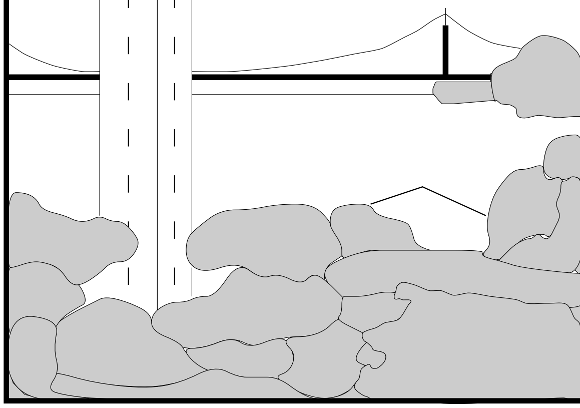
Modeling Annotated Data

David M. Blei
blei@cs.berkeley.edu
Computer Science Division
University of California
Berkeley, CA 94720, USA

Michael I. Jordan
jordan@cs.berkeley.edu
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA



Report No. UCB/CSD-02-1202

January, 2003

Computer Science Division (EECS)
University of California
Berkeley, California 94720

Modeling Annotated Data

David M. Blei
blei@cs.berkeley.edu
Computer Science Division
University of California
Berkeley, CA 94720, USA

Michael I. Jordan
jordan@cs.berkeley.edu
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

January, 2003

Abstract

We consider the problem of modeling annotated data—data with multiple types where the instance of one type (such as a caption) serves as a description of the other type (such as an image). We describe three hierarchical probabilistic mixture models that are aimed at such data, culminating in the Corr-LDA model, a latent variable model that is effective at modeling the joint distribution of both types and the conditional distribution of the annotation given the primary type. We take an empirical Bayes approach to finding parameter estimates and conduct experiments in held-out likelihood, automatic annotation, and text-based image retrieval using the Corel database of images and captions.

1 Introduction

Traditional methods of information retrieval and organization are tailored to effectively represent and analyze a document in (high-dimensional) word-space. Modern algorithms for these tasks, however, must address that a document is not merely a collection of words, but can be a collection of related text, images, audio, and cross-references. Performing tasks such as retrieval, classification, and clustering requires methods of representing multiple types of high-dimensional data and successfully associating similar data of different types.

In previous work, generative probabilistic models have played an important role in modeling this type of data, often referred to as *dyadic* or *relational.* Several models have been developed that jointly cluster the different types, finding latent variable representations that capture low-dimensional probabilistic relationships among interacting sets of variables (Barnard et al., 2002; Brochu and de Freitas, 2003; Cohn and Hofmann, 2001; Taskar et al., 2001).

In many problems, however, the overall goal is to find a *conditional* relationship between types, and in such cases improved performance may be found in methods with a more discriminative flavor. In particular, correlated data sets often involve *annotation*, in which the annotation type serves as a description of the primary type. Examples include images and their captions, papers and their bibliographies, and genes and their functions. The (discriminative) goal is to build a model that can provide a description of an unannotated instance of the primary type (*e.g.*, label a new image or find images that are relevant to a query of words). Yet much of the problem in such domains is to find good low-dimensional representations on which such conditional relationships can be built, a task that has more commonly been addressed within the generative tradition.

In this paper, we explore these issues in the context of modeling images and their captions. We build a set of increasingly sophisticated models for data in this domain, culminating in *correspondence latent Dirichlet allocation* (Corr-LDA), a model that finds conditional relationships between latent variable representations of sets of image regions and sets of words. We show that, in this class of models, only Corr-LDA succeeds in providing both an excellent fit of the joint data and an effective conditional model of the caption given an image. We demonstrate its use in automatic image annotation, automatic region annotation, and text-based image retrieval.

2 Image/Caption data

Our work has focused on images and their captions from the Corel database. Following previous work (Barnard et al., 2002), each image is segmented into regions by the N-cuts algorithm (Shi and Malik, 2000). For each region, we compute a set of real-valued features representing visual properties such as size, position, color, texture, and shape. Each image and its corresponding caption is represented as a pair $(\mathbf{r}, \mathbf{w})$. The first element $\mathbf{r} = \{r_1, \ldots, r_N\}$ is a collection of feature vectors associated with the regions of the image. The second element $\mathbf{w} = \{w_1, \ldots, w_M\}$ is the collection of words of the caption.

We consider hierarchical probabilistic models of image/caption data which involve mixtures over underlying discrete and continuous variables. Conditional on the values of these latent variables, the region feature vectors are assumed to be distributed as a multivariate Gaussian distribution with diagonal covariance, and the caption words are assumed to be distributed as a multinomial distribution over the vocabulary.

We are interested in models that can perform three tasks: modeling the joint distribution of an image and its caption, modeling the conditional distribution of words given an image, and modeling the conditional distribution of words given a particular region of an image. The first task can be useful for clustering and organizing a large database of images. The second task is useful for automatic image annotation and text-based image retrieval. The third task is useful for automatically labeling and identifying a particular region of an image.

3 Generative models

In this section, we describe two hierarchical mixture models of image/caption data, noting their strengths and limitations with respect to the three tasks described above. We introduce a third model, *correspondence latent Dirichlet allocation*, in which the underlying probabilistic assumptions are appropriate for all three tasks.

3.1 A Gaussian-multinomial mixture model

We begin by considering a simple finite mixture model—the model on which most previous work in probabilistic modeling of multi-type data is based (Barnard et al., 2002; Brochu and de Freitas, 2003; Taskar et al., 2001). In this model—the *Gaussian-multinomial mixture* (GM-Mixture) shown in Figure 1—a single discrete latent variable z is used to represent a joint clustering of an image and its caption. As shown in the figure, an image/caption is assumed to be generated by first choosing a value of z, and then repeatedly sampling M region descriptions r_m and N caption words w_n conditional on the chosen value of z. The variable z is sampled once per image/caption, and is held fixed during the process of generating its components. The joint distribution of the hidden factor z and the image/caption $(\mathbf{r}, \mathbf{w})$ is:

$$p(z, \mathbf{r}, \mathbf{w}) = p(z \,|\, \lambda) \textstyle\prod_{n=1}^{N} p(r_n \,|\, z, \mu, \sigma) \prod_{m=1}^{M} p(w_m \,|\, z, \beta). \qquad (1)$$

Given a number of desired factors K and a corpus of images/captions, the parameters of a GM-Mixture model can be estimated by the EM algorithm[1] This yields K Gaussian distributions over features and K multinomial distributions over words

[1]In Section 4.2.1, we consider Bayesian versions of all of our models in which some the parameters are endowed with prior distributions.

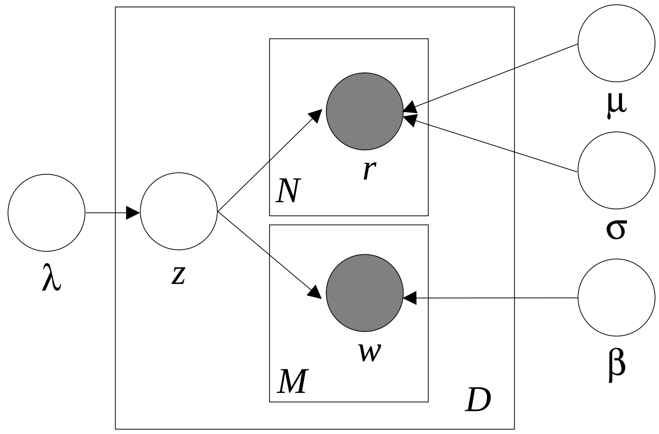


Figure 1: The GM-Mixture model of images and captions. Following the standard graphical model formalism, nodes represent random variables and edges indicate possible dependence. Shaded nodes are observed random variables; unshaded nodes are latent random variables. The joint distribution can be obtained from the graph by taking the product of the conditional distribution of nodes given their parents (see Eq. 1). Finally, the box around a random variable is a "plate," a notational device to denote replication. The box around r denotes N replicates of r (this gives the product in Eq. 1).

which together describe a clustering of the images/captions. Since each image and its caption are assumed to have been generated conditional on the same factor, the resulting multinomial and Gaussian parameters will correspond. An image with high probability under a certain factor will likely contain a caption with high probability in the same factor.

Let us consider the three tasks under this model. First, the joint probability of an image/caption can be computed by simply marginalizing out the hidden factor z from Eq. (1). Second, we can obtain the conditional distribution of words given an image by invoking Bayes' rule to find $p(z \mid \mathbf{r})$ and marginalizing out the hidden factor:

$$p(w \mid \mathbf{r}) = \textstyle\sum_z p(z \mid \mathbf{r}) p(w \mid z).$$

Finally, we would like to compute a region-specific distribution over words. This task, however, is beyond the scope of the GM-Mixture model. Conditional on the latent factor variable z, regions and words are generated independently, and the correspondence between specific regions and specific words is necessarily ignored.

3.2 Gaussian-Multinomial LDA

The *latent Dirichlet allocation* (LDA) model is a latent variable model that allows factors to be allocated repeatedly within a given document or image (Blei et al., 2003). Thus, different words in a document or different regions in an image can come from different underlying factors, and the document or image as a whole can be viewed as

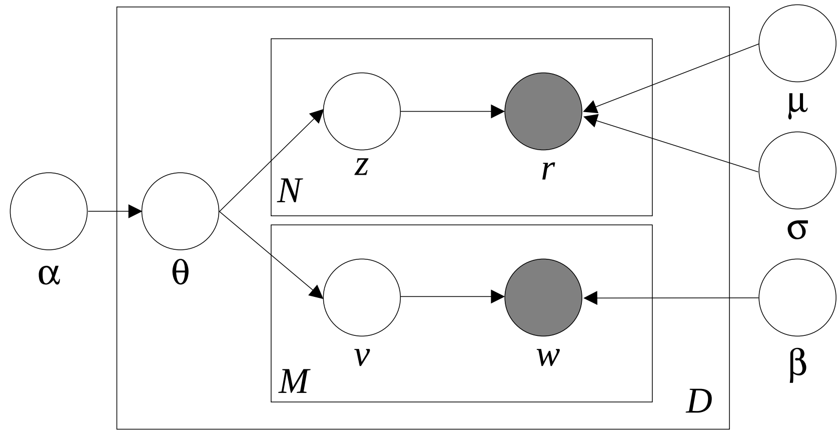


Figure 2: The GM-LDA model of images and captions. As opposed to the GM-Mixture model in Figure 1, each word and image region can potentially be drawn from a different hidden factor.

containing multiple "topics." This idea can be applied directly to the joint modeling of images and captions.

Gaussian-multinomial LDA (GM-LDA), shown in Figure 2, assumes the following generative process:

1. Sample a Dirichlet random variable θ; this provides a probability distribution over the latent factors, $\mathrm{Mult}(\theta)$.
2. For each of the N image regions:
   (a) Sample $z_n \sim \mathrm{Mult}(\theta)$.
   (b) Sample a region description r_n conditional on z_n.
3. For each of the M words:
   (a) Sample $v_m \sim \mathrm{Mult}(\theta)$.
   (b) Sample w_m conditional on v_m.

Note the importance of the plates. Within an image, all the region descriptions and words are generated with θ held fixed; the latent factors for each word and region description can (potentially) vary. Each new image/caption is generated by again selecting from the Dirichlet variable θ and repeating the entire process. Thus, we can view θ as a high-level representation of the ensemble of image/caption pairs in terms of a probability distribution over factors that each image/caption can be assembled from.

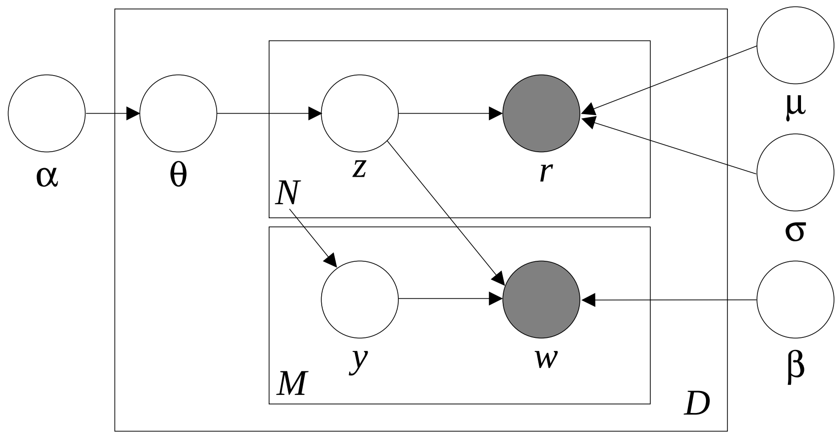


Figure 3: The graphical model representation of the CORR-LDA model. Note that the variables y_m are conditioned on N, the number of image regions.

The resulting joint distribution on an image/caption is given as follows:

$$p(\mathbf{r}, \mathbf{w}) = \int p(\theta \,|\, \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \,|\, \theta) p(r_n \,|\, z_n, \mu, \sigma) \right) \left(\prod_{m=1}^{M} \sum_{v_m} p(v_m \,|\, \theta) p(w_m \,|\, v_m, \beta) \right) d\theta.$$

As in the simpler LDA model, it is intractable to compute the conditional distributions of latent variables given observed data under this joint distribution, but efficient variational inference methods are available to compute approximations to these conditionals (see Barnard et al. 2002 for details). Furthermore, we can use variational inference methods to find the conditional probability $p(w \,|\, \mathbf{r})$ needed for image annotation/retrieval and the conditional probability $p(w \,|\, \mathbf{r}, r_n)$ needed for region labeling.

LDA provides significant improvements in predictive performance over simpler mixture models in the domain of text data (Blei et al., 2003), and we expect for GM-LDA to provide similar advantages over GM-MIXTURE. Indeed, we will see in Section 5 that GM-LDA does model the image/caption data better than GM-MIXTURE. We will also see, however, that good models of the joint probability of images and captions do not necessarily yield good models of the *conditional* probabilities that are needed for automatic annotation, text-based image retrieval, and region labeling. We will argue that this is due to the lack of a dependency between the latent factors z_n and v_m which respectively generated the images and their captions. In the next section, we turn to a model that aims to correct this problem.

3.3 Correspondence LDA

We introduce *correspondence LDA* (Corr-LDA) as a model that combines the flexibility of GM-LDA with the associability of GM-Mixture. With this model, we achieve simultaneous dimensionality reduction in the representation of region descriptions and words, while also modeling the conditional correspondence between their respective reduced representations.

Corr-LDA is depicted in Figure 3. The model can be viewed in terms of a generative process that first generates the region descriptions and subsequently generates the caption words. In particular, we first generate N region descriptions r_n from an LDA model. Then, for each of the M caption words, one of the regions is selected from the image and a corresponding caption word w_m is drawn, conditioned on the factor that generated the selected region.

Formally, let $\mathbf{z} = \{z_1, z_2, \ldots, z_N\}$ be the latent factors that generate the image, and let $\mathbf{y} = \{y_1, y_2, \ldots, y_M\}$ be discrete indexing variables that take values from 1 to N with equal probability. Conditioned on N and M, a K-factor Corr-LDA model assumes the following generative process for an image/caption $(\mathbf{r}, \mathbf{w})$:

1. Generate $\theta \sim \text{Dir}(\theta \mid \alpha)$.
2. For each image region r_n, $n \in \{1, \ldots, N\}$:
   (a) Generate $z_n \sim p(z_n \mid \theta)$
   (b) Generate $r_n \sim p(r \mid z_n, \mu, \sigma)$ from a multivariate Gaussian distribution conditioned on z_n.
3. For each caption word w_m, $m \in \{1, \ldots, M\}$:
   (a) Generate $y_m \sim \text{Unif}(1, \ldots, N)$
   (b) Generate $w_m \sim p(w \mid y_m, \mathbf{z}, \beta)$ from a multinomial distribution conditioned on the z_{y_m} factor.

Corr-LDA thus specifies the following joint distribution on image regions, caption words, and latent variables:

$$
\begin{aligned}
p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{y}) = p(\theta \mid \alpha) & \left(\prod_{n=1}^{N} p(z_n \mid \theta) p(r_n \mid z_n, \mu, \sigma) \right) \\
& \left(\prod_{m=1}^{M} p(y_m \mid N) p(w_m \mid y_m, \mathbf{z}, \beta) \right).
\end{aligned}
$$

The independence assumptions of the Corr-LDA model are a compromise between the extreme correspondence enforced by the GM-Mixture model, where the entire image and caption are conditional on the same factor, and the lack of correspondence in the GM-LDA model, where the image regions and caption words can conceivably be conditional on two disparate sets of factors. Under the Corr-LDA model, the regions of the image can be conditional on any ensemble of factors but the words of the caption must be conditional on factors which are present in the image. In effect, our model captures the notion that the image is generated first and the caption annotates the image.

Finally, note that the correspondence implemented by Corr-LDA is not a one-to-one correspondence, but is more flexible: all caption words could come from a subset of the image regions, and multiple caption words can come from the same region.

4 Inference and estimation

In this section, we describe approximate inference and parameter estimation for the Corr-LDA model. As a side effect of the inference method, we can compute approximations to our three distributions of interest : $p(w \mid \mathbf{r})$, $p(w \mid \mathbf{r}, r_n)$, and $p(\mathbf{w}, \mathbf{r})$.

4.1 Variational inference

Exact probabilistic inference for Corr-LDA is intractable; as before, we avail ourselves of variational inference methods (Jordan et al., 1999) to approximate the posterior distribution over the latent variables given a particular image/caption.

In particular, we define the following factorized distribution on the latent variables:

$$q(\theta, \mathbf{z}, \mathbf{y}) = q(\theta \mid \gamma) \left(\prod_{n=1}^{N} q(z_n \mid \phi_n) \right) \left(\prod_{m=1}^{M} q(y_m \mid \lambda_m) \right),$$

with free (variational) parameters γ, ϕ, and λ. Each variational parameter is appropriate to its respective random variable. Thus γ is a K-dimensional Dirichlet parameter, ϕ_n are N K-dimensional multinomial parameters, and λ_m are M N-dimensional multinomial parameters.

Following the general recipe for variational approximation, we minimize the KL-divergence between this factorized distribution and the true posterior thus inducing a dependence on the data $(\mathbf{r}, \mathbf{w})$. Taking derivatives with respect to the variational parameters, we obtain the following coordinate ascent algorithm:

1. Update the posterior Dirichlet parameters:

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}.$$

2. For each region, update the posterior distribution over factors. Note that this update takes into account the likelihood that each caption word was generated

by the same factor as the region:

$$\phi_{ni} \propto p(r_n \mid z_n = k, \mu, \sigma) \exp\{\mathrm{E}_q[\log \theta_i \mid \gamma]\}$$
$$\exp\left\{\textstyle\sum_{m=1}^{M} \lambda_{mn} \log p(w_m \mid y_m = n, z_m = i, \beta)\right\},$$

where $\mathrm{E}_q[\log \theta_i \mid \gamma] = \Psi(\gamma_i) - \Psi(\sum \gamma_j)$, and Ψ is the digamma function which can be approximated by Taylor expansion (Abramowitz and Stegun, 1970).

3. For each word, update the approximate posterior distribution over regions:

$$\lambda_{mn} \propto \exp\left\{\textstyle\sum_{i=1}^{K} \phi_{ni} \log p(w_m \mid y_m = n, z_n = i, \beta)\right\}.$$

These update equations are invoked repeatedly until the change in KL divergence is small.

With the approximate posterior in hand, we can find a lower bound on the joint probability, $p(\mathbf{w}, \mathbf{r})$, and also compute the conditional distributions of interest: $p(w \mid \mathbf{r})$ and $p(w \mid \mathbf{r}, r_n)$. In annotation, we approximate the conditional distribution over words as follows:

$$p(w \mid \mathbf{r}) \approx \sum_{n=1}^{N} \sum_{z_n} q(z_n \mid \phi_n) p(w \mid z_n, \beta).$$

In region labeling, the distribution over words conditioned on an image and a region is approximated by:

$$p(w \mid \mathbf{r}, r_n) \approx \sum_{z_n} q(z_n \mid \phi_n) p(w \mid z_n, \beta).$$

4.2 Parameter estimation

Given a corpus of image/caption data, $\mathcal{D} = \{(\mathbf{r}_1, \mathbf{w}_1), \ldots, (\mathbf{r}_D, \mathbf{w}_D)\}$, we find maximum likelihood estimates of the model parameters with a *variational EM* procedure that maximizes the lower bound on the log likelihood of the data induced by the variational approximation described above. In particular, the E-step computes the variational posterior for each image and caption given the current setting of the parameters. The M-step subsequently finds maximum likelihood estimates of the model parameters from expected sufficient statistics taken under the variational distribution. The variational EM algorithm alternates between these two steps until the bound on the expected log likelihood converges.

4.2.1 Smoothing with empirical Bayes

In Section 5, we show that overfitting can be a serious problem, particularly when working with the conditional distributions for image annotation. We deal with this issue by taking a more thoroughgoing Bayesian approach, imposing a prior distribution on the word multinomial parameters β. We represent β as a matrix whose columns are the K multinomial parameters for the latent factors. We treat each column as a sample from an exchangeable Dirichlet distribution $\beta_i \sim \mathrm{Dir}(\eta, \eta, \ldots, \eta)$ where η is a scalar parameter.

In place of a point estimate, we now have a smooth posterior $p(\beta \mid \mathcal{D})$. A variational approach can again be used to find an approximation to this posterior distribution (Attias, 2000). Introducing variational Dirichlet parameters ρ_i for each of the K multinomials, we find that the only change to our earlier algorithm is to replace the maximization with respect to β with the following variational update:

$$\rho_{ij} = \eta + \textstyle\sum_{d=1}^{D} \sum_{m=1}^{M} 1(w_{dm} = j) \sum_{n=1}^{N} \phi_{ni} \lambda_{mn},$$

and we replace all instances of β in the variational inference algorithm by $\exp\{\mathrm{E}[\log \beta \mid \rho]\}$.

Bayesian methods often assume a noninformative prior which, in the case of the exchangeable Dirichlet, means setting $\eta = 1$. With K draws from the prior distribution, however, we are in a good position to take the empirical Bayes perspective (Morris, 1983) and compute a maximum likelihood estimate of η. This amounts to a variant of the ML procedure for a Dirichlet with expected sufficient statistics under ρ. Analogous smoothing algorithms are also readily derived for the GM-Mixture and GM-LDA models.

5 Results

In this section, we present an evaluation of all three models on 7000 images and captions from the Corel database. We held out 25% of the data for testing purposes and used the remaining 75% to estimate parameters. Each image is segmented into 6-10 regions and is associated with 2-4 caption words. The vocabulary contains 168 unique terms.

5.1 Test set likelihood

To evaluate how well a model fits the data, we computed the per-image average negative log likelihood of the test set on all three models for various values of K. A model which better fits the data will assign a higher likelihood to the test set (*i.e.*, lower numbers are better in negative likelihood).

Figure 5 illustrates the results. As expected, GM-LDA provides a much better fit than GM-Mixture. Furthermore, Corr-LDA provides as good a fit as GM-LDA. This is somewhat surprising since GM-LDA is a less constrained model. However, both models have the same number of parameters; their similar performance indicates

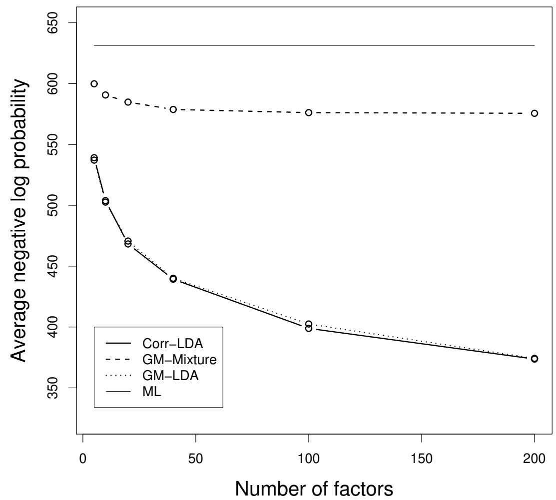


Figure 4: The per-image average negative log probability of the held-out test set as a function of the number of hidden factors (lower numbers are better). The horizontal line is the model that treats the regions and captions as an independent Gaussian and multinomial, respectively.

that, on average, the number of hidden factors used to model a particular image is adequate to model its caption.[2]

5.2 Automatic annotation

Given a segmented image without its caption, we can use the mixture models described in Section 3 to compute a distribution over words conditioned on the image, $p(w \mid \mathbf{r})$. This distribution reflects a prediction of the missing caption words for that image.

5.2.1 Caption perplexity

To measure the annotation quality of the models, we computed the *perplexity* of the given captions under $p(w \mid \mathbf{r})$ for each image in the test set. Perplexity, which is used in the language modeling community, is equivalent algebraically to the inverse of the geometric mean per-word likelihood (again, lower numbers are better):

$$\text{perplexity} = \exp\{-\textstyle\sum_{d=1}^{D} \sum_{m=1}^{M_d} \log p(w_m \mid \mathbf{r}_d) / \sum_{d=1}^{D} M_d\}.$$

[2]Empirically, when $K = 200$, we find that in only two images of the test set does the GM-LDA model use more hidden factors for the caption than it does for the image.

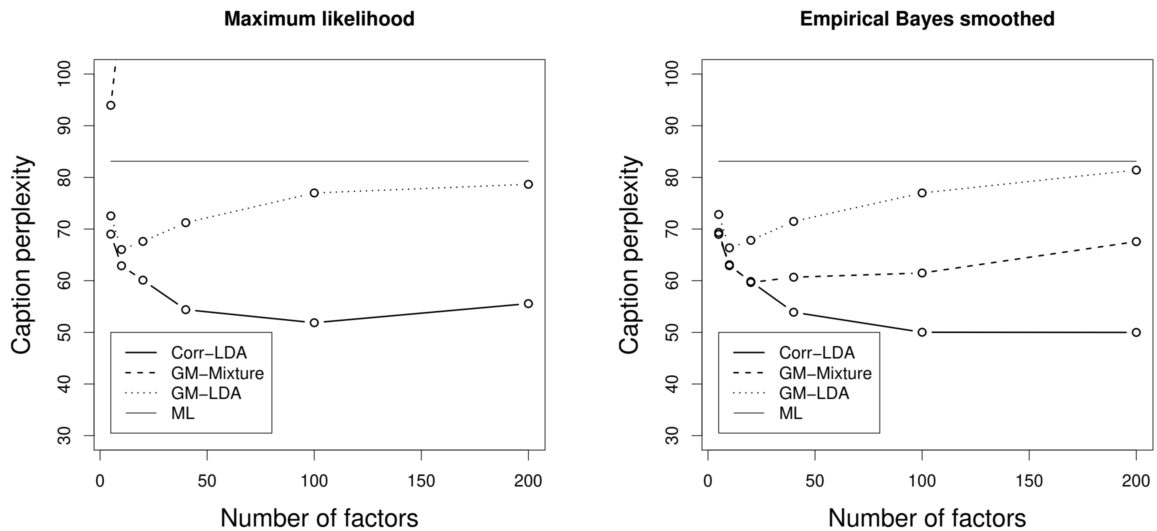


Figure 5: (Left) Caption perplexity on the test set for the ML estimates of the models (lower numbers are better). Note the serious overfitting problem in GM-Mixture (values for K greater than five are off the graph) and the slight overfitting problem in Corr-LDA. (Right) Caption perplexity for the empirical Bayes smoothed estimates of the models. The overfitting problems in GM-Mixture and Corr-LDA have been corrected.

Figure 5 (Left) shows the perplexity of the held-out captions under the maximum likelihood estimates of each model for different values of K. We see that overfitting is a serious problem in the GM-Mixture model, and its perplexity immediately grows off the graph (e.g., when $K = 200$, the perplexity is 2922). Note that in related work (Barnard et al., 2002), many of the models considered are variants of GM-Mixture and rely heavily on an ad-hoc smoothing procedure to correct for overfitting.

Figure 5 (Right) illustrates the caption perplexity under the smoothed estimates of each model using the empirical Bayes procedure from Section 4.2.1. The overfitting of GM-Mixture has been corrected. Once smoothed, it performs better than GM-LDA despite the GM-LDA model's superior performance in joint likelihood.

We found that GM-LDA does not provide good conditional distributions for two reasons. First, it is "over-smoothed." Computing $p(w \mid \mathbf{r})$ requires integrating a somewhat diffuse posterior (due to the small number of regions) over all the factor dimensions. Thus, the factors to which each region is associated are essentially washed out and, as K gets large, the model's performance approaches the performance of the simple ML estimate of the caption words.

Second, GM-LDA easily allows caption words to be generated by factors that did not contribute to generating the image regions (e.g., when $K = 200$, 54% of the

Figure 6: Example images from the test set and their automatic annotations under different models.

caption words in the test set are assigned to factors that do not appear in their corresponding images). With this freedom, the estimated conditional Gaussian parameters do not necessarily reflect regions that are correctly annotated by the corresponding conditional multinomial parameters. While it better models the joint distribution of words and regions, it fails to model the relationship between them.

Most notably, Corr-LDA finds much better predictive distributions of words than either GM-LDA or GM-Mixture. It provides as flexible a joint distribution as GM-LDA but guarantees that the image captions are generated by a subset of the factors that appear in their corresponding images. Furthermore, by allowing caption words to be allocated to different factors, the Corr-LDA model achieves superior performance to the GM-Mixture which is constrained to associating the entire image/caption to a single factor. Thus, with Corr-LDA, we can achieve a competetive fit of the joint distribution and find superior conditional distributions of words given images.

5.2.2 Annotation examples

Figure 6 shows ten sample annotations — the top five words from $p(w \mid \mathbf{r})$ — computed by each of the three models for $K = 200$. These examples illustrate the limitations

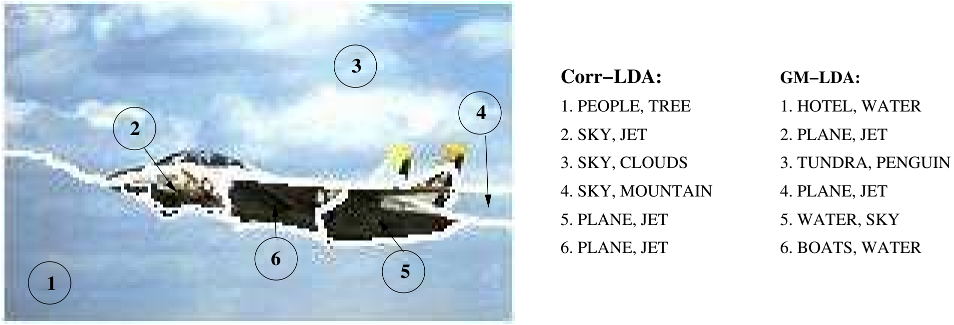


Figure 7: An automatic region labeling example from the test set.

and power of the probabilistic models described in Section 3 when used for a practical discriminative task.

The GM-LDA model, as shown quantitatively in the previous section, gives the least impressive performance of the three. First, we see washing out effect described above by the fact that many of the most common words in the corpus — words like "water" and "sky" — occur in the predicted captions for all of the pictures. Second, the predicted caption rarely predicts the object or objects that are in the picture. For example, it misses "jet" in the picture captioned *clouds, jet, plane*, a word that both other models predict with high accuracy.

The GM-Mixture model performs better than GM-LDA, but we see how this model relies on the average image features and can fail to predict a region in the image that may not generally occur in other similar images. For example, it omits "tree" from the picture captioned *scotland, water* since the trees are only a small part on the left side of the frame. Furthermore, the GM-Mixture predicts completely incorrect words if the average features do not easily correspond to a common theme. For example, the background of *fish, reefs, water* is not the usual blue and GM-Mixture predicts words like "fungus", "tree", and "flowers."

Finally, as reflected by the term perplexity results above, the Corr-LDA model gives the best performance and correctly labels most of the example pictures. Unlike the GM-Mixture model, it can consider each region of a photo independently and assign it to a different cluster. The final distribution over words reflects the ensemble of clusters which were assigned to the image regions. Thus, the Corr-LDA model finds the trees in the picture labeled *scotland, water* and can correctly identify the fish, even without its usual blue background.

As described in Section 3, the Corr-LDA and GM-LDA models can furthermore compute a region-based distribution over words, $p(w \mid \mathbf{r}, r_n)$. Figure 7 illustrates a sample region labeling on an image in the test set. Though both models hypothesize the word "plane" and "jet," Corr-LDA places them in the reasonable regions 2, 5, and 6 while GM-LDA places them in regions 2 and 4. Furthermore, Corr-LDA

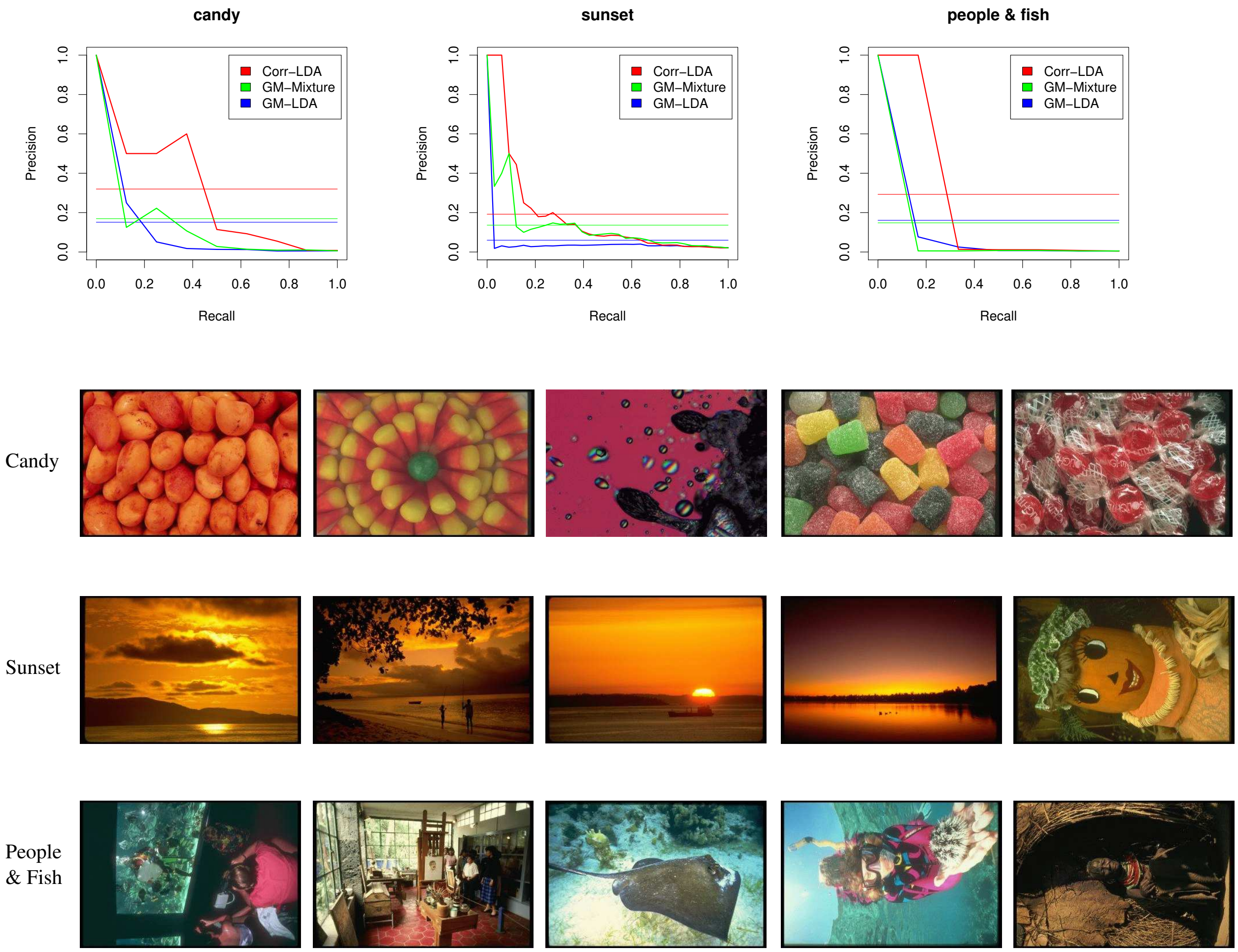


Figure 8: Three examples of text-based image retrieval. (Top) Precision/recall curves for three queries. The horizontal lines are the mean precision for each model. (Bottom) Top five results from a 100-factor Corr-LDA model for the same three queries.

recognizes the top region as "sky, clouds" while GM-LDA provides the enigmatic "tundra, penguin."[3]

5.3 Text-based image retrieval

There has been significant computer science research in *content-based image retrieval*, using a particular query image (possibly a sketch or primitive graphic) to find matching relevant images (Goodrum, 2000; Wang et al., 2001). Another line of research,

[3]We cannot quantitatively evaluate this task (*i.e.*, compute the region perplexity) because our data does not provide ground-truth for the region labels.

multimedia information retrieval, uses representations of different data types (such as text and images) to retrieve documents that contain both (Meghini et al., 2001).

Less attention, however, has been focused on *text-based image retrieval*, an arguably more difficult task where a user submits a text query to find matching images for which there is no related text. Previous approaches have essentially treated this task as a classification problem, handling specific queries from a vocabulary of about five words (Naphade and Huang, 2001). In contrast, by using the conditional distribution of words given an image, our approach can handle an open-ended vocabulary and arbitrary queries.

We use a unique form of the language modeling approach (Ponte and Croft, 1998) where the document language models are derived from images rather than words. To obtain an image-specific distribution over words, we use the conditional distribution $p(w \mid \mathbf{r})$ that we can compute for each of the models described in Section 3. This distribution provides us with a description of an image in word-space which we use to find images that are similar to the words of the query.

More formally, denote an N-word query by $\mathbf{q} = \{q_1, \ldots, q_N\}$. For each image $\mathbf{r}_i$, its score relative to the query is:

$$\text{score}_i = \textstyle\prod_{n=1}^{N} p(q_n \mid \mathbf{r}_i),$$

where $p(q_n \mid \mathbf{r}_i)$ is the probability of the nth query word under the distribution $p(w \mid \mathbf{r}_i)$. After computing the score for each image, we return a ranked list of images in descending order.

Figure 8 illustrates three queries performed on the three models with 200 factors and the held-out test set of images. We consider an image to be relevant if its true caption contains the query words (recall that we make no reference to the true caption in the retrieval process). As illustrated by the precision/recall curves, the Corr-LDA model achieves superior retrieval performance. It is particularly strong with difficult queries such as “people and fish.” In this example, there are only six relevant images of the 2000 image test set and two of them appear in the top five. This is due to the ability of Corr-LDA to assign different regions to different clusters. The model can independently learn the salient features of “fish” and “people” and effectively combine them to perform retrieval.

6 Summary

We have developed Corr-LDA, a powerful model for annotated data that combines the advantages of probabilistic clustering for dimensionality reduction with explicit modeling of the conditional distributions from which data annotations are generated. We have shown on image/caption data that this model can achieve a competitive joint likelihood and superior conditional distribution of words given an image.

Corr-LDA provides a clean probabilistic model for performing various information tasks associated with multi-type data such as images/captions. We have demonstrated its use in automatic image annotation, automatic image region annotation, and text-based image retrieval. It is important to note that this model is not specially tailored for image/caption data. Given good features, the Corr-LDA model can handle any kind of annotated data such as video/closed-captions, music/text, and gene/descriptions.

Acknowledgments

The authors would like to express their sincere thanks to David Forsyth and Kobus Barnard for numerous conversations and the well-curated data on which this work was based. This work was supported by a grant from the Intel Corporation, the National Science Foundation (NSF grant IIS-9988642), and the Multidisciplinary Research Program of the Department of Defense (MURI N00014-00-1-0637). David M. Blei was additionally supported by a grant from the Microsoft Corporation.

References

Abramowitz, M. and Stegun, I. (1970). *Handbook of Mathematical Functions*. Dover, New York.

Attias, H. (2000). A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*.

Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., and Jordan, M. (2002). Matching words and pictures. *Journal of Machine Learning Research*. (in press).

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

Brochu, E. and de Freitas, N. (2003). Name that song: A probabilistic approach to querying on music and text. In *Advances in Neural Information Processing Systems 15*.

Cohn, D. and Hofmann, T. (2001). The missing link—A probabilistic model of document content and hypertext connectivity. In *Advances in Neural Information Processing Systems 13*.

Goodrum, A. (2000). Image information retrieval: An overview of current research. *Informing Science*, 3(2):63–67.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Meghini, C., Sebastiani, F., and Straccia, U. (2001). A model of multimedia information retrieval. *Journal of the ACM (JACM)*, 48(5):909–970.

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–65. With discussion.

Naphade, M. and Huang, T. (2001). A probabilistic framework for semantic video indexing, filtering and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151.

Ponte, J. and Croft, B. (1998). A language modeling approach to information retrieval. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):888–905.

Taskar, B., Segal, E., and Koller, D. (2001). Probabilistic clustering in relational data. In *IJCAI-01*, pages 870–876.

Wang, J., Li, J., and Wiederhold, G. (2001). SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–983.