, can be operationalized as a probability distribution over a set of topics.

5 Modeling

Latent Dirichlet Allocation For ease of exposition, we first describe the basics of Latent Dirichlet Allocation (LDA; Blei et al. 2003), a probabilistic model of text generation and then move on to discuss our models which make use of probabilities estimated by LDA.

LDA can be represented as a three level hierarchical Bayesian model. Given a corpus consisting of M documents, Blei et al. (2003) define the generative process for a document d as follows:

1. Choose $\theta|\alpha \sim Dir(\alpha)$
2. For $n \in 1, 2, ..., N$:
   (a) Choose topic $z_n|\theta \sim Mult(\theta)$
   (b) Choose a word $w_n|z_n, \beta_{1:K} \sim Mult(\beta_{z_n})$

The mixing proportion over topics θ is drawn from a Dirichlet prior with parameters α whose role is to create a smoothed topic distribution. Once α and β are sampled, then each document is generated according to the topic proportions $z_{1:K}$ and word probabilities over topics β. The probability of a document d in a corpus is defined as:

$$P(d|\alpha,\beta) = \int_{\theta} P(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_k} P(z_k|\theta) P(w_n|z_k,\beta) \right) \mathrm{d}\theta$$

Computing the posterior distribution $P(\theta, z|d, \alpha, \beta)$ of the hidden variables given a document is intractable in general. However, a variety of approximate inference algorithms have been proposed in the literature including variational inference which our model adopts (Blei et al., 2003). In this case, training an LDA model on a document collection will give two sets of parameters, the word probabilities given topics, $p(w|z_{1:K})$, and the topic proportions given documents, $P(z_{1:K}|d)$. The latter are document-specific, whereas the former represent the set of topics learned from the document collection.

Given a trained model, it is possible to do inference on an unseen document d_{new}:

$$p(w|d_{new}) \approx \sum_{k=1}^{K} P(w|z_k) \frac{\gamma_k}{\sum_{j=1}^{K} \gamma_j} \tag{1}$$

where $P(w|z_{1:K})$ are word probabilities over topics $z_{1:K}$ estimated during model training, and $\gamma_{1:K}$ are variational Dirichlet parameters obtained during inference on the new document (and can be considered as the posteriors of topic proportions over documents).

Image Annotation In the standard image annotation setting, a hypothetical model is given an image I and a set of keywords W, and must find the subset W_I ($W_I \subseteq W$) which appropriately describes image I:

$$W_I^* = \arg\max_W P(W|I) \tag{2}$$

The keywords are usually assumed to be conditionally independent on each other, so Equation (2) simplifies to:

$$W_I^* = \arg\max_W \prod_{w \in W} P(w|I) \tag{3}$$

Each entry in our database is an image-caption-document tuple (I, C, D). In this setting, we must find the subset of keywords W_I which appropriately describe image I with the help of the accompanying document D:

$$W_I^* = \arg\max_{W_t} P(W_t|I, D) \tag{4}$$

Here, W_t denotes a set of textual words (we use the subscript t to discriminate from the visual words which are not part of the model's output). We also assume that the keywords are conditionally independent of each other:

$$W_I^* = \arg\max_{W_t} P(W_t|I, D) = \arg\max_{W_t} \prod_{w_t \in W_t} P(w_t|I, D) \tag{5}$$

Since I and D are represented jointly as the concatenation of textual and visual terms, we may intuitively simplify the problem and use the mixed document representation d_{Mix} directly in estimating the conditional probabilities $P(w_t|I, D)$:

$$P(w_t|I, D) \approx P(w_t|d_{Mix}) \tag{6}$$

Substituting Equation (6) into (5) yields:

$$W_I^* \approx \arg\max_{W_t} \prod_{w_t \in W_t} P(w_t|d_{Mix}) \qquad (7)$$

As mentioned earlier, we assume that the image and its associated text are generated by a mixture of latent topics which we infer using LDA. Specifically, we select the number of topics K and apply the LDA algorithm to a corpus consisting of documents $\{d_{Mix}\}$ in order to obtain the multimodal word distributions over topics $P(w|z_{1:K})$, and the estimated posterior of the topic proportions over documents $P(z_{1:K}|d_{Mix})$. We infer the topic proportions $P(z_{1:K}|d_{Mix_{new}})$ on a new document-image pair $d_{Mix_{new}}$ approximately using Equations (1) and (7):[4]

$$\begin{aligned} W_I^* &\approx \arg\max_{W_t} \prod_{w_t \in W_t} P(w_t|d_{Mix}) \qquad (8) \\ &= \arg\max_{W_t} \prod_{w_t \in W_t} \sum_{k=1}^{K} P(w_t|z_k) P(z_k|d_{Mix}) \\ &\approx \arg\max_{W_t} \prod_{w_t \in W_t} \sum_{k=1}^{K} P(w_t|z_k) \frac{\gamma_k}{\sum_{j=1}^{K} \gamma_j} \end{aligned}$$

where $P(w_t|z_k)$ are obtained during training, and $\gamma_{1:K}$ are inferred on the image-document test pair.

However, note that for an unseen image d_I and accompanying document d_D, the estimated topic proportions are solely based on variational inference which is an approximate algorithm. In order to render the model more robust, we smooth the topic proportions $P(z_{1:K}|d_{Mix})$ with probabilities based on a single modality:

$$\begin{aligned} &P^*(z_{1:K}|d_{Mix}) \approx \qquad (9) \\ &q_1 P(z_{1:K}|d_{Mix}) + q_2 P(z_{1:K}|d_D) + q_3 P(z_{1:K}|d_I) \end{aligned}$$

where $P(z_{1:K}|d_D)$ and $P(z_{1:K}|d_I)$ are inferred on d_D and d_I, respectively, and q_1, q_2, q_3 are smoothing parameters (which we tune experimentally on held-out data); q_3 is a shorthand for $(1 - q_1 - q_2)$.

In sum, calculating $P(W_t|I,D)$ boils down to estimating the probabilities $P(w_t|d_{Mix})$ according to Equations (8) and (9) which we obtain using the LDA model. We train LDA on the document collection $\{d_{Mix}\}$ and use inference to obtain the topic distributions of unseen image-document pairs. In the end, we obtain a ranked list of textual words w_t, the n-best of which are the annotations for image I.

Text Illustration Previous text illustration models are based on Corel-like databases with manual image descriptions (Barnard and Forsyth, 2001; Blei and Jordan, 2003) or instance-based learning using complex learning schemes (Joshi et al., 2006). Here, we present a relatively simple model, again under the topic mixture framework.

Given a test document D and a candidate image database $I_{1...N}$ with captions C, we must find the image or images which best describe the document. We can simply compute the probability of each visual term in the vocabulary given D by marginalizing over the document topics $z_{1:K}$:

$$P(w_v|D) = \sum_{z_{1:K}} P(w_v|z_k) P(z_k|d_D) \qquad (10)$$

where w_v is a visual term and $P(w_v|z_k)$ the probability of w_v given topic z_k (as estimated on the training set).

Equation (10) delivers a ranked list of visual terms according to a given document. We could multiply these probabilities together mirroring Equation (7), however this is not reliable. In contrast to textual words, for which we may infer whether they are linguistically meaningful (e.g., by resorting to their part of speech), there is no easy way of knowing which visual words are important. Relying solely on frequency is not reliable either, as frequent visiterms may simply represent features common in all images (e.g., most images have some white color). To avoid a bias towards frequent but potentially irrelevant visual words, we output a fixed number of visual terms and select the image with the highest overlap as the correct illustration.

6 Experimental Setup

In this section we discuss our experimental design for assessing the performance of the models presented above. We give details on our training procedure and parameter estimation, describe our features, and present the baseline methods used for comparison with our models.

Data We evaluated the image annotation and text illustration tasks on the dataset described in Section 3. Documents and captions were part-of-speech tagged and lemmatized with Tree Tagger (Schmid, 1994). We excluded from the vocabulary low frequency words (appearing fewer than five times) and words other than nouns, verbs, and adjectives. For the image annotation task we follow the data split used in Feng and Lapata (2008). The training set contains 2,881 image-caption-document tuples; 240 tuples are reserved for development and 240 for

[4] During training, the model has access to all three elements (I,C,D), so the mixed document d_{Mix} is the concatenation of the visual terms and words in the caption and document. During testing, the model is given an image and its accompanying document, so d_{Mix} contains words based on I and D, but not C.

testing. Our text illustration experiments, used 2,881 image-caption-document tuples for training. For the purposes of simulating a real story picturing engine environment, we created a large image pool of 450 image-caption pairs and tested on 300 of them.

Model Parameters For each image we extracted 150 (on average) SIFT features. These were quantized into a discrete set of visual terms using K-means. We varied K from 100 to 2,000. We trained the LDA topic model on the multimodal document collection $\{d_{Mix}\}$ and varied the number of topics from 15 to 1,000. The hyperparameter α was initialized to 0.1; the β probabilities were initialized randomly. The maximum number of iterations for variational inference was set to 1,000. We tuned the smoothing parameters q_1, q_2, and q_3 (see Equation (9)) on the development set. The best values were $q_1 = 0.84$, $q_2 = 0.12$, and $q_3 = 0.04$ (for both tasks). As the number of visual terms and topics are interrelated we exhaustively examined all possible combinations on the development set. We obtained best results on image annotation with 1,000 topics and 750 visual terms. On text illustration the best parameters were 1,000 topics and 2,000 visual terms.

Baselines For the image annotation experiments, we compared our model against the following baselines. Firstly, we trained a vanilla LDA model on the document collection without taking the images into account. This model estimates $P(w_t|D) = \sum_{k=1}^{K} P(w_t|z_k)P(z_k|D)$, the probability of textual word w_t given text document D. We assume that the most probable words are the captions for the accompanying image. Our second baseline is the extended relevance model (Feng and Lapata, 2008) that also takes the document into account but crucially assumes that the process of generating the images is independent from the process of generating its keywords.

We also compared our approach with two closely related latent variable models (developed for image-caption pairs), a PLSA-based model (Monay and Gatica-Perez, 2007) and CorrLDA (Blei and Jordan, 2003). The former model is an asymmetric version of PLSA; it estimates the topic structure solely from the textual modality and keeps it fixed for the visual modality. The technique is similar to *folding-in* (Hofmann, 2001), the standard PLSA procedure for inference in unseen documents and allows modeling an image as a mixture of latent topics that is defined only by one modality (in this case the caption words). CorrLDA first generates image regions from a Gaussian multinomial distribution parametrized with Dirichlet priors. Then, for each annotation word, it uniformly selects a region from the image and generates a word according to the topic used to generate that region. We optimized the parameters for both models on the development set. For CorrLDA, we followed the mean-field variational inference strategy proposed in Blei (2004). The optimal number of topics for PLSA, was 200 (with 2000 visual terms) and for CorrLDA 50.

For the text illustration experiments, the proposed model was compared with three baselines. The first one is essentially word overlap. We select the image whose caption has the largest number of words in common with the test document. The second one is a straightforward implementation of the vector space model (Salton and McGill, 1983) where documents and captions are represented by vectors whose components correspond to term-document co-occurrences. We followed common practice in weighting terms by their tf-idf values, and used the cosine similarity measure to find the image whose caption is most similar to the test document. Our third baseline uses a text-based LDA model to estimate document-caption similarity probabilistically, through topic sharing. The images most relevant to a document are found by maximizing the conditional probability of the candidate captions C given the document d_D: $P(C|d_D) = \prod_{w_c \in C} \sum_{k=1}^{K} P(w_c|z_k)P(z_k|d_D)$ (where w_c are the caption words, $P(w_c|z_k)$ the conditional distribution of each w_c given a topic z_k, and $P(z_k|d_D)$ the conditional distribution of z_k given d_D, the document we wish to illustrate.

Evaluation In the image annotation task we follow the evaluation methodology proposed in Duygulu et al. (2002). We are given an un-annotated image I and asked to automatically produce the n-best keywords. For all models discussed here, we report results with the top 10 annotation words using precision, recall and F1. In the text illustration task, we are given a test document d and a pool of candidate images $I_{1...N}$ with captions $C_{1...N}$. The model is expected to find an image from the candidate pool that matches the test document. We use equation (10) to output a ranked list of M_I visual terms. The image having the highest overlap with the top 30 visual terms is selected as the illustration for the test document. All illustration models were evaluated using top 1 accuracy, which is the percentage of successfully matched image-document pairs in the test set.

Model	Top 10		
	Precision	Recall	F1
CorrLDA	5.33	11.80	7.36
TxtLDA	7.30	16.90	10.20
PLSA	10.26	22.60	14.12
ExtRel	14.70	27.90	19.80
MixLDA	16.30	33.10	21.60

Table 2: Automatic image annotation results.

TxtLDA	
Afghanistan, Taleban, soldier, British, zone, kill, force, Microsoft, **troop**, NATO	police, Burgess, time, letter, **crash**, case, death, operation, investigation, jail
MixLDA	
Afghanistan, **troop**, Blair, British, NATO, **helicopter**, soldier, support, **operation**, commander	**Diana**, police, case, **crash**, **Princess**, report, **death**, inquest, **Paris**, Burgess
Caption	
Troops need more Chinook helicopters to carry out operations	Princess Diana died in a car crash in Paris in 1997

Figure 1: Annotations generated by the TxtLDA and MixLDA models. Words in bold face indicate exact matches. The original captions are in the last row.

7 Results

Our results on the image annotation task are summarized in Table 2. Here, we compare our own model (MixLDA) which is trained on both visual and textual information against an LDA model based solely on textual information (TxtLDA), an extended version of the Continuous Relevance model that also exploits collateral document information (ExtRel; Feng and Lapata 2008), a PLSA model that prioritizes the textual over visual modality (Monay and Gatica-Perez, 2007), and CorrLDA (Blei and Jordan, 2003) which does the opposite. We performed significance testing on F1 using stratified shuffling (Noreen, 1989), an instance of assumption-free approximative randomization testing.

Let us first discuss the performance of TxtLDA and MixLDA. These two models are closely related — they both rely on the probabilities $P(w_t|d)$ to generate the image keywords — save one important difference. MixLDA uses a concatenated representation of words and visual features assuming that the two modalities have equal importance in defining the latent space, whereas TxtLDA considers only the textual modality. Our results show that MixLDA outperforms TxtLDA in terms of precision (by 9%), recall (by 16.2%). MixLDA improves F1 by 11.4%, and the difference is significant ($p < 0.01$).

PLSA significantly ($p < 0.01$) improves upon TxtLDA. The key difference is the visual information which makes up (to a certain extent) for the lack of richer textual data. Interestingly, CorrLDA performs significantly ($p < 0.01$) worse than both PLSA and TxtLDA. Recall that in CorrLDA word topic assignments are drawn from the image regions which are in turn drawn from a Gaussian distribution. Although this modeling choice delivers better results on the simpler Corel dataset, it does not seem able to capture the characteristics of our images which are noisier and more complex. Moreover, CorrLDA assumes that annotation keywords *must correspond* to image regions. This assumption is too restrictive in our setting where a single keyword may refer to many objects or persons in an image (e.g.,the word *badminton* is used to collectively describe an image depicting players, shuttlecocks, and rackets). As an aside, it is interesting to note, that neither PLSA nor CorrLDA achieve better results, when modified to take the captions *and* associated documents into account. PLSA scores are in the same ballpark (see Table 2), whereas CorrLDA performs worse, F1 decreases by 2%.

The extended relevance model improves considerably upon TxtLDA, CorrLDA, and PLSA but is significantly worse ($p < 0.01$) than MixLDA. On the surface, MixLDA seems similar to ExtRel, both models take advantage of visual and textual information. ExtRel smooths the conditional probability of a word given an image with the conditional probability of the same word given the document and uses an LDA model (trained on the document collection) to remove non-topical keywords from the model's output. MixLDA is conceptually simpler, LDA is the actual model rather than a post-processing step, and exploits the synergy between visual and textual information more directly. Topics are created based on both modalities which are treated on an equal footing. Compared to ExtRel, MixLDA improves precision by 1.6%, recall by 5.2% and the overall F1 by 1.8%.

Figure 1 illustrates examples of annotations gen-

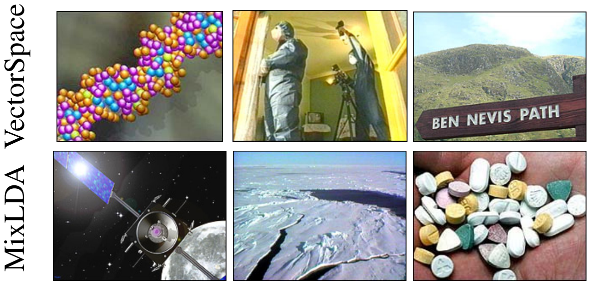
Model	Accuracy
TxtLDA	31.0
Overlap	31.3
VectorSpace	38.7
MixLDA	57.3

Table 3: Text Illustration results.

erated by TxtLDA and MixLDA for two images. For comparison, we also show the goldstandard image captions. Note that TxtLDA fails to generate any words relating to the objects shown in the image. It finds primarily words relating to the topics of the associated articles such as *troops* and *crash*. On the contrary, MixLDA is more successful at identifying the depicted objects, since it takes visual information into account.

Table 3 presents our results on the automatic text illustration task. Here, we compare our multimodal topic model (MixLDA) against three text-based baselines, namely word overlap (Overlap) a standard vector space model (VectorSpace), and TxtLDA. We examined whether differences in performance are statistically significant using a χ^2 test. As can be seen, MixLDA significantly ($p < 0.01$) outperforms these models by a wide margin (accuracy is 57.3% for MixLDA vs. 31.0% for TxtLDA, 38.7% for the vector space model, and 31.3% for word overlap). These results are encouraging given the simplicity of our model. They also indicate that substantial mileage can be gained by taking into account the visual modality.

Figure 2 shows the three best illustrations found by MixLDA and VectorSpace (incidentally, Overlap delivered the same ranking as VectorSpace). The images are presented in ranked order, i.e., the first image was given a higher score than the second one, etc. The document discusses Smart 1 Probe, a lunar satellite about to end its mission by crashing onto the moon's surface. MixLDA identifies an image depicting this satellite. The second best picture is also relevant, it resembles the moon's surface. The VectorSpace model does not find any related images, the first one is a DNA image, the second one depicts policemen at a crime scene and the third one Ben Nevis, the highest mountain in the British Isles.



Europe's lunar satellite, the Smart 1 probe, is about to end its mission by crashing onto the Moon's surface. It will be a spectacular end for the robot which has spent the past 16 months testing innovative and miniaturized space technologies. Smart 1 has also produced detailed maps of the Moon's chemical make-up.

Figure 2: Top-3 illustrations for document in bottom row.

8 Conclusions

In this paper we have presented a probabilistic approach for automatic image annotation and text illustration. Our model postulates that visual terms and words are generated by common (hidden) topics and is trained on a dataset consisting of images available on the Internet, their captions, and associated news articles. The annotations are implicit and the dataset is representative of the scale, diversity, and difficulty of real-world image collections. Overall, our results show that the model is robust to the noise inherent in such data. It improves upon competitive approaches that prioritize one modality over the other or exploit them indirectly. We also show that with minor modifications the model can be used to automatically illustrate a document with an appropriate image. Our method shows promise for multimodal search and image retrieval and other applications which have been traditionally text-based. An interesting future direction involves generating actual sentence descriptions rather than isolated keywords. Another relevant application is summarization. Our results suggest that taking visual information into account could potentially create more focused and accurate summaries.

The model presented here could be further improved in two ways. Firstly, we could allow an infinite number of topics and develop a nonparametric version that *learns* how many topics are optimal. Secondly, our model is based on word unigrams, and in this sense takes very little linguistic knowledge into account. Recent developments in topic modeling could potentially rectify this, e.g., by assuming that each word is generated by a distribution that combines document-specific topics and parse-tree-specific syntactic transitions (Boyd-Graber and Blei, 2009).

References

Barnard, K., P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. 2002. Matching words and pictures. *Journal of Machine Learning Research* 3:1107–1135.

Barnard, K. and D. Forsyth. 2001. Learning the semantics of words and pictures. In *Proceedings of the 8th International Conference on Computer Vision*. Vancouver, BC, pages 408–415.

Blei, D. 2004. *Probabilistic Models of Text and Images*. Ph.D. thesis, U.C. Berkeley, Division of Computer Science.

Blei, D. and M. Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference*. Toronto, ON, pages 127–134.

Blei, D., A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bosch, A., A. Zisserman, and X. Munoz. 2008. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(4):712–727.

Boyd-Graber, J. and D. Blei. 2009. Syntactic topic models. In *Proceedings of the 22nd Conference on Advances in Neural Information Processing Systems*. MIT, Press, Cambridge, MA, pages 185–192.

Chai, C. and C. Hung. 2008. Automatically annotating images with keywords: A review of image annotation systems. *Recent Patents on Computer Science* 1:55–68.

Duygulu, P., K. Barnard, J. de Freitas, and D. Forsyth. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision*. Copenhagen, Danemark, pages 97–112.

Fei-Fei, L. and P. Perona. 2005. A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society Washington, DC, volume 2, pages 524–531.

Feng, S., V. Lavrenko, and R. Manmatha. 2004. Multiple Bernoulli relevance models for image and video annotation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. Washington, DC, pages 1002–1009.

Feng, Y. and M. Lapata. 2008. Automatic image annotation using auxiliary text information. In *Proceedings of ACL-08: HLT*. Columbus, OH, pages 272–280.

Hawking, D., N. Craswell, P. Thistlewaite, and D. Harman. 1999. Results and challenges in web search evaluation. *Computer Networks* 31(11):1321–1330.

Hofmann, T. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 41(2):177–196.

Joshi, D., J.Z. Wang, and J. Li. 2006. The story picturing engine—a system for automatic text illustration. *ACM Transactions on Multimedia Computing, Communications, and Applications* 2(1):68–89.

Lavrenko, V., R. Manmatha, and J. Jeon. 2003. A model for learning the semantics of pictures. In *Proceedings of the 17th Conference on Advances in Neural Information Processing Systems*. MIT, Press, Cambridge, MA.

Lowe, D. 1999. Object recognition from local scale-invariant features. In *Proceedings of International Conference on Computer Vision*. IEEE Computer Society, pages 1150–1157.

Monay, F. and D. Gatica-Perez. 2007. Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(10):1802–1817.

Pan, J., H. Yang, P. Duygulu, and C. Faloutsos. 2004. Automatic image captioning. In *Proceedings of the 2004 International Conference on Multimedia and Expo*. Taipei, pages 1987–1990.

Salton, G. and M.J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, pages 44–49.

Smeulders, A. W., M. Worring, S. Santini, A. Gupta, and R. Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12):1349–1380.

Tang, J. and P. H. Lewis. 2007. A study of quality issues for image auto-annotation with the Corel data-set. *IEEE Transactions on Circuits and Systems for Video Technology* 17(3):384–389.

Vailaya, A., M. Figueiredo, A. Jain, and H. Zhang. 2001. Image classification for content-based indexing. *IEEE Transactions on Image Processing* 10:117–130.

Wang, C., D. Blei, and L. Fei-Fei. 2009. Simultaneous image classification and annotation. In *Proceedings of CVPR*. Miami, FL, pages 1903–1910.

Westerveld, T. and A. P. de Vries. 2003. Experimental evaluation of a generative probabilistic image retrieval model on ‘easy’ data. In *Proceedings of the SIGIR Multimedia Information Retrieval Workshop*. Toronto, ON.

Zhao, R. and W. I. Grosky. 2003. Video shot detection using color anglogram and latent semantic indexing: From contents to semantics. In B. Furht and O. Marques, editors, *Handbook of Video Databases: Design and Applications*, CRC Press, pages 371–392.