

Topic Regression Multi-Modal Latent Dirichlet Allocation for Image Annotation

Duangmanee Putthividhya
UCSD
9500 Gilman Drive
La Jolla, CA 92307
putthi@ucsd.edu

Hagai T. Attias
Golden Metallic, Inc.
P. O. Box 475608
San Francisco, CA 91147
htattias@goldenmetallic.com

Srikantan S. Nagarajan
UCSF
513 Parnassus Avenue
San Francisco, CA 94143
sri@radiology.ucsf.edu

Abstract

We present topic-regression multi-modal Latent Dirichlet Allocation (tr-mmLDA), a novel statistical topic model for the task of image and video annotation. At the heart of our new annotation model lies a novel latent variable regression approach to capture correlations between image or video features and annotation texts. Instead of sharing a set of latent topics between the 2 data modalities as in the formulation of correspondence LDA in [2], our approach introduces a regression module to correlate the 2 sets of topics, which captures more general forms of association and allows the number of topics in the 2 data modalities to be different. We demonstrate the power of tr-mmLDA on 2 standard annotation datasets: a 5000-image subset of COREL and a 2687-image LabelMe dataset. The proposed association model shows improved performance over correspondence LDA as measured by caption perplexity.

1. Introduction

Image and video retrieval has long been an important area of research in computer vision. Traditional methods in multimedia retrieval have focused on a query-by-example paradigm, also known as content-based retrieval, where users submit an image or video query, and the system returns an item in the database closest in some distance measure to the query. As today’s multimedia content becomes increasingly multi-modal with texts accompanying images and videos in the form of content description, transcribed text, or captions, current state-of-the-art multimedia search technology relies heavily on these collateral annotation texts to identify and retrieve images and video. Besides the fact that users often prefer the use of textual queries over examples, an important benefit of such an approach is the high-level semantic retrieval, e.g. retrieval of abstract concepts, that could not be achieved with low-level visual cues used in most query-by-example systems.

With annotation texts playing an increasingly vital role in modern multimedia retrieval systems, a relevant question

one might ask is how to deal with numerous fast-growing user-generated content that often lacks descriptive annotation texts which would enable accurate semantic retrieval to be performed. The traditional solution is to employ manual labeling—a process that is costly and unscalable to large-scale repositories. With recent unprecedented availability of image and video data online, there is a growing demand to bypass the human intervention and develop automated tools that can generate semantic descriptors of multimedia content—automatic annotation systems. Given a database of images and their corresponding annotation words that can be used for training, the task of an automatic annotation algorithm is to learn patterns of image-text (or video-text) association so that when presented with an un-annotated image, the system can accurately infer the missing annotation.

Previous work on image and video annotation can be broadly summarized into 2 groups. The first line of work considers a discriminative approach and cast the problem of automatic annotation as a classification problem, treating annotation words as class labels [6, 10]. By directly modeling the conditional distribution of image features given annotation words (class-conditional density), this line of approach makes no attempt to uncover correlation structures in annotation texts that can be useful when predicting annotation. Another set of techniques consider probabilistic latent variable models for the task of multimedia annotation. By postulating the existence of a small set of hidden factors that govern the association between the 2 data types, the latent variable representations learned under such an assumption are ensured to be useful in predicting the missing captions given the corresponding image. Several modeling variations have been proposed, see [1, 2, 9, 8], with the different forms of probability distribution assumed for caption words (multinomial vs. bernoulli) and image features (mixture of Gaussian vs. non-parameteric density estimation).

In the specific case of statistical topic models applied to the task of image annotation, the seminal work of Blei *et al.* [2] proposed 2 association models—multi-modal LDA (mmLDA) and correspondence LDA (cLDA)—which ex-

tend the basic Latent Dirichlet Allocation (LDA) model to learn the joint distribution of texts and image features. In order to capture correlations between the two modalities, the association models use a set of shared latent variables to represent the underlying causes of cross-correlations in the data. Indeed, with the same forms of probabilistic distributions assumed, the two models differ only in their choices of shared latent variables. In mmLDA [1, 2], the mean topic proportion variable is shared between the 2 modalities, potentially leading to an undesirable scenario where some topics are used entirely to explain either image features or caption words. To ensure the same sets of topics are used to generate corresponding data in the 2 modalities, under cLDA each caption word directly shares a hidden topic variable with a randomly selected image region. Better prediction results are reported in [2] as a result of a tighter association enforced in sharing the hidden topics.

In this work, we are interested in alternative methods in capturing statistical association between image and text. Instead of sharing the hidden topics as in the formulation of cLDA, our model which we call Topic-Regression Multi-Modal Latent Dirichlet Allocation (tr-mmLDA), learns 2 separate sets of hidden topics and a regression module which allows one set of topics to be linearly predicted from the other. More specifically, we introduce a linear Gaussian regression module where the proportion of topic variable for annotation texts is the response variable and is modeled as a noise corrupted version of a linear combination of the image topic variable. Inspired by the good predictive performance of the supervised LDA model [4], we adopt the empirical image topic frequency covariates to ensure that only the topics that actually occur in the image modality are used in predicting caption texts. Our proposed formulation can capture varying degrees of correlations between the two data modalities and allows the number of hidden topics in images to be different from that of caption texts. We derive an efficient variational inference algorithm which relies on a mean-field approximation to handle intractable posterior computations. To demonstrate the predictive power of the new association model, we compare image annotation performance on 2 standard datasets: a 5000-image subset of COREL and 2687-image 8-category subset of the LabelMe dataset. Our results are indeed more superior to cLDA as measured by caption perplexity.

The paper is organized as follows. Section 2 describes the representation of images and caption texts used in all topic models discussed in this work. In section 3, we review the association models of mmLDA and cLDA and discuss their strengths and weaknesses with respect to an image annotation task. We describe the details of our proposed model and show the derivation of our variational inference algorithm. In section 5, we present experimental results on an image annotation task and conclude the paper.

2. Data Representation and Notations

Inspired by the success of the recent work on scene modeling [7, 12], we borrow a tool from statistical text document analysis and adopt a bag-of-words representation for both image and text. In such a representation, word ordering is ignored and a document is simply reduced to a vector of word count. A multimedia document consisting of an image and the corresponding caption text is thus summarized in our representation as a pair of vectors of word counts. An image word is denoted as a unit-basis vector r of size T_r with exactly one non-zero entry representing the membership to only one word in a dictionary of T_r words. A caption word w_n is similarly defined for a dictionary of size T_w . An image is a collection of N word occurrences denoted by $\mathbf{R} = \{r_1, r_2, \dots, r_N\}$; the caption text is a collection of M word occurrences denoted by $\mathbf{W} = \{w_1, w_2, \dots, w_M\}$. A training set of D image-caption pairs is denoted as $\{\mathbf{R}_d, \mathbf{W}_d\}, d \in \{1, 2, \dots, D\}$.

3. Probabilistic Models

All the topic models discussed in this work builds on Latent Dirichlet Allocation (LDA) [5] which is a powerful generative model for modeling words in documents. Under LDA, words in the same document are allowed to exhibit characteristics from multiple components (topics). A document, which is a collection of words, is then summarized in terms of the components' overall relative influences on the collection. To this end, LDA employs 2 sets of latent variables for each document as seen in Fig 1(a): (i) discrete-valued hidden variables z_n which assign each word to one of the K components and (ii) a latent variable θ that represents the random proportion of the components' influence in the document. In more specific terms, LDA decomposes the distribution of word counts for each document into contributions from K topics and model the proportion of topics as a Dirichlet distribution, while each topic, in turn, is a multinomial distribution over words.

3.1. Multi-modal LDA (mmLDA) and Correspondence LDA (cLDA)

In order to extend the basic LDA model to learn the joint correlations between data of different types, a traditional solution under a probabilistic framework is to assume the existence of a small set of shared latent variables that are the common causes of correlations between the 2 modalities. This is precisely the design philosophy behind multi-modal LDA (mmLDA) and correspondence LDA (cLDA) [2], which extend LDA to describe the joint distributions of image and caption words in multimedia documents. While adopting the same core assumption of LDA in allowing words in a document to be generated from multiple topics, the two extension models differ in their choices of latent

variables being shared between images and texts. Originally proposed in [1], mmLDA postulates that the mean topic proportion variable θ is the common factor that generates the two types of words. By forcing the topic proportion to be the same in image and caption modality, the 2 sets of Multinomial topic parameters therefore are assumed to correspond. However, the decision to share θ between the two data modalities implies that image and caption words become independent conditioned on θ . Indeed without the plate notation as depicted in Fig 1(b), it is not hard to see that the association of mmLDA assumes that image and caption words are exchangeable—a key assumption which allows words from the two data modalities to potentially be generated from non-overlapping sets of hidden topics. As K becomes large, annotation experiments on the COREL dataset in [2] show that more than 50% of the caption words are assigned to topics that do not occur in the corresponding images, rendering the knowledge about the image modality essentially half useless at predicting the missing caption words. The flexibility of mmLDA provides a good fit for the joint distribution of the data but is a bad fit for a prediction task, hence a poor annotation performance.

To ensure that only the set of topics that actually generate the image words are those used in generating the caption words, correspondence LDA (cLDA) as seen in Fig 1(c) was designed so that image is the primary modality and is generated first; conditioned on the topics used in the image, caption words are then generated. More specifically, by forcing each caption word to directly share a hidden topic with a randomly selected image word, cLDA guarantees that the topics in caption texts are indeed a subset of the topics that occur in the corresponding image. Note that while each caption word is restricted to be associated with one particular image region, the association of cLDA does allow the same image region to be associated with multiple caption words, accounting for the scenario where more than one caption words are used to describe a single object in the image.

Despite a good annotation performance as reported in [2], the constrained association of cLDA proves to be too restrictive in practice. When dealing with annotation words that globally describe the scene as a whole, the association model that restricts each caption word to one image region can be very inaccurate. A more powerful association model should indeed allow the captions words to be influenced by topics from all image regions as well as those from a particular subset of regions. In addition, by sharing the discrete-valued latent topic variables directly between the two modalities, cLDA provides no mechanism to allow different number of topics to be used in modeling images and caption texts. In the next section, we will describe a more flexible association model that address these limitations of cLDA while still maintaining a good predictive power.

3.2. Topic-regression Multi-modal LDA (tr-mmLDA)

To get past the issues associated with sharing latent variables between data modalities, in this work we explore a novel approach in modeling correlations between data of different types. Instead of using a set of shared latent variables to explain correlations in the data, we propose a latent variable regression approach to correlate latent variables of the two modalities; designed with the prediction task in mind, our framework thus allows latent variables of one type to be predicted from latent variables of another type. In the specific case of extending LDA to learn correlations between image and caption words, we propose a formulation that uses 2 separate topic models one for each data modality and introduce a regression module to correlate the 2 sets of hidden topics. To this end, we draw insights from several recent topic models [3, 11] and adopt a linear Gaussian regression module which takes in the image topic proportion as its input and target the hidden topic proportion for annotation texts as the response variable (in line with the task of predicting captions given an image). Our approach is similar in spirit to the way topic correlations are captured in Independent Factor Topic Models in [11] by explicitly modeling the independent sources and linearly combining them to obtain the correlated topic vectors. In our case, the hidden sources of topic correlations in caption data correspond to the hidden topics of the image modality.

Our model which we call a topic-regression multi-modal Latent Dirichlet Allocation (tr-mmLDA) has the graphical representation as shown in Fig 1(d). Given K image topics and L text topics, from an image side we have an LDA model with hidden topics $\mathbf{Z} = \{z_1, z_2, \dots, z_N\}$ and topic proportion θ . A real-valued topic proportion variable for caption text $\mathbf{x} \in \mathcal{R}^L$ is given by: $\mathbf{x} = \mathbf{A}\bar{\mathbf{z}} + \mu + \mathbf{n}$, where \mathbf{A} is an $L \times K$ regression coefficients matrix, μ is a vector of the mean parameters, $\mathbf{n} \sim \mathcal{N}(\mathbf{n}; 0, \mathbf{\Lambda})$ is a zero-mean uncorrelated Gaussian noise with a diagonal precision matrix $\mathbf{\Lambda}$. Instead of regressing over the mean topic proportion variable θ as done in mmLDA, we follow the formulation in supervised LDA in [4, 14] and adopt the empirical topic frequency covariates $\bar{\mathbf{z}} = \frac{1}{N} \sum_n z_n$ as an input into our regression module so that the topic proportion of annotation data depends directly on the actual topics that do occur in the image. Given \mathbf{x} , the topic proportion of caption text η is deterministically obtained via a softmax transformation of \mathbf{x} , *i.e.* the probability of observing topic l is given by $\eta_l = \frac{\exp(x_l)}{\sum_{k=1}^L \exp(x_k)}$. The generative process of tr-mmLDA for an image-caption pair with N visual words and M caption words is given as follows:

- Draw an image topic proportion $\theta | \alpha \sim \text{Dir}(\alpha)$
- For each image word $r_n, n \in \{1, 2, \dots, N\}$
 1. Draw topic assignment $z_n = k | \theta \sim \text{Mult}(\theta_k)$
 2. Draw visual word $r_n = t | z_n = k \sim \text{Mult}(\beta_{kt}^r)$
- Given the empirical image topic proportion $\bar{\mathbf{z}} = \frac{1}{N} \sum_{n=1}^N z_n$, we sample a real-valued topic proportion vari-

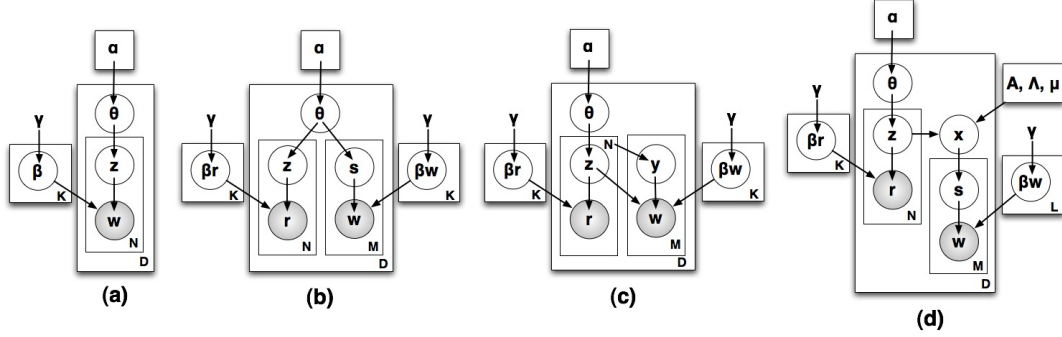


Figure 1. Graphical model representations for (a) Latent Dirichlet Allocation (LDA) and 3 extensions of LDA for the task of image annotation: (b) Multi-modal LDA (mmLDA) (c) correspondence LDA (cLDA) (d) Topic-regression Multi-modal LDA (tr-mmLDA).

able for caption text: $\mathbf{x}|\bar{\mathbf{z}}, \mathbf{A}, \mu, \Lambda \sim \mathcal{N}(\mathbf{x}; \mathbf{A}\bar{\mathbf{z}} + \mu, \Lambda)$.

- Compute topic proportion $\eta_l = \frac{\exp(x_l)}{\sum_{k=1}^L \exp(x_k)}$.
- For each caption word $w_m, m \in \{1, 2, \dots, M\}$
 1. Draw topic assignment $s_m = l|\eta \sim \text{Mult}(\eta_l)$
 2. Draw caption word $w_m = t|s_m = l \sim \text{Mult}(\beta_{it}^w)$

The formulation of tr-mmLDA can be seen as linking the LDA model for images and the IFTM model [11] for caption texts using a linear regression module, which is a flexible way of capturing correlations in the data. Under this framework, varying degrees of correlations can be captured by adapting the regression coefficients in the matrix \mathbf{A} accordingly. When the 2 data modalities are independent, the coefficients in \mathbf{A} are driven close to 0 and tr-mmLDA is reduced to 2 independent topic models one for each data modality. As correlations between the 2 data types grows stronger, more regression coefficients in the matrix \mathbf{A} will take values further away from 0. Indeed, correspondence LDA (cLDA) can be derived as a special case of tr-mmLDA by fixing the regression coefficients matrix \mathbf{A} to an identity matrix (assuming $K = L$) and setting the diagonal entries of precision matrix Λ to ∞ , which has the effect of forcing the empirical topic proportions in the 2 data modalities to be identical. Note that as the regression coefficient matrix \mathbf{A} moves away from an identity matrix (with more non-zero off-diagonal entries), tr-mmLDA allows the hidden topics from more than one image region to collectively exert influence on each caption word, which depicts a more accurate relationship for annotation words that globally describe the scene as a whole. Our framework in capturing correlations is thus more flexible than cLDA and allow more general forms of correlation to be modeled.

One important additional benefit in employing 2 sets of hidden topics is the flexibility in allowing the number of topics in the two data modalities to be different. Indeed as we shall show in the experimental results, different statistical structures in image features and caption texts often result in different optimal number of topics when fitting a topic model to each modality separately. By restricting the number of the topics to be the same, we might end up with

K that overfits the data in one modality while underfits the other modality. Our regression approach to model correlations gives the flexibility in exploring the optimal numbers of topics for each data modality separately.

4. Variational EM

To learn parameters of tr-mmLDA that maximizes the likelihood of the training data, we employ the Expectation Maximization (EM) framework that iteratively estimates the model parameters of latent variable models. Using Jensen’s inequality, the E step of the EM algorithm derives an auxiliary function which tightly lower-bounds the data likelihood function to allow for a more simple optimization to be performed in the M step. Indeed, for most probabilistic models involving latent variables, computing the exact posterior distribution over latent variables to obtain a tight likelihood lower-bound in the E step becomes computationally intractable. In variational EM, we replace the exact inference in the E step with an approximate inference algorithm. Variational EM framework thus alternates between computing a strict likelihood lower bound in the variational E step, and maximizing the bound to obtain a new parameter estimate in the M step.

4.1. Variational Inference

To infer the posterior over hidden variables, we begin with the expression of the true log-likelihood for an image-caption pair $\{\mathbf{W}, \mathbf{R}\}$:

$$\log p(\mathbf{W}, \mathbf{R}|\Psi) \geq \int q(\mathbf{Z}, \theta, \mathbf{x}, \mathbf{S}) \{ \log p(\mathbf{W}, \mathbf{R}, \mathbf{Z}, \theta, \mathbf{x}, \mathbf{S}|\Psi) - \log q(\mathbf{Z}, \theta, \mathbf{x}, \mathbf{S}) \} d\mathbf{Z}d\theta d\mathbf{x}d\mathbf{S} = \mathcal{F}, \quad (1)$$

where Ψ denotes the model parameters for tr-mmLDA $\{\beta^r, \beta^w, \gamma, \mathbf{A}, \mu, \Lambda\}$. Using the concavity of the log function, we apply Jensen’s inequality and derive a lower bound of the log-likelihood as seen in (1). Indeed, equality holds when the posterior over the hidden variables $q(\mathbf{Z}, \theta, \mathbf{x}, \mathbf{S})$ equals the true posterior $p(\mathbf{Z}, \theta, \mathbf{x}, \mathbf{S}|\mathbf{W}, \mathbf{R})$. Like in LDA, computing the exact joint posterior is computationally intractable; we employ a mean-field varia-

tional approximation to approximate the joint posterior distribution with a variational posterior in a factorized form: $p(\mathbf{Z}, \theta, \mathbf{x}, \mathbf{S} | \mathbf{w}, \mathbf{R}) \approx \prod_n q(z_n) \prod_m q(s_m) q(\theta) q(\mathbf{x})$. With such a posterior, the RHS of (1) becomes a strict lower bound of the data likelihood. The goal of the variational E step now is to find within a family of factorized distributions the variational posterior that maximizes the lower bound. Writing out the likelihood lower bound \mathcal{F} on the right hand side of (1), we obtain the following expression:
$$\mathcal{F} = \sum (E[\log p(r_n | z_n, \beta^r)] + E[p(z_n | \theta)]) + E[\log p(\theta | \alpha)] + \sum_n (E[\log p(w_m | s_m, \beta^w)] + E[\log p(s_m | \mathbf{x})]) + E[\log p(\mathbf{x} | \bar{\mathbf{z}}, \mathbf{A}, \mathbf{\Lambda}, \mu)] + \mathcal{H}(q(\mathbf{Z})) + \mathcal{H}(q(\theta)) + \mathcal{H}(q(\mathbf{S})) + \mathcal{H}(q(\mathbf{x})), \quad (2)$$

where the expectations are taken with respect to the factorized posteriors. $\mathcal{H}(p(x))$ denotes the entropy of $p(x)$. The fifth expectation term in (2)

$$E_{q(\mathbf{x})}[\log p(s_m = l | \mathbf{x})] = E_{q(\mathbf{x})}[x_l - \log(\sum_j e^{x_j})] \quad (3)$$

contains a normalization term from the softmax operation that will be difficult to evaluate in closed-form, regardless of the form of the variational posterior $q(\mathbf{x})$. We make use of convex duality and represents a convex function ($-\log(\cdot)$ function) as a point-wise supremum of linear functions. More specifically, the log normalization term is replaced with adjustable lower bounds parameterized by a convex variational parameter ξ :

$$x_l - \log \sum_j e^{x_j} \geq x_l - \log \xi - \frac{1}{\xi} \sum_j e^{x_j} + 1. \quad (4)$$

Under the diagonality assumption of $\mathbf{\Lambda}$ and the use of convex variational bound of the log-normalizer term in (4), the free-form maximization of \mathcal{F} w.r.t $q(\mathbf{x})$ results in the variational posterior $q(\mathbf{x})$ automatically taking on a factorized form $q(\mathbf{x}) = \prod_l q(x_l)$. However, $q(x_l)$ obtained by the free-form maximization is not in the form of a distribution that we recognize. We thus approximate $q(x_l)$ as a Gaussian distribution: $q(x_l) \sim \mathcal{N}(x_l; \bar{x}_l, \gamma_l^{-1})$ with mean parameter \bar{x}_l and precision γ_l . To simplify the notation, we denote $q(z_n = k)$ as ϕ_{nk} and $q(s_m = l)$ as η_{ml} . Since the prior $p(\theta | \alpha)$ is a Dirichlet distribution, which is a conjugate prior to the multinomial distribution, we can conclude that the posterior $q(\theta)$ is also a Dirichlet distribution. More specifically, we denote the posterior Dirichlet parameters as $\tilde{\alpha}$: $q(\theta) \sim \text{Dir}(\tilde{\alpha})$. By taking the expectation with respect to the variational posterior, we can write out the terms in the lower bound \mathcal{F} explicitly as a function of the variational parameters. The first two terms of \mathcal{F} in (2) are given by:

$$\sum_{n,k} \phi_{nk} \sum_t 1(r_n = t) \log \beta_{kt}^r + \sum_{n,k} \phi_{nk} E_{q(\theta)}[\log \theta_k] \quad (5)$$

where $E_{q(\theta)}[\log \theta_k] = \Psi(\tilde{\alpha}_k) - \Psi(\sum_j \tilde{\alpha}_j)$, with $\Psi(x)$ denoting the first derivative of the log-gamma function $\frac{\partial \log \Gamma(x)}{\partial x}$. The third term in (2) can be written as:

$$\log \Gamma(\sum_j \alpha_j) - \sum_j \log \Gamma(\alpha_j) + \sum_j (\alpha_j - 1) E_{q(\theta)}[\log \theta_j].$$

By evaluating the expectation with respect to a Gaussian posterior $q(x_j) \sim \mathcal{N}(x_j; \bar{x}_j, \gamma_j)$, we have that $E_{q(x_j)}[e^{x_j}] = e^{\bar{x}_j + \frac{0.5}{\gamma_j}}$ and the fourth and fifth expectation terms in (2) can be written as:

$$\sum_{m,l} \eta_{ml} \sum_t 1(w_m = t) \log \beta_{lt}^w + \sum_{m,l} \eta_{ml} \bar{x}_l - M \log \xi - \frac{M}{\xi} \sum_j e^{\bar{x}_j + \frac{0.5}{\gamma_j}} + M. \quad (6)$$

Making use of the following expectation $E[\mathbf{x}^\top \mathbf{\Lambda} \mathbf{x}] = \bar{\mathbf{x}}^\top \mathbf{\Lambda} \bar{\mathbf{x}} + \text{tr}(\mathbf{\Lambda} \mathbf{\Gamma}^{-1})$, the sixth term in (2) is given by:

$$-\frac{1}{2} ((\bar{\mathbf{x}} - \mu)^\top \mathbf{\Lambda} (\bar{\mathbf{x}} - \mu) + \text{tr}(\mathbf{\Lambda} \mathbf{\Gamma}^{-1})) - 2(\bar{\mathbf{x}} - \mu)^\top \mathbf{\Lambda} \mathbf{A} E[\bar{\mathbf{z}}] + E[\bar{\mathbf{z}}^\top \mathbf{A}^\top \mathbf{\Lambda} \mathbf{A} \bar{\mathbf{z}}], \quad (7)$$

where $E[\bar{\mathbf{z}}] = \frac{1}{N} \sum_n \phi_n$ and $E[\bar{\mathbf{z}}^\top \mathbf{A}^\top \mathbf{\Lambda} \mathbf{A} \bar{\mathbf{z}}]$ is evaluated to be $\text{tr}(\mathbf{A}^\top \mathbf{\Lambda} \mathbf{A} \frac{1}{N^2} (\sum_n \text{diag}(\phi_n) + \sum_n \phi_n \sum_{m \neq n} \phi_m^\top))$.

To update these variational parameters, we employ a coordinate ascent algorithm where we update one set of parameters while holding the rest fixed. By computing the gradient of \mathcal{F} w.r.t. ϕ_n and set the derivative to 0, we obtain the following update rule for ϕ_n :

$$\log \phi_n = \sum 1(r_n = t) \log \beta_{t}^r + E[\log \theta] + \frac{1}{N} \mathbf{A}^\top \mathbf{\Lambda} (\bar{\mathbf{x}} - \mu) - \frac{1}{2N^2} \text{diag}(\mathbf{A}^\top \mathbf{\Lambda} \mathbf{A}) - \frac{1}{N^2} \mathbf{A}^\top \mathbf{\Lambda} \mathbf{A} \sum_{m \neq n} \phi_m. \quad (8)$$

The variational parameters $\eta_{ml}, \xi, \tilde{\alpha}_k$ can be similarly re-estimated, resulting in the following closed-form updates:

$$\log \eta_{ml} = \sum 1(w_m = t) \log \beta_{lt}^w + \bar{x}_l, \quad (9)$$

$$\xi = \sum_t e^{\bar{x}_t + \frac{0.5}{\gamma_t}}, \quad (10)$$

$$\tilde{\alpha}_k = \sum_n \phi_{nk} + \alpha_k. \quad (11)$$

To update the parameters of the variational posterior $q(x_l) \sim \mathcal{N}(x_l; \bar{x}_l, \gamma_l)$, we differentiate \mathcal{F} w.r.t. \bar{x}_l and obtain the following expression for the gradient:

$$\frac{\partial \mathcal{F}}{\partial \bar{x}_l} = \sum_m \eta_{ml} - \frac{M}{\xi} e^{\frac{0.5}{\gamma_l} + \bar{x}_l} - \lambda_l (x_l - \mu_l - \mathbf{a}_l^\top E[\bar{\mathbf{z}}]). \quad (12)$$

However, the value of \bar{x}_l that makes the gradient vanish cannot be obtained in closed-form. We employ a Newton

algorithm that finds a zero-crossing solution for (12) efficiently. First, by substituting $\sum_m \frac{\eta_{ml}}{\lambda_l} - \bar{x}_l + a_l^\top E[\bar{\mathbf{z}}] + \mu_l$ with t_l , we can re-write the expression in (12) as follows:

$$t_l e^{t_l} = \frac{M}{\xi \lambda_l} e^{\frac{0.5}{\gamma_l}} \cdot e^{\frac{\sum_m \eta_{ml}}{\lambda_l} + a_l^\top E[\bar{\mathbf{z}}] + \mu_l} = u_l. \quad (13)$$

The Newton update rule for t_l is thus given by:

$$t_l^n = t_l^o + \frac{u_l e^{-t_l^o} - t_l^o}{t_l^o + 1}. \quad (14)$$

Starting from a good initial solution, the Newton algorithm converges in just a few iterations. The precision parameter γ_l can be similarly updated using a fast Newton algorithm, with the gradient given as:

$$\frac{\partial \mathcal{F}}{\partial \gamma_l^{-1}} = -\frac{\lambda_l}{2} - \frac{M}{2\xi} e^{\bar{x}_l} \cdot e^{\frac{\gamma_l}{2}} + \frac{1}{2\gamma_l^{-1}} = 0. \quad (15)$$

4.2. Parameter Estimation

Closed-form parameter updates can be obtained for all our model parameters. Again, by taking a derivative of the lower-bound objective function \mathcal{F} w.r.t the regression parameters and set the derivative to 0, the re-estimation equations can be written as follow:

$$\mathbf{A} = \left(\sum_d (\bar{\mathbf{x}}_d - \mu) E[\bar{\mathbf{z}}_d]^\top \right) \left(\sum_d E[\bar{\mathbf{z}}_d \bar{\mathbf{z}}_d^\top] \right)^{-1}, \quad (16)$$

$$\mu = \frac{1}{D} \sum_d (\bar{\mathbf{x}}_d - \mathbf{A} E[\bar{\mathbf{z}}_d]), \quad (17)$$

$$\mathbf{\Lambda}^{-1} = \frac{1}{D} \sum_d \left((\bar{\mathbf{x}}_d - \mu)(\bar{\mathbf{x}}_d - \mu)^\top + \mathbf{\Gamma}_d^{-1} - \mathbf{A} E[\bar{\mathbf{z}}_d] \bar{\mathbf{x}}_d^\top \right). \quad (18)$$

The Multinomial parameters for each topic can be re-estimated with the following update rules:

$$\beta_{lt}^w = \frac{\sum_{d=1}^D \sum_m \eta_{ml}^d 1(w_m^d = t)}{\sum_t \sum_{d=1}^D \sum_m \eta_{ml}^d 1(w_m^d = t)}, \quad (19)$$

$$\beta_{kt}^r = \frac{\sum_{d=1}^D \sum_n \phi_{nk}^d 1(r_n^d = t)}{\sum_t \sum_{d=1}^D \sum_n \phi_{nk}^d 1(r_n^d = t)}. \quad (20)$$

5. Experimental Results

We test our model on 2 standard datasets for image annotation: COREL and LabelMe datasets. The 5,000 image subset of the COREL dataset is the same subset used in the annotation experiments in [6]. This subset contains 50 classes of images, with 100 images per class. Each image in the collection is reduced to size 117×181 (or 181×117). 4,500 images are used in training (90 per class), and 500 for testing (10 per class). Each image is treated as a collection of 20×20 patches obtained by sliding a window with a 20-pixel interval, resulting in 45 patches per image.

For the LabelMe dataset, following the work in [14], we use the 8-category subset which contains 2,687 images from the classes: ‘coast’, ‘forest’, ‘highway’, ‘inside city’, ‘mountain’, ‘open country’, ‘street’, and ‘tall building’. 80% of the data in each class (2,147 images total) are used for training and 20% for testing (540 total). Each image is of size 256×256 . Again we use a 20×20 patch with a 20-pixel interval, resulting in 144 patches per image.

Following the work in [7], we use the 128-dim SIFT descriptor computed on 20×20 gray-scale patches. In addition, we follow the work in [13] and add additional 36-dim robust color descriptors which have been designed to complement the SIFT descriptors extracted from the gray-scale patches. We run k-means on a collection of 164-dim features to learn a dictionary of $T_r = 256$ visual words. To account for different statistics in the two image datasets, two visual word dictionaries are learned separately.

5.1. Caption Perplexity

To measure the quality of annotations predicted by the models, we follow [2] and adopt caption perplexity as a performance measure. The essential quantity that we need to compute is the conditional probability of caption words given a test image $p(w|\mathbf{R})$, which is computed with respect to the variational factorized posterior. As seen in the definition in (21), perplexity is indeed the inverse of the geometric mean likelihood, which implies that the model that gives higher conditional likelihood will lead to a lower perplexity (hence the lower the number, the better the model).

$$\text{Perp} = \exp \left(-\frac{\sum_{d=1}^D \sum_{m=1}^{M_d} \log p(w_m|\mathbf{R}_d)}{\sum_d M_d} \right) \quad (21)$$

As seen in Fig 2(a), we show caption perplexity for the COREL dataset as a function of the number of text topics L when K is fixed at 50 and 100. Generally, perplexity decreases as L increases and we do not have observe problems with over-fitting. Showing as a baseline in the magenta graph is the average perplexity of around 135 words when the regression coefficient matrix \mathbf{A} is set to 0, *i.e.* the 2 data modalities are independent. When \mathbf{A} is properly learned from the data, perplexity is reduced on average by over 50 words. A similar pattern is again observed in the plot of perplexity as a function of K while holding L fixed in Fig 2(b). tr-mmLDA appears quite robust to over-fitting as model complexity increases.

We compare the predictive performance of cLDA and tr-mmLDA. To see if over-fitting is an issue for cLDA, we observe how perplexity (as a function of K) changes as the number of patches in each image N changes. To obtain images with varying values of N , we sub-sample N patches from all the patches to represent each image. As shown in Fig 3(a) for a small value of $N = 20$, cLDA

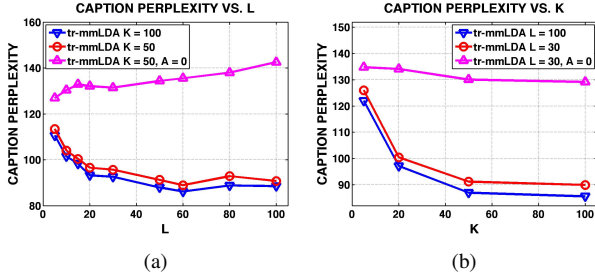


Figure 2. COREL dataset: (a) Caption perplexity as a function of L . (b) Caption perplexity as a function of K .

seriously overfits the data as perplexity increases as K increases. With larger N , overfitting becomes less of an issue but tr-mmLDA still outperforms cLDA.

Indeed, the severe overfitting in cLDA can be directly attributed to the restrictive association between the 2 modalities. When N is small, since cLDA enforces that the topics used in the caption modality must be a subset of the topics occurring in the image modality, a small value of N implies that the words in each document will be assigned to only a small number of topics (clusters). For a large value of K , in order not to have empty clusters (topics), a large number of documents will be required. With less than 2200 training documents in the LabelMe dataset, the topic parameters for caption texts will therefore be estimated poorly. For the COREL dataset with a larger training set (4500 documents), the problem of over-fitting does not become severe until we reduce N down to 10, as shown in Fig 3(b). Since tr-mmLDA imposes no such restriction with regards to the number of topics used for the caption modality, over-fitting becomes less of an issue using the same dataset size.

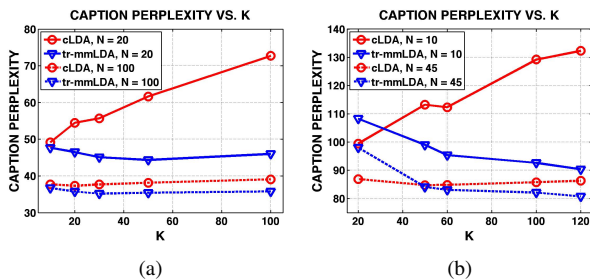


Figure 3. Perplexity as a function of K as N increases for (a) LabelMe (b) COREL dataset.

5.2. Example Annotation and Topics

We compare examples of caption topics from the LabelMe dataset learned using cLDA and tr-mmLDA. Like other previous topic models, we examine the Multinomial parameters and employ 10 most probable caption words under each topic to represent the topic. We learn 50 topics using cLDA and found around 50% of those topics learned has the word *car* in its top 10 caption words, while only 4 out of 50 topics learned using tr-mmLDA contains the word

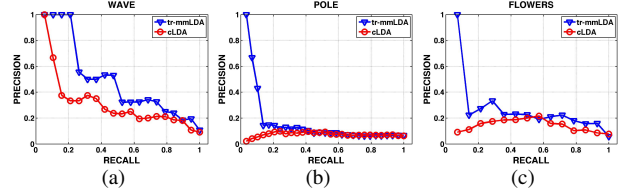


Figure 4. Precision-recall curve for 3 single word queries: ‘wave’, ‘pole’, ‘flower’, comparing cLDA (red) and tr-mmLDA (blue).

car, see Figure 5. The *car* example used here illustrates that topics learned using cLDA are found to contain more general terms, while tr-mmLDA uncover the 4 contexts *car* appears in the dataset: class highway, street, inside city, tall building. Fig 5 shows examples of more superior quality of annotation inferred by our model (compared to cLDA) because of the superior topic parameters learned.

Table 1. Example topics learned using cLDA (top panel) and tr-mmLDA (bottom panel)

Topic 1	<i>car</i> , sky, road, tree, building, mountain, trees
Topic 2	building, window, <i>car</i> , person, buildings, skyscraper
Topic 5	sky, road, <i>car</i> , fence, mountain, trees, sign
Topic 6	<i>car</i> , building, buildings, person, sidewalk, cars, walking
Topic 14	window, building, <i>car</i> , door, person, pane, road
Topic 15	sky, tree, building, mountain, <i>car</i> , road, trees
Topic 16	buildings, <i>car</i> , sky, tree, cars, building, road
Topic 17	<i>car</i> , road, sign, trees, street light, highway
Topic 20	<i>car</i> , road, highway, freeway, sign, trees, streetlight
Topic 26	building, <i>car</i> , buildings, sky, mountain, sidewalk, road
Topic 27	<i>car</i> , building, buildings, person, sidewalk, road, tree
Topic 15	<i>car</i> , road, sign, trees, highway, freeway, sky
Topic 32	<i>car</i> , buildings, building, sidewalk, cars, road, sky
Topic 39	<i>car</i> , road, <i>car</i> back, <i>car</i> top back, van, <i>car</i> right, car left
Topic 48	balcony, shop, building, door, <i>car</i> , terrace, light

5.3. Text-based Retrieval

We can also use the annotation model of tr-mmLDA and cLDA to perform image retrieval on a database of un-annotated images using word queries. This retrieval method is called text-based retrieval, to contrast with content-based retrieval where queries are given in the form of examples. Given a single word query, we perform retrieval by ranking the test images according to the probability that each image will be annotated with the query word. More specifically, the score used in ranking is $p(w|\mathbf{R}_{test})$ which can be computed using variational posterior inferred for each test document. Table 2 shows top 4 images from the LabelMe dataset that have been retrieved, rank ordered by the score $p(w|\mathbf{R})$, using tr-mmLDA with query words ‘hill’ and ‘buildings’. Figure 4 shows precision-recall curves for 3 single word queries, comparing the rank order of the retrieved images using cLDA and tr-mmLDA. Our model generally yields higher precisions at the same recall values for all the 3 queries and give a better overall retrieval performance.

6. Conclusion

In this work, we propose topic-regression multi-modal LDA, a novel statistical topic model for the task of image



Figure 5. Examples of predicted annotation on LabelMe dataset.

annotation. The main novelty of our model is the latent variable regression approach to capture correlations between image features and annotation texts. Instead of sharing the hidden topics directly between the 2 data modalities as in cLDA, we propose a formulation that keeps 2 sets of hidden topics and incorporate a linear regression module to correlate them. Our approach can capture varying degrees of correlations and allow the number of topics in image and annotation text to be different. Experimental results on image annotation show that the association model of tr-mmLDA has an edge over correspondence LDA in predicting annotation as seen in the superior annotation and retrieval quality.

References

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

Table 2. Top 4 images (with no captions) retrieved using single word queries. The queries *hill* and *buildings* are used for images in the top row and bottom row accordingly.



[2] D. M. Blei and M. I. Jordan. Modeling annotated data. In *ACM SIGIR*, 2003.

[3] D. M. Blei and J. D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

[4] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *Neural Information Processing Systems (NIPS)*, 2007.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[7] L. Fei-fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[8] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[9] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

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

[11] D. Putthividhya, H. Attias, and S. Nagarajan. Independent factor topic models. In *International Conference on Machine Learning (ICML)*, 2009.

[12] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *International Conference on Computer Vision (ICCV)*, 2005.

[13] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *ECCV*, 2006.

[14] C. Wang, D. M. Blei, and L. Fei-fei. Simultaneous image classification and annotation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.