Learning Cross-modality Similarity for Multinomial Data

Yangqing Jia
UC Berkeley EECS
jiayq@eecs.berkeley.edu

Mathieu Salzmann
TTI-Chicago
salzmann@ttic.edu

Trevor Darrell
UC Berkeley EECS
trevor@eecs.berkeley.edu

Abstract

Many applications involve multiple-modalities such as text and images that describe the problem of interest. In order to leverage the information present in all the modalities, one must model the relationships between them. While some techniques have been proposed to tackle this problem, they either are restricted to words describing visual objects only, or require full correspondences between the different modalities. As a consequence, they are unable to tackle more realistic scenarios where a narrative text is only loosely related to an image, and where only a few image-text pairs are available. In this paper, we propose a model that addresses both these challenges. Our model can be seen as a Markov random field of topic models, which connects the documents based on their similarity. As a consequence, the topics learned with our model are shared across connected documents, thus encoding the relations between different modalities. We demonstrate the effectiveness of our model for image retrieval from a loosely related text.

1. Introduction

Many real-world applications involve multi-modal data, where information arises from different sources, such as images, text, or speech. In this paper, we focus on images with loosely related narrative text descriptions, which are a natural way of providing rich information about the image, not restricted to exploiting words associated to visible objects. Figure 1 gives an example of this: Objects irrelevant to the description of the image, such as *sky* and *cranes*, are not present in the text, while non-visual words, such as *launch*, *maiden flight* and *accomplish*, strongly help understanding the image at a high level. Even though existing techniques have tackled the problem of leveraging text associated with images, they typically assume the text to contain mostly words describing visible objects. As a consequence, they are not able to exploit the entire information present in a narrative text.

Combining multiple sources of information can be traced back to multiple-kernel learning [9]. Recently, fusing text and image information has received much attention. Several approaches [1, 3, 22, 20, 2] have proposed general probabilistic models to tackle the multi-modal scenario for tasks such as object detection, recognition, and scene understanding. However, these approaches are restricted to using only words matching visual objects in the images. These words typically correspond to category labels, or tags, and richer information in the text is discarded.

"A timed exposure of the first Space Shuttle mission, STS-1, at Launch Pad A, Complex 39, turns the space vehicle and support facilities into a night-time fantasy of light. To the left of the Shuttle are the fixed and the rotating service structures."

Figure 1. Example of an image with a loosely related, narrative description from wikipedia.

In the text processing community, topic models, such as Latent Dirichlet Allocation (LDA) [5], have proved effective at discovering the underlying topics in text documents, and thus at modeling more than single words. To this end, they learn the groups of semantically consistent words that generate the training data. Topic models were extended to the image domain by replacing text words with local image descriptors [19, 18]. The resulting models have been successfully applied to problems such as scene classification and content-based image retrieval. Modeling spatial interactions across topics in an LDA model has recently been addressed for image segmentation [23] by defining a spatial graph over the topic activations of local image patches.

While LDA is effective in these single modality scenarios, it does not directly apply to the multi-modal case. In particular, LDA does not provide a mechanism to model the relationships between topics coming from different modalities. To address this issue, other models have been developed. For instance, Correspondence LDA (Corr-LDA) [3] was proposed to capture the topic-level relations between images and text annotations. Corr-LDA assumes a one-to-one correspondence between the topics of each modality. In other words, each image topic must have a corresponding text topic. To generalize over this, a topic regression multi-

2011 IEEE International Conference on Computer Vision
978-1-4577-1102-2/11/$26.00 ©2011 IEEE

modal LDA was recently proposed [13]. This model learns a regression from the topics in one modality to those in the other. As a result, it does not have a one-to-one correspondence between each individual topic, but between the sets of topics describing each modality in a document. Unfortunately, this still assumes that each image is associated with a text description. Furthermore, in practice, these types of models have only been applied to the case where all the words in the description have a visual interpretation. In more realistic scenarios where images and text are loosely related, these models would therefore neglect most of the text information.

In this paper, we introduce a model that addresses the two above-mentioned issues. In particular, our model is able to leverage the information of non-visual words in a text loosely related to an image. Furthermore, we do not require to be given pairs of corresponding image and text, but only employ the notion of similarities between two documents containing a single modality. As a consequence, our model can exploit the availability of only a few image-text pairs, together with image-image and text-text similarities, to learn the intrinsic relationships between images and text.

More specifically, our model can be seen as a Markov random field (MRF) over LDA topic models. Each node of the MRF represents the topic model for a particular document, which contains a single modality. The edges in the graph encode the similarity between two documents containing either the same modality, or different ones. Each document is then generated not only from its own topics, but also from the topics of the documents connected to it. Learning our model therefore yields topics that are shared across several documents. As a consequence, when two linked documents contain different modalities, our model learns the relations between these modalities. We name our model Multi-modal Document Random Field (MDRF).

We demonstrate the benefit of our approach over existing multi-modal LDA models on the task of retrieving images from loosely related text descriptions.

2. Modeling Multi-modal Data

In this section, we first briefly review LDA and its multi-modal extensions, and then explain the generative process of our model.

2.1. LDA and Corr-LDA Revisited

Latent Dirichlet Allocation [5] is a generative probabilistic model for collections of discrete data. In general, LDA aims to discover the topics that generate the documents in a corpus, as well as the topic proportion for each document. More specifically, following the notation in Figure 2, the topic proportion $\boldsymbol{\theta}_d$ for a particular document d follows a Dirichlet distribution with parameter α. Given $\boldsymbol{\theta}_d$, a particular topic z_{dn} is drawn from a multinomial distribution,

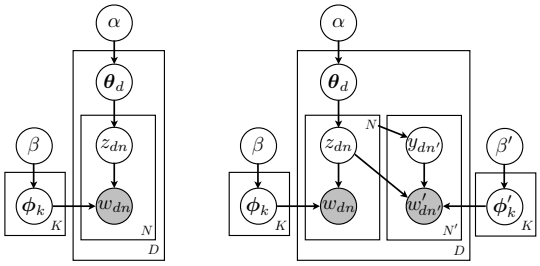


Figure 2. Graphical models of LDA (left) and Corr-LDA (right).

and in turn, a word w_{dn} from the corresponding topic-word multinomial distribution $\boldsymbol{\phi}_k$, which is drawn from a Dirichlet distribution with prior β. This defines the marginal probability for a document as

$$p(\mathbf{w}_d|\alpha,\beta) = \int p(\boldsymbol{\theta}_d|\alpha) \times \left(\prod_{n=1}^{N}\sum_{z_{dn}} p(z_{dn}|\boldsymbol{\theta}_d)p(w_{dn}|z_{dn},\beta)\right) \mathrm{d}\boldsymbol{\theta}_d \,, \quad (1)$$

The probability distribution for the whole document corpus is taken as the product of the probability of each document.

Correspondence LDA [3] was introduced to account for the availability of multiple modalities in the LDA framework. In particular, it tackles the problem of modeling annotated images. The image part is modeled using standard LDA. To generate the text, a region indicator $y_{dn'}$ is drawn from a uniform distribution over $\{1,\cdots,N\}$, and used in conjunction with the image topic assignment z_{dn} to draw the text words w'_{dn} from a multinomial distribution with dirichlet prior β'. From this, it can be seen that Corr-LDA treats the two modalities differently: The text topics are sampled from the empirical distribution of the image topics. Thus if a topic is not discovered from the images, this topic won't be available to generate the text. As mentioned before, this limits the applicability of Corr-LDA in scenarios where the text is more loosely related to the images.

To generalize over this requirement for one-to-one topic correspondence, the topic regression multi-modal LDA model was recently proposed [13]. In essence, this model learns a linear mapping between the topics proportions for one modality and the topics proportions for the other. As in Corr-LDA, the text modality can then be generated from the topic proportions computed for the image modality. However, the dependencies between the topics is weaker than in the Corr-LDA case. Instead of relying on a multinomial distribution to generate the topics, the model uses a logistic normal distribution, as the correlated topic model [4]. This generalizes over the Corr-LDA model, but has the drawback of making inference more complicated, since there are $O(K^2)$ additional parameters to learn. More importantly, this still assumes that image-text pairs are available for all

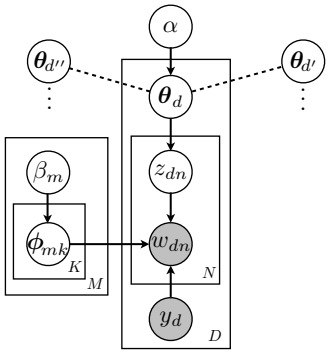


Figure 3. The Graphical model of the Multi-modal Document Random Field model. The dashed edges denote the similarities between different documents.

the documents. As described below, our model addresses this issue by considering the notion of similarities between two documents, thus having weaker requirements for the documents available at training.

2.2. Multi-modal Document Random Field Model

Our paper focuses on learning a generative topic model from a set of documents $\mathcal{D} = \{(y_1, \mathbf{w}_1), (y_2, \mathbf{w}_2), \cdots, (y_D, \mathbf{w}_D)\}$. Each document $(y_d, \mathbf{w}_d)$ contains an index $y_d \in \{1, 2, \cdots, M\}$ selecting one modality among M possible ones, and a set $\mathbf{w}_d$ of words drawn from the vocabulary of this particular modality. Without loss of generality, we assume that each word w_{dn} $(1 \leq n \leq N_d)$ takes a discrete value in $\{1, 2, \cdots, V_m\}$, where V_m is the vocabulary size of the m-th modality. Note that as opposed to Corr-LDA and other existing multi-modal topic models, we do not assume a full set of corresponding documents across the different modalities. In other words, we do not assume that there exists a corresponding text document for each image document. Instead, we assume that we are given a document-level similarity graph $\mathcal{G} = (\mathcal{D}, \mathcal{E})$, where $\mathcal{E}$ is a set of edges modeling the similarity between different documents. If there is an edge $e = (i, j)$ between document i and document j, the two documents are considered similar. Note that this is a weaker requirement than one-to-one correspondences, since the graph might not contain all image-text pairs, and allows for more general similarities, such as image-image ones. As we show below, this serves as a weakly-supervised information to help us discover the topics shared across documents and modalities.

Figure 3 depicts the graphical model of our approach, where α and $\beta_{1\ldots M}$ are the hyperparameters for the Dirichlet priors. In this graphical model, each document is represented with an LDA model. In addition to this, we model the relationships between pairs of documents with the similarity graph $\mathcal{G}$. This graph defines a Markov random field over the documents. For each edge $e = (i, j)$ in the graph, we define the potential function

$$\psi(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \exp\left(-\lambda f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\right), \tag{2}$$

where $f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ is a distance measure between two documents, and λ is the parameter that controls the peakyness of the potential function, which can be interpreted as the strength of the similarity. Several distance measures can be employed, the simplest of which is the Euclidean distance. Here, we choose the symmetric KL-divergence defined as

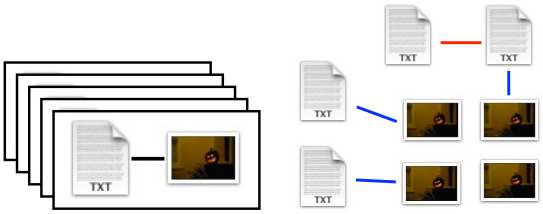
$$\begin{aligned} f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) &= \frac{1}{2}(D_{KL}(\boldsymbol{\theta}_i || \boldsymbol{\theta}_j) + D_{KL}(\boldsymbol{\theta}_j || \boldsymbol{\theta}_i)) & (3) \\ &= \frac{1}{2}\sum_{k=1}^{K}\left(\theta_{ik}\log\frac{\theta_{ik}}{\theta_{jk}} + \theta_{jk}\log\frac{\theta_{jk}}{\theta_{ik}}\right). & (4) \end{aligned}$$

From a generative perspective, each document d in modeled by first generating a topic distribution $\boldsymbol{\theta}_d$, and then sampling the words of that document given $\boldsymbol{\theta}_d$. Similarly as in LDA, we generate $\boldsymbol{\theta}_d$, from a Dirichlet prior. However, in addition to this prior, the topic distribution also depends on the random field. More specifically, given the hyperparameters, the number of topics K, the graph $\mathcal{G}$, and the vocabulary size V_m for each modality, the generative procedure goes through the following steps:

1. For each topic k in each modality m, sample the V_m dimensional word distribution $\boldsymbol{\phi}_{mk} \sim \text{Dir}(\boldsymbol{\phi}|\beta_m)$.
2. Sample the D topic proportions $\boldsymbol{\theta}_{1\ldots D}$ from the distribution
$$p(\boldsymbol{\theta}_{1\ldots D}|\alpha, \mathcal{G}) = \frac{1}{Z}\exp(-\lambda\sum_{i,j\in\mathcal{E}} f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j))\prod_{d=1}^{D}\text{Dir}(\boldsymbol{\theta}_d|\alpha),$$
where Z is a normalization constant.
3. For each document d, sample its modality y_d from a uniform distribution over $\{1, \cdots, M\}$.
4. For each word w_{dn}:
   (a) Sample a topic $z_{dn} \sim \text{Multi}(z|\boldsymbol{\theta}_d)$;
   (b) Sample a word $w_{dn} \sim \text{Multi}(w|\boldsymbol{\phi}_{y_d z_{dn}})$.

From this procedure, and by defining $\boldsymbol{\Phi}$ as the set of word-distribution parameters, the joint probability of a document corpus given similarities between the documents can be written as

$$\begin{aligned} &p(\mathcal{D}, \boldsymbol{\theta}_{1\ldots D}, \mathbf{z}_{1\ldots D}, \boldsymbol{\Phi}|\alpha, \beta_{1\ldots M}, \mathcal{G}) \\ &= \frac{1}{Z}\prod_{m=1}^{M}\prod_{k=1}^{K}\text{Dir}(\boldsymbol{\phi}_{mk}|\beta_k)\exp\left(-\lambda\sum_{i,j\in\mathcal{E}} f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\right) \quad (5) \\ &\times\prod_{d=1}^{D}\text{Dir}(\boldsymbol{\theta}_d|\alpha)\left(\prod_{n=1}^{N_d}\text{Multi}(z_{dn}|\boldsymbol{\theta}_d)\text{Multi}(w_{dn}|\boldsymbol{\phi}_{y_d z_{dn}})\right). \end{aligned}$$



(a) Previous Models (b) Our Model

Figure 4. Comparison of existing multi-modal LDA models with our model. While previous models define documents in a "superdocument" fashion, our model assumes a single modality per document.

2.3. Relation to Existing Models

In general, our model is a natural extension of LDA to the multi-modal case. The key contribution of our model is the document random field, which enables us to capture the similarities between documents from different modalities. Note that our definition of a document is different from existing multi-modal LDA models, who define a document to be a *super-document* that contains one sub-document for each modality. As depicted in Fig. 4, defining documents to be single-modal enables us to utilize those without cross-modality correspondences, or supervised intra-document similarities. We will show in the experiments that such flexibility is particularly helpful when correspondence information is scarce.

The idea of fusing the Markov Random Field and LDA has been shown in [23]. However, in this approach, a random field is built within each document on the topic level, in order to capture the spatial relationships between topic assignments. Our model builds the random field on the document level instead, and tackles the problem of multi-modal data and document similarities.

From a different perspective, our model can be seen as learning a joint latent space for documents containing different modalities. The similarities between documents are enforced in the joint latent space in a weakly supervised manner. Learning shared latent spaces across modalities has been an active research topic in human pose estimation [17, 6, 16] and image domain transfers [15]. However, the existing methods focus on dense, real-valued feature spaces and are typically designed for Gaussian distributions. Our work, on the other hand, explores the possibility of finding shared information in the context of integer-valued multinomial distributions.

In the single-modality case, several methods such as the Hierarchical Dirichlet Process [21] and Pachinko Allocation [10] have shown that a deeper topic structure may better capture the underlying semantics of the corpus. The potential discrepancy between image topics and text topics, as raised by [13], can be tackled by assuming topic correspondence at a deeper level. While our model uses LDA as the generative procedure of the data, a deep topic model can be naturally employed. This will be the topic of future work.

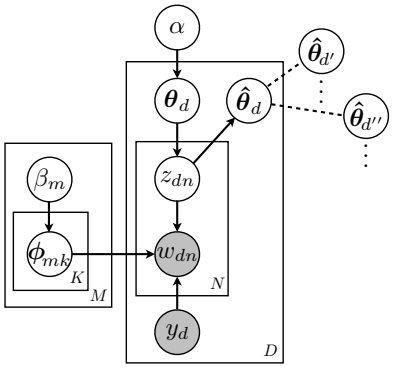


Figure 5. Empirical-MDRF for efficient inference.

3. Learning the Model

In this section we describe our learning strategy for the MDRF model. The hidden variables of the model are the multinomial distribution parameters $\boldsymbol{\Phi}$ and the topic assignments for all the documents. We assume a symmetric Dirichlet prior for the topic distribution and word distribution, and take $\beta_{1\ldots M}$ to be identical for all the modalities. Similarly as in LDA, exact inference is in general intractable. We therefore need to resolve to one of the usual approximate inference methods, such as variational inference [5], expectation propagation [11], or Gibbs sampling [7]. Here, we use Gibbs sampling, since it has proved effective at avoiding local optima, while yielding relatively simple algorithms.

3.1. Empirical-MDRF for Efficient Inference

The general MDRF model is able to capture the document similarities via the random field. However, inference with this random field is generally difficult as the topic distributions for multiple documents are coupled. Inspired by Corr-LDA, instead of enforcing similarity on $\boldsymbol{\theta}_d$s, we introduce an empirical topic distribution $\hat{\boldsymbol{\theta}}_d$ for each document d, and construct the graph on these distributions. This yields the generative model depicted in Figure 5. We call this model the Empirical-MDRF and will use it for all the experiments in this paper.

Specifically, given a set of topic assignments $\mathbf{z}_d$ in document d, the empirical topic distribution $\hat{\boldsymbol{\theta}}_d$ is computed as

$$\hat{\theta}_{dk} = \frac{n_{dk}^{(d)} + \alpha}{\sum_{k=1}^{K} n_{dk}^{(d)} + K\alpha}, \tag{6}$$

where $n_{dk}^{(d)}$ is the number of occurrences of topic k in document d. Note that we introduced a smoothness factor in the computation of $\hat{\boldsymbol{\theta}}_d$. This leads to a more robust estimation in

practice, when we need to compare the similarity between two documents. In fact, $\hat{\boldsymbol{\theta}}_d$ is the maximum likelihood estimate of the underlying multinomial distribution given the observation $\mathbf{z}_d$ sampled from the Dirichlet-multinomial distribution

$$p(\mathbf{z}|\alpha) = \int_{\boldsymbol{\theta}} \text{Multi}(\mathbf{z}|\boldsymbol{\theta})\text{Dir}(\boldsymbol{\theta}|\alpha)\text{d}\boldsymbol{\theta} \ . \tag{7}$$

The joint distribution of this empirical model is similar to that of the original MDRF model. However, as we will show in the next subsection, inference in the empirical MDRF model can be performed via an efficient collapsed Gibbs sampling algorithm.

3.2. Gibbs Sampling

For an excellent discussion about Gibbs sampling for LDA-like probabilistic models, we refer the reader to [8]. In this paper, we employ a collapsed Gibbs sampling algorithm. To this end, we marginalize out $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$, and only perform Gibbs sampling on the $\mathbf{z}$s. More specifically, we sample a topic assignment for one word based on its conditional probability given the observations and the topic assignments for the other words, and by integrating out the multinomial distributions with parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$. For document d containing modality $y_d = m$, the probability of the topic assignment of word w being k given the corpus $\mathcal{D}$, the parameters α and β, and the topic assignments for the other words $\mathbf{z}_{-w}$ is expressed as

$$\begin{aligned} &P(z = k|\mathcal{D}, \mathbf{z}_{-w}, \alpha, \beta) \propto \\ &\frac{n_{dk}^{(d)} + \alpha}{\sum_{k=1}^{K} n_{dk}^{(d)} + K\alpha} \times \frac{n_{kw}^{(m)} + \beta_y}{\sum_{w=1}^{V_m} n_{kw}^{(m)} + V_m\beta_m} \\ &\times \prod_{d',(d,d')\in\mathcal{E}} \exp\left(\lambda f(\hat{\boldsymbol{\theta}}_{d,-z}, \hat{\boldsymbol{\theta}}_{d'}) - \lambda f(\hat{\boldsymbol{\theta}}_{d,z=k}, \hat{\boldsymbol{\theta}}_{d'})\right) \ , \end{aligned} \tag{8}$$

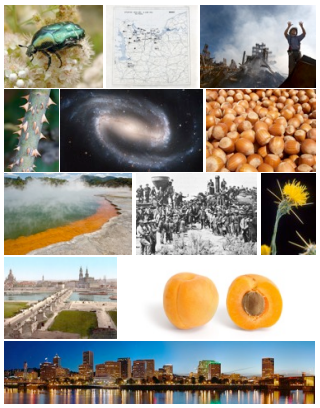
where $n_{kw}^{(m)}$ is the number of occurrences of word w in topic k for modality m, both excluding the current word. $\hat{\boldsymbol{\theta}}_{d,-z}$ is the empirical topic distribution for document d excluding the current word, and $\hat{\boldsymbol{\theta}}_{d,z=k}$ is the empirical topic distribution for document d when the topic for the current word is k. The first two terms in this equation are identical to those in LDA, and the last term encodes the conditional probability introduced by the random field.

3.3. Parameter Estimation

For all the topic models, determining the hyperparameters of the Dirichlet distributions is an important issue. While empirically optimal parameter settings are available for LDA [7] when applied to text processing, such parameter settings might not be optimal for other modalities such as images. Finding the optimal parameters for our method by performing a grid-search is also prohibitive. Therefore, we seek to automatically learn these parameters from the training data. This has been shown to be possible when the latent topic assignments are fixed (i.e., in a slice during the Gibbs sampling procedure) [12]. For instance, for fixed latent variables, the hyperparameter α is obtained by iteratively carrying out the update rule

$$\alpha \leftarrow \frac{\alpha\left[\left(\sum_{d=1}^{D}\sum_{k=1}^{K}\Psi(n_{dk}^{(d)} + \alpha)\right) - DK\Psi(\alpha)\right]}{K\left[\left(\sum_{d=1}^{D}\Psi(\sum_{k=1}^{K}(n_{dk}^{(d)} + \alpha))\right) - D\Psi(K\alpha)\right]} \ , \tag{9}$$

where $\Psi(\cdot)$ is the digamma function $\Psi(x) = \frac{d}{dx}\ln\Gamma(x)$. The other parameters, $\beta_{1\ldots M}$, are updated in a similar fashion. In practice, hyperparameter update is performed every few Gibbs sampling steps. To prevent over-fitting to the current slice, we only run a limited number of iterations for each update (1 in our experiments).



"world, species, united, states, found, north, american, image, convert, common, large, long, located, city, war, native, small, family, century, largest, national, water, time, light, river, plant, popular, designed"

Figure 6. Representative subset of the images in the POTD dataset, and of the words that appear most frequently in the text corpus.

4. Experiments

In this section, we empirically show the effectiveness of our model on the task of multi-modal image retrieval. Since most existing multi-modal datasets are limited to annotations that describe visible object names only, we collected a new dataset containing richer and looser text descriptions of the images. We first describe this dataset and the experimental settings, and then present our results and compare them against those obtained with LDA and Corr-LDA.

4.1. The Wikipedia POTD Dataset

The wikipedia "Picture of the day" website[1] provides a collection of daily featured pictures. Together with the images, a short paragraph of about 100 words gives a brief and

[1] http://en.wikipedia.org/wiki/Wikipedia:Picture_of_the_day

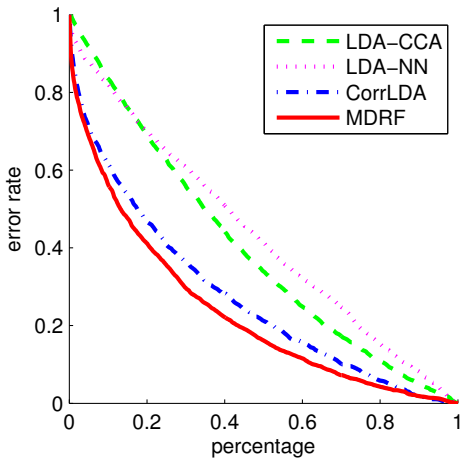


Figure 7. Average error rate as a function of the percentage of the ranked list considered for retrieval. Curves closer to the axes represents better performance. See the text for more details.

Method	AUC value
LDA-NN	43.15 ± 1.95
LDA-CCA	39.44 ± 2.27
Corr-LDA	26.94 ± 1.87
MDRF	$\mathbf{23.14 \pm 1.49}$

Table 1. Average area under the curve (AUC) (in percentage) and standard deviations for the curves in Figure 7. A smaller value indicates a better performance.

loose description of the picture. Figure 6 shows several representative images and words from the dataset. Note that both the pictures and the descriptions cover a wide variety of topics ranging from celestial pictures to historical photos. Furthermore, the words are beyond the scope of simple visual objects present in the images.

For our experiment, we collected the daily pictures and their corresponding descriptions from Nov 1, 2004 to Oct 30, 2010. After removing non-image data (e.g., movie files) and text that could not be parsed, we obtained a total of 1987 image-text pairs. We used rainbow[2] to tokenize the text and kept the words that appeared more than 3 times in the whole corpus. This resulted in a vocabulary of 3562 words. For the images, we computed densely sampled SIFT features over 16×16 grids. Each image was resized so that approximately 400 features were sampled per image. We randomly chose a subset of 50,000 SIFT features and ran k-means to obtain 1,000 clusters. These clusters were used to vector-quantize the SIFT features, thus yielding 1,000 discrete visual words. The dataset can be downloaded at http://www.eecs.berkeley.edu/∼jiayq/wikipedia_potd/.

[2]http://www.cs.cmu.edu/∼mccallum/bow/rainbow/

Method	Percentage
LDA-NN	30.10
LDA-CCA	30.98
Corr-LDA	53.30
MDRF	**58.84**

Table 2. Percentage of images correctly retrieved in the first 20% of the ranked list.

4.2. Retrieval Protocol

To test our model and to compare it against existing methods, we consider the problem of multi-modal image retrieval. More specifically, given a text query, we aim to find images that are most relevant to it. For each text in the test set, we rank the test images using either our approach, or a competing method. To this end, for Corr-LDA and for our method, we learn the topic distributions $\boldsymbol{\theta}_i$ for each test image. Given a text query $\mathbf{w} = w_1, w_2, \cdots, w_N$, the score for each image is then defined as

$$s_i = p(\mathbf{w}|\boldsymbol{\theta}_i) = \prod_{n=1}^{N} p(w_n|\boldsymbol{\theta}_i) \; . \tag{10}$$

Note that the marginal probabilities $p(w_n|\boldsymbol{\theta}_i)$ for all words can be pre-computed for each image during learning time, so no marginalization is necessary during query time. An alternative to this would be to compute the text-topic distribution and measure the KL-divergence between this distribution and the image-topic distribution. However, this requires an inference step for each query, which is time-consuming. Instead, the score described above is deterministic and can be performed in $O(N)$ time.

Since there is only one ground-truth match for each image/text, to evaluate the performance we rely on the position of the ground-truth image in the ranked list obtained. More specifically, an image is considered correctly retrieved if it appears in the first t percent of the list created from its corresponding text. Sweeping through all the text queries gives us an error rate that is dependent on t, which is shown in Figure 7.

To obtain statistically valid error measures, we split the data into 10 folds, and test on each fold with the remaining 9 as training data. For LDA and Corr-LDA, all the hyper-parameters can be learned directly from the training data as described in Section 3.3. Our method uses an additional parameter λ for the document random field. To set this parameter, we performed a grid-search using cross validation on the first 9 folds. The optimal value for λ was kept unchanged for all the other partitions. For all the methods, we fixed the number of topics to 64. This number was found to work best for LDA and Corr-LDA, while our method was not significantly affected by the number of topics. We set the burn-in period for Gibbs sampling to 1,000 iterations.

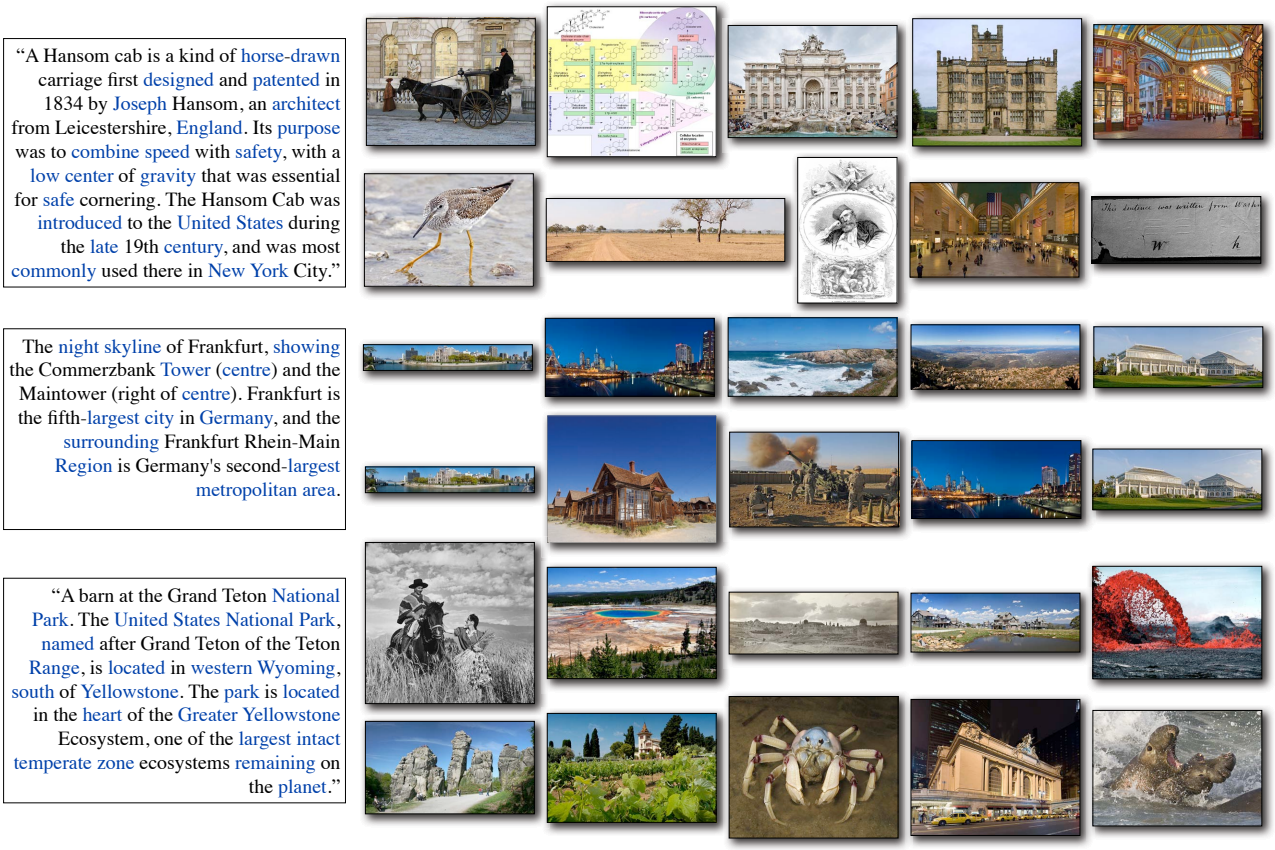


Figure 8. Three typical image retrieval results. For each example, we show the query text, the top 5 images returned by our algorithm (top row), and the top 5 images returned by Corr-LDA (bottom row). The words that are in the vocabulary are colored in blue. For space consideration, the results of the LDA baselines are not shown here.

We compare our method against Corr-LDA and two LDA-based baselines[3]. In the two latter cases, LDA models are trained separately for images and text. Retrieval is then performed using either nearest-neighbors (LDA-NN), or CCA (LDA-CCA) [14]. For LDA-NN, we compute the nearest neighbor of the query text among the training texts, take the corresponding training image, and build the ranked list of test images using the symmetric KL-divergence between the image topic distributions. LDA-CCA learns the individual projections of the image and text topic distributions to a joint latent space in which the correlation between those distributions is maximum. The ranked list is then obtained from the distances between the test images and the query text in this latent space. For each experiment, we searched for the dimensionality of the CCA latent space that gave the best results.

[3]No reference implementation of the topic regression MMLDA [13] is available. We implemented a Gibbs sampling version of the algorithm, which performed worse than Corr-LDA. Since our implementation might be different from the original one that uses variational inference, we do not report its performance here. A potential explanation is that topic regression MMLDA has a large number of parameters to learn, making it less robust on small training sets such as ours.

4.3. Results

We now present our results on the POTD dataset. Figure 7 depicts the retrieval errors averaged over the 10 partitions for all the methods. In Table 1, we report the area under the curve (AUC) values for those errors. A t-test with threshold 0.01 revealed that the difference between our results and the others is significant. Since in information retrieval, it is always valuable to have related documents appear as early as possible in the ranked list, we also report the percentage of the images correctly retrieved in the first 20% of the ranked list in Table 2. Compared to Corr-LDA, about 5% more documents on average are accurately retrieved by our method.

Figure 8 shows several illustrative examples of the retrieval results, using text from the POTD pages. Qualitatively, it can be observed that our model captures the general topics represented in both the images and the text better than Corr-LDA. For instance, in the third query, our model captures the fact that the national parks mentioned in the text are closely related to nature and outdoor scenes. In the first

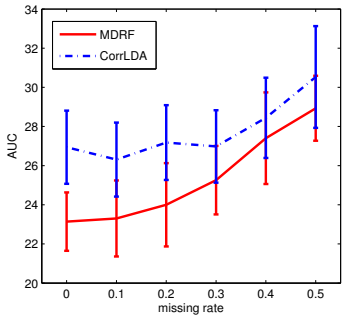


Figure 9. Average AUC value as a function of the percentage of missing correspondences.

query, our model relates the city names in the text to urban images, whereas Corr-LDA cannot capture this connection, since city names do not correspond to visible objects in an image.

Finally, to test the robustness of our algorithm against missing correspondence information, we removed a subset of the correspondences between images and text when learning the models. CorrLDA is not able to use the part of data that do not have correspondence information present, while our method can process sparse similarity information inherently. More specifically, we assume that t percent of the correspondence in the training corpus are unknown, and vary t from 0 to 50 in our experiments. The average AUC value versus the proportion of missing correspondences is shown in Figure 9. It can be observed that our method consistently outperforms CorrLDA. Furthermore, note that in the limit where no correspondences are available, Corr-LDA could not be applied at all. In contrast, our model would still learn topics that generate the documents well, although they would not necessarily model the cross-similarities.

5. Conclusion

In this paper, we have proposed a new probabilistic model that learns cross-modality similarities from a document corpus containing multinomial data. While existing methods require full correspondence between the modalities, our MDRF model defines a Markov random field on the document level that allows modeling more flexible document similarities. As a result, our model learns a set of shared topics across the modalities. By applying our model to the task of image retrieval from wikipedia data, where the narrative text is only loosely related to the images, we have shown that our method outperforms existing techniques, which assume the text to contain visual objects only. In the future, we intend to study the use of deeper topic structures, such as Pachinko Allocation [10], to better capture the semantics shared among the documents.

References

[1] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, and M. Jordan. Matching words and pictures. *JMLR*, 3:1107–1135, 2003. 1

[2] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *CVPR*, 2004. 1

[3] D. Blei and M. Jordan. Modeling annotated data. In *SIGIR*, 2003. 1, 2

[4] D. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35, 2007. 2

[5] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003. 1, 2, 4

[6] C. Ek. Shared Gaussian Process Latent Variable Models. *Ph.D. Thesis*, 2009. 4

[7] T. Griffiths. Finding scientific topics. *Proceedings of the National Academy of Science*, 101:5228–5235, 2004. 4, 5

[8] G. Heinrich. Parameter estimation for text analysis. *Technical Report*, 2005. 5

[9] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan. Learning the kernel matrix with semidefinite programming. *JMLR*, 5:27–72, 2004. 1

[10] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML*, 2006. 4, 8

[11] T. Minka. Expectation propagation for approximate Bayesian inference. In *UAI*, 2001. 4

[12] T. Minka. Estimating a Dirichlet distribution. *Technical report, MIT, 2003*, 2003. 5

[13] D. Putthividhy, H. Attias, and S. Nagarajan. Topic regression multi-modal Latent Dirichlet Allocation for image annotation. In *CVPR*, 2010. 2, 4, 7

[14] N. Rasiwasia, J. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A New Approach to Cross-Modal Multimedia Retrieval. In *ACM MM*, 2010. 7

[15] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 4

[16] M. Salzmann, C. Ek, R. Urtasun, and T. Darrell. Factorized orthogonal latent spaces. In *AISTATS*, 2010. 4

[17] A. Shon, K. Grochow, A. Hertzmann, and R. Rao. Learning shared latent structure for image synthesis and robotic imitation. In *NIPS*, 2005. 4

[18] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. In *ICCV*, 2005. 1

[19] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 1

[20] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, 2010. 1

[21] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 4

[22] C. Wang, D. Blei, and F. Li. Simultaneous image classification and annotation. In *CVPR*, 2009. 1

[23] B. Zhao, L. Fei-Fei, and E. Xing. Image Segmentation with Topic Random Fields. In *ECCV*, 2010. 1, 4