Learning Representations for Multimodal Data with Deep Belief Nets

Nitish Srivastava NITISH@CS.TORONTO.EDU
University of Toronto, Toronto, ON. M5S 3G4 Canada

Ruslan Salakhutdinov RSALAKHU@UTSTAT.TORONTO.EDU
University of Toronto, Toronto, ON. M5S 3G4 Canada

Abstract

We propose a Deep Belief Network architecture for learning a joint representation of multimodal data. The model defines a probability distribution over the space of multimodal inputs and allows sampling from the conditional distributions over each data modality. This makes it possible for the model to create a multimodal representation even when some data modalities are missing. Our experimental results on bi-modal data consisting of images and text show that the Multimodal DBN can learn a good generative model of the joint space of image and text inputs that is useful for filling in missing data so it can be used both for image annotation and image retrieval. We further demonstrate that using the representation discovered by the Multimodal DBN our model can significantly outperform SVMs and LDA on discriminative tasks.

1. Introduction

Information in the real world comes through multiple input channels. Images are associated with captions and tags, videos contain visual and audio signals, sensory perception includes simultaneous inputs from visual, auditory, motor and haptic pathways. While each input modality conveys additional information, the information content of any modality is unlikely to be independent of the others. For example, images of forests and landscapes are strongly associated with tags like *nature* and *scenery*.

Presented at the *ICML Representation Learning Workshop*, Edinburgh, Scotland, UK, 2012.

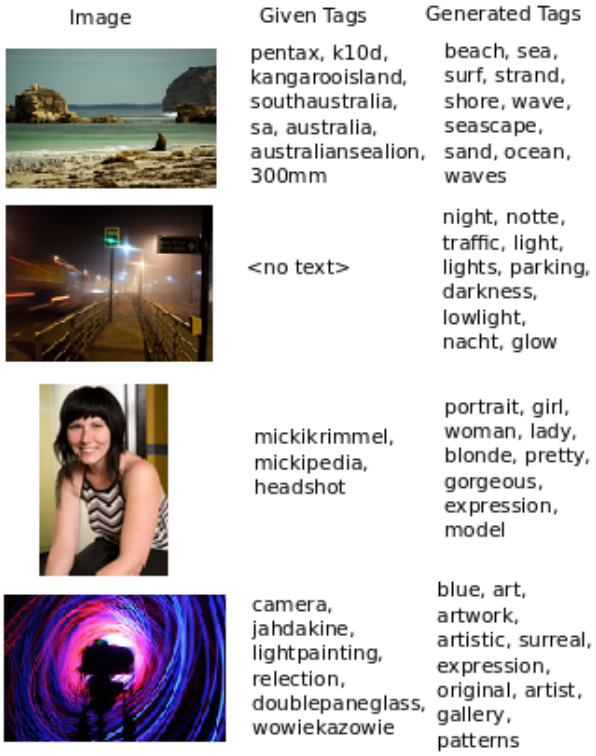

Figure 1. Examples of data from the MIR Flickr Dataset, along with text generated from the Deep Belief Net by sampling from $P(v_{txt}|v_{img}, \theta)$

The goal of this work is to learn a representation that takes this association into account. At the same time, the model must be able to handle missing data modalities so that the same kind of representation can be extracted even when some input channels are not available. One way to achieve this is by learning a joint density model over the space of multimodal inputs. Missing modalities can then be handled by sampling from the implied conditional distributions over missing modalities given the observed modalities. For example, we can use a large collection of user-tagged images to learn a distribution over images and text

$P(v_{img}, v_{txt}|\theta)$ such that it is easy to sample from $P(v_{txt}|v_{img}, \theta)$ and from $P(v_{img}|v_{txt}, \theta)$ so that we can do image annotation (Figure 1) and image retrieval (Figure 2). In addition, it is also desirable that the representation be useful for discriminative tasks, such as object recognition.

Before we describe our model in detail, it is useful to note why such a model is required. In many applications, observations come from different input channels each of which has a different representation and correlational structure. For example, text is usually represented as sparse word count vectors whereas an image is represented using pixel intensities or outputs of feature extractors which are real-valued and dense. This makes it much harder to discover relationships across modalities than relationships among features of the same modality. There is a lot of structure in the input but it is difficult to discover the highly non-linear relationships that exist between features across different modalities. Moreover, these observations are noisy and may have missing values. Using our probabilistic model, it will be possible to discover joint latent representations that capture relationships across various modalities. Different modalities typically carry different kinds of information. For example, people often caption an image to say things that may not be obvious from the image itself, such as the name of the person or place in the picture. It would not be possible to discover a lot of useful information about the world unless we do multimodal learning.

In this paper, we propose a model based on Deep Belief Nets (Hinton & Salakhutdinov, 2006). The key idea is to first use separate modality-friendly latent variable models to learn low-level representations of each data modality independently. For doing this we can leverage a large supply of unlabeled data to separately learn good generative models for each modality. Indeed, for many domains, including text retrieval, speech perception, and machine vision, unlabeled data is readily available. While the inputs to each of these separate models will typically belong to different modalities, our model will learn latent representations that are similar in form and correlational structure. The latent representations for different modalities can then be concatenated to form a multimodal input. Higher-order latent variables can then be used to model the distribution over this input. The posteriors over the higher-order variables can then be used to represent the multimodal input.

There have been several approaches to learning from multimodal data. In particular, Huiskes et al. (2010) showed that using captions, or tags, in addition to

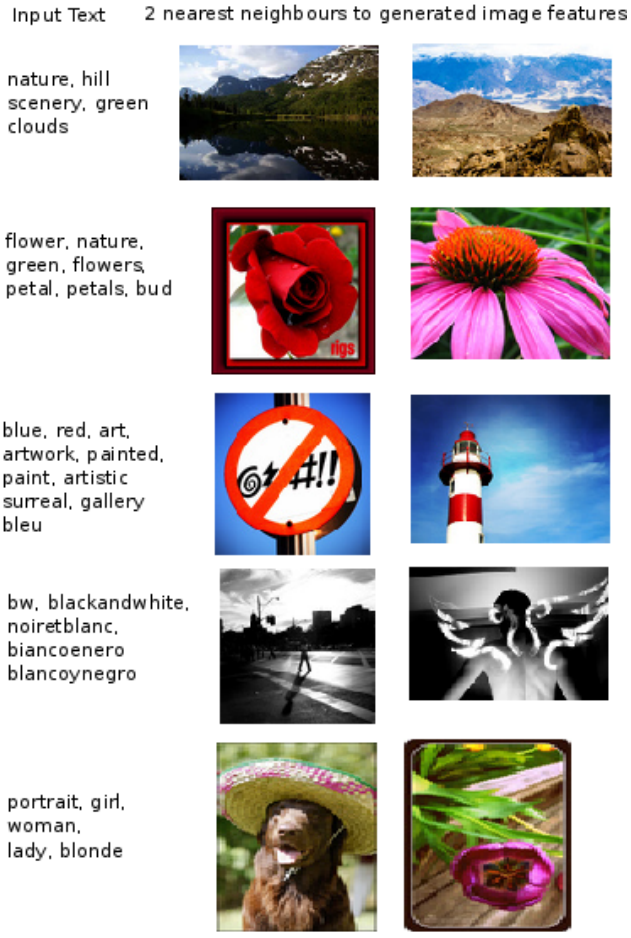

Figure 2. Examples of images retrieved using features generated from a Deep Belief Net by sampling from $P(v_{img}|v_{txt}, \theta)$

standard low-level image features significantly improves classification accuracy of SVM and LDA (Linear Discriminant Analysis) models. A similar approach of Guillaumin et al. (2010), based on multiple kernel learning framework, further demonstrated that an additional text modality can improve the accuracy of SVMs on various object recognition tasks. However, all of these approaches are discriminative by nature and cannot make use of large amounts of unlabeled data or deal easily with noisy or missing input modalities.

On the generative side, Xing et al. (2005) used dual-wing harmoniums to build a joint model of images and text, which can be viewed as a linear RBM model with Gaussian hidden units together with Gaussian and Poisson visible units. Most similar to our work is the recent approach of Ngiam et al. (2011) that used a deep autoencoder for speech and vision fusion. There are, however, several crucial differences. First, in this work we focus on integrating together very different data modalities: sparse word count vectors, and real-

valued dense image features. Second, we develop a Deep Belief Network as a generative model as opposed to unrolling the network and finetuning it as an autoencoder. While both approaches have lead to interesting results in several domains, using a generative model is important here as it allows our model to easily handle missing data modalities.

2. Background: RBMs and Their Generalizations

2.1. Restricted Boltzmann Machines

A Restricted Boltzmann Machine is an undirected graphical model with visible units $\mathbf{v} \in \{0,1\}^D$ and hidden units $\mathbf{h} \in \{0,1\}^F$ with each visible unit connected to each hidden unit. The model defines an energy function $E : \{0,1\}^{D+F} \rightarrow \mathbb{R}$

$$E(v,h;\theta) = -\mathbf{a}^\top \mathbf{v} - \mathbf{b}^\top \mathbf{h} - \mathbf{v}^\top W \mathbf{h},$$

where $\theta = \{\mathbf{a}, \mathbf{b}, W\}$ are the model parameters. The joint distribution over the visible and hidden units is defined by:

$$P(\mathbf{v},\mathbf{h};\theta) = \frac{1}{\mathcal{Z}(\theta)} \exp\left(-E(\mathbf{v},\mathbf{h};\theta)\right), \qquad (1)$$

where $\mathcal{Z}(\theta)$ is the normalizing constant.

2.2. Gaussian RBM

Consider modelling visible real-valued units $\mathbf{v} \in \mathbb{R}^D$ and let $\mathbf{h} \in \{0,1\}^F$ be binary stochastic hidden units. The energy of the state $\{\mathbf{v},\mathbf{h}\}$ of the Gaussian RBM is defined as follows:

$$E(\mathbf{v},\mathbf{h};\theta) = \sum_{i=1}^{D} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i=1}^{D}\sum_{j=1}^{F} \frac{v_i}{\sigma_i} W_{ij} h_j - \sum_{j=1}^{F} a_j h_j,$$

where $\theta = \{\mathbf{a}, \mathbf{b}, \mathbf{W}, \sigma\}$ are the model parameters. This leads to the following conditional distribution:

$$P(v_i|\mathbf{h};\theta) = \mathcal{N}\left(b_i + \sigma_i \sum_{j=1}^{F} W_{ij} h_j, \sigma_i^2\right) \qquad (2)$$

2.3. Replicated Softmax Model

The Replicated Softmax Model Salakhutdinov & Hinton (2009) is useful for modelling sparse count data, such as word count vectors in a document. Let $\mathbf{v} \in \mathbb{N}^K$ be a vector of visible units where v_k counts the number of times word k occurs in the document with the vocabulary of size K. Let $\mathbf{h} \in \{0,1\}^J$ be binary stochastic hidden topic features. The energy of the state $\{\mathbf{v},\mathbf{h}\}$ is defined as follows

$$E(v,h;\theta) = -\sum_{k=1}^{K}\sum_{j=1}^{J} W_{kj} h_j v_k - \sum_{k=1}^{K} v_k b_k - M \sum_{j=1}^{J} h_j a_j$$

where $\theta = \{\mathbf{a}, \mathbf{b}, \mathbf{W}\}$ are the model parameters and $M = \sum_k v_k$ is the total number of words in a document. The leads to the following conditional distribution:

$$P(v_k = 1|\mathbf{h};\theta) = \frac{\exp(-b_k + \sum_{j=1}^{J} W_{kj} h_j)}{\sum_{k'=1}^{K} \exp(-b_{k'} + \sum_{j=1}^{J} W_{k'j} h_j)} \qquad (3)$$

For all of the above models, exact maximum likelihood learning is intractable. In practice, efficient learning is performed by following an approximation to the gradient of the Contrastive Divergence (CD) objective (Hinton, 2002).

3. Multimodal Deep Belief Network

We illustrate the construction of a multimodal DBN using an image-text bi-modal DBN as our running example. Let $\mathbf{v}_m \in \mathbb{R}^D$ denote an image and $\mathbf{v}_t \in \mathbb{N}^K$ denote a text input. Consider modelling each data modality using a separate two-layer DBN (see Fig. 3). The probability that each DBN model assigns to a visible vector is:

$$P(\mathbf{v}_m) = \sum_{\mathbf{h}^{(1)},\mathbf{h}^{(2)}} P(\mathbf{h}^{(2)},\mathbf{h}^{(1)}) P(\mathbf{v}_m|\mathbf{h}^{(1)}) \qquad (4)$$

$$P(\mathbf{v}_t) = \sum_{\mathbf{h}^{(1)},\mathbf{h}^{(2)}} P(\mathbf{h}^{(2)},\mathbf{h}^{(1)}) P(\mathbf{v}_t|\mathbf{h}^{(1)}) \qquad (5)$$

The image-specific DBN uses Gaussian RBM to model the distribution over real-valued image features, whereas text-specific DBN uses Replicated Softmaxes to model the distribution over word count vectors. The conditional probabilities of the visibles given hiddens used in Eqs 4, 5 are as shown in Eqs 2, 3 respectively.

To form a multimodal DBN, we combine the two models by learning a joint RBM on top of them. The resulting graphical model is shown in Fig. 3, right panel. The joint distribution can be written as:

$$\begin{aligned} P(\mathbf{v}_m,\mathbf{v}_t) = & \sum_{\mathbf{h}_m^{(2)},\mathbf{h}_t^{(2)},\mathbf{h}^{(3)}} P(\mathbf{h}_m^{(2)},\mathbf{h}_t^{(2)},\mathbf{h}^{(3)}) \times \\ & \sum_{\mathbf{h}_m^{(1)}} P(\mathbf{v}_m|\mathbf{h}_m^{(1)}) P(\mathbf{h}_m^{(1)}|\mathbf{h}_m^{(2)}) \times \\ & \sum_{\mathbf{h}_t^{(1)}} P(\mathbf{v}_t|\mathbf{h}_t^{(1)}) P(\mathbf{h}_t^{(1)}|\mathbf{h}_t^{(2)}). \end{aligned} \qquad (6)$$

The parameters of this mulitmodal DBN can be learned approximately by greedy layer-wise training using CD.

Note that the Multimodal DBN can be described as a composition of unimodal pathways. Each pathway is learned separately in a completely unsupervised fashion, which allows us to leverage a large supply of unlabeled data. Any number of pathways each with any

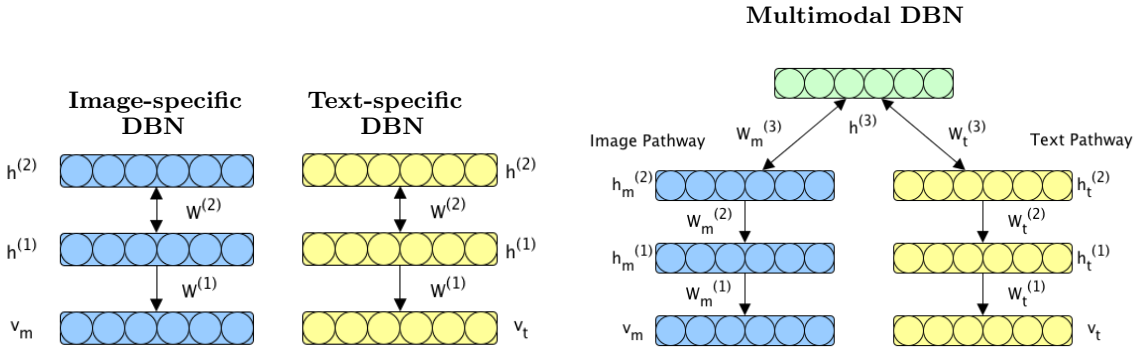

Figure 3. **Left:** Image-specific two-layer DBN that uses a Gaussian model to model the distribution over real-valued image features. **Middle:** Text-specific two-layer DBN that uses a Replicated Softmax model to model its distribution over the word count vectors. **Right:** A Multimodal DBN that models the joint distribution over image and text inputs.

number of layers could potentially be used. The type of lower RBMs in each layer could be different, accounting for different kinds of input distributions, as long as the final hidden representations at the end of each pathway are of the same type.

The intuition behind our model is as follows. Each data modality may have very different statistical properties which make it difficult for a shallow model to directly find correlations across modalities. The purpose of the independent modality-friendly models (Eq 4, 5) is to learn higher-level representations that remove such modality-specific correlations so that the top level RBM is presented with features that are relatively "modality-free", i.e., they are more alike in terms of their statistical properties than the original inputs were. In other words, given the original inputs, it is easy to say which represents images and which represents text using their sparsity and correlational structure. But, looking at the higher-level hidden features in the DBNs, it is more difficult to make such a distinction. Hence, the top-level joint RBM can pick up cross-modal relationships easily.

3.1. Generative Tasks

As argued in the introduction, many real-world applications will often have one or more of its modalities missing. We can infer missing values by drawing samples from the conditional model, which would allow us to properly use all input channels.

As an example, consider generating text conditioned on a given image[1] $\mathbf{v}_m$. We first infer the values of the hidden variables $\mathbf{h}_m^{(2)}$ in the image pathway by forward propagating $\mathbf{v}_m$ through to the last hidden layer. Conditioned on $\mathbf{h}_m^{(2)}$ at the top level RBM, we can perform alternating Gibbs sampling using the following conditional distributions:

$$P(\mathbf{h}^{(3)}|\mathbf{h}_m^{(2)}, \mathbf{h}_t^{(2)}) = \sigma\left(W_m^{(3)}\mathbf{h}_m^{(2)} + W_t^{(3)}\mathbf{h}_t^{(2)} + \mathbf{b}\right), \quad (7)$$

$$P(\mathbf{h}_t^{(2)}|\mathbf{h}^{(3)}) = \sigma\left(W_t^{(3)\top}\mathbf{h}^{(3)} + \mathbf{a}_t\right), \quad (8)$$

where $\sigma(x) = 1/(1+e^{-x})$. The sample $\mathbf{h}_t^{(2)}$ can then be propagated back through the text pathway to generate a distribution over the softmax vocabulary. This distribution can then be used to sample words.

3.2. Discriminative Tasks

The model can also be used for classification tasks by adding a discriminative layer of weights on top of the Multimodal DBN and finetuning the network to optimize a cross-entropy objective. In our experiments we use a simple logistic classifier to do 1-vs-all classification and finetune the model with stochastic gradient descent.

4. Experiments

4.1. Dataset and Feature Extraction

The MIR Flickr Data set (Huiskes & Lew, 2008) was used in our experiments. The data set consists of 1 million images retrieved from the social photography website Flickr along with their user assigned tags. The collection includes images released under the Creative Commons License. Among the 1 million images, 25,000 have been annotated for 24 concepts including object categories such as *bird*, *tree*, *people* and scene categories like *indoor*, *sky* and *night*. For 14 of them, a stricter labelling was done in which an image was assigned an annotation only if the corresponding category was salient in the image. This leads to a total of 38 classes. Each image may belong to several classes. The unlabeled 975,000 images were used only for pretraining the DBN. We use 15,000 images for training and 10,000 for testing, following (Huiskes et al., 2010). Mean Average Precision (MAP) is used as the perfor-

[1] Generating image features conditioned on text can be done in a similar way.

mance metric. Results are averaged over 10 random splits of training and test sets.

There are more than 800,000 distinct tags in the dataset. In order to keep the text representation manageable, each text input was represented using a vocabulary of the 2000 most frequent tags. After restricting to this vocabulary, the average number of tags associated with an image is 5.15 with a standard deviation of 5.13. There are 128,501 images which do not have any tags out of which 4,551 are in the labelled set. Hence about 18% of the labelled data does not have any tags. Word counts w were replaced with $\lceil \log(1+w) \rceil$. We concatenated Pyramid Histogram of Words (PHOW) features (Bosch et al., 2007), Gist (Oliva & Torralba, 2001) and MPEG-7 descriptors (Manjunath et al., 2001) (EHD, HTD, CSD, CLD, SCD) to get a 3857 dimensional representation of images. Each dimension was mean-centered. PHOW features are bags of image words obtained by extracting dense SIFT features over multiple scales and clustering them.

4.2. Model Architecture and Learning

The image pathway consists of a Gaussian RBM with 3857 visible and 1000 hidden units, followed by another layer of 1000 hidden units. The text pathway consists of a Replicated Softmax Model with 2000 visible and 1000 hidden units followed by another layer of 1000 hidden units. The joint layer also contains 1000 hidden units. The model was not found to be very sensitive to the choice of these hyperparameters.

We pretrained each pathway with greedy layer-wise CD1. The variance of each Gaussian unit was fixed to be its empirical variance in the training set. For discriminative tasks, we perform 1-vs-all classification using logistic regression on the last layer of hidden units in the joint model. The entire network was finetuned with stochastic gradient descent for each of the 38 classes separately since the class labels overlap. We split the 15K training set into 10,000 for training and 5,000 for validation.

4.3. Discriminative Aspect

In our first set of experiments, we evaluate the multimodal DBN as a discriminative model. Table 1 shows the results of our comparison with Linear Discriminant Analysis (LDA) and Support Vector Machines (SVMs) (Huiskes et al., 2010). The LDA and SVM models were trained using the labelled data on concatenated image and text features. Moreover, SIFT-based features were not used. Hence, to make a fair comparison, we first trained our model without using unlabeled data and using a similar set of features (i.e., excluding our SIFT-based features). We call this model **DBN-Lab**. Table 1 shows that the DBN-Lab model already outperforms its competitor SVM and LDA models across many classes. DBN-Lab achieves a MAP (mean Average Precision over 38 classes) of 0.503. This is compared to 0.475 and 0.492 achieved by SVM and LDA models.

To quantify the effect of using unlabeled data, we next trained a Multimodal DBN that used all of 975,000 unlabeled examples. We call this model **DBN-Unlab**. The only difference between the DBN-Unlab and DBN-Lab models is that DBN-Unlab used unlabeled data during its pretraining stage. The input representation for both models remained the same. Not surprisingly, the DBN-Unlab model significantly improved upon DBN-Lab almost across all classes, achieving a MAP of 0.532. Next, we trained a third model, called **DBN**, that used SIFT-based features along with unlabeled data. Table 1 shows that using SIFT features provided additional gains in model performance, achieving a MAP of **0.563**.

We also compare to an autoencoder that was initialized with the DBN weights and finetuned as proposed in Ngiam et al. (2011) **AUTOENCODER**. It performs much better than SVM and LDA getting a MAP of 0.547. It does better than the DBN model on some categories, however, on average it does not do as well. Notice that the autoencoder model does quite well on object-level categories such as *bird*, *car* and *food*.

There are several scenarios in which one may want to use the multimodal DBN for classification. The simplest is the case where images and associated tags are available for both training and testing. However, it is often the case that some training and test cases may not have tags at all. For example, in our setting, 18% of the labelled data has no text input. One way to deal with this problem is to simply use a text input of 0 in cases where there are no tags. All the models discussed till now correspond to this scenario. i.e., the training and test sets are used as given, (with a zero text input when no tags are present).
There is an alternative way of dealing with missing text. The generative model defined by the DBN can be used to infer a text input conditioned on the image input. This reconstructed text can then be used to fill in the missing text. To see whether this method of completing missing data is useful for classification, we train discriminative models using the training set as given but at test time, missing text data is filled in using the method described in section 3.1. We call this model **DBN-Recon**. Mean-field inference was used in

Table 1. Comparison of AP scores of various Mutlimodal DBNs with SVM and LDA models on the MIR Flickr Dataset.

LABELS	ANIMALS	BABY	BABY*	BIRD	BIRD*	CAR	CAR*	CLOUDS	CLOUDS*	DOG
RANDOM	0.129	0.010	0.005	0.030	0.019	0.047	0.015	0.148	0.054	0.027
LDA	0.537	**0.285**	**0.308**	0.426	0.500	0.297	0.389	0.651	0.528	**0.621**
SVM	0.531	0.200	0.165	0.443	0.520	0.339	0.434	0.695	0.434	0.607
DBN-LAB	0.498	0.129	0.134	0.184	0.255	0.309	0.354	0.759	0.691	0.342
DBN-UNLAB	**0.633**	0.096	0.088	0.431	0.499	0.310	0.422	0.730	0.658	0.568
AUTOENCODER	0.602	0.156	0.121	**0.461**	**0.547**	**0.366**	**0.526**	0.735	0.684	0.605
DBN	0.625	0.115	0.128	0.382	0.459	0.341	0.486	0.772	**0.739**	0.457
DBN-RECON	0.632	0.135	0.190	0.412	0.506	0.346	0.440	**0.796**	0.730	0.513
LABELS	DOG*	FEMALE	FEMALE*	FLOWER	FLOWER*	FOOD	INDOOR	LAKE	MALE	MALE*
RANDOM	0.024	0.247	0.159	0.073	0.043	0.040	0.333	0.032	0.243	0.146
LDA	**0.663**	0.494	0.454	0.560	0.623	0.439	0.663	0.258	0.434	0.354
SVM	0.641	0.465	0.451	0.480	0.717	0.308	0.683	0.207	0.413	0.335
DBN-LAB	0.376	0.540	0.478	0.593	0.679	0.447	0.750	0.262	0.503	0.406
DBN-UNLAB	0.598	0.555	0.505	**0.645**	0.718	0.484	0.745	0.246	0.479	0.395
AUTOENCODER	0.642	0.557	0.542	0.613	0.723	**0.558**	0.730	0.271	0.491	0.388
DBN	0.515	**0.588**	**0.564**	0.643	**0.765**	0.491	0.754	**0.281**	0.522	0.436
DBN-RECON	0.567	0.588	0.545	0.616	0.757	0.482	**0.757**	0.266	**0.529**	**0.442**
LABELS	NIGHT	NIGHT*	PEOPLE	PEOPLE*	PLANT_LIFE	PORTRAIT	PORTRAIT*	RIVER	RIVER*	SEA
RANDOM	0.108	0.027	0.415	0.314	0.351	0.157	0.153	0.036	0.006	0.053
LDA	0.615	0.420	0.731	0.664	0.703	0.543	0.541	0.317	**0.134**	0.477
SVM	0.588	0.450	0.748	0.565	0.691	0.480	0.558	0.158	0.109	0.529
DBN-LAB	0.655	0.483	0.800	0.730	0.791	0.642	0.635	0.263	0.110	0.586
DBN-UNLAB	0.674	0.467	0.826	0.764	0.791	0.630	0.627	0.244	0.051	0.588
AUTOENCODER	0.657	0.464	0.791	0.742	0.769	0.655	0.656	0.240	0.016	0.608
DBN	**0.698**	0.567	**0.837**	**0.788**	**0.823**	0.691	0.690	**0.351**	0.103	**0.647**
DBN-RECON	0.684	**0.585**	0.836	0.780	0.819	**0.696**	**0.693**	0.296	0.077	0.644
LABELS	SEA*	SKY	STRUCTURES	SUNSET	TRANSPORT	TREE	TREE*	WATER	**MEAN**	
RANDOM	0.009	0.316	0.400	0.085	0.116	0.187	0.027	0.133	0.124	
LDA	0.197	0.800	0.709	0.528	0.411	0.515	0.342	0.575	0.492	
SVM	0.201	0.823	0.695	0.613	0.369	0.559	0.321	0.527	0.475	
DBN-LAB	0.259	0.873	0.787	0.648	0.406	0.660	0.483	0.629	0.503	
DBN-UNLAB	0.245	0.860	0.786	0.636	0.421	0.596	0.511	0.675	0.532	
AUTOENCODER	0.357	0.836	0.761	0.625	0.460	0.641	0.513	0.683	0.547	
DBN	0.359	**0.888**	**0.811**	**0.679**	**0.464**	0.679	0.539	0.703	0.563	
DBN-RECON	**0.419**	0.885	0.811	0.670	0.443	**0.679**	**0.546**	**0.712**	**0.566**	

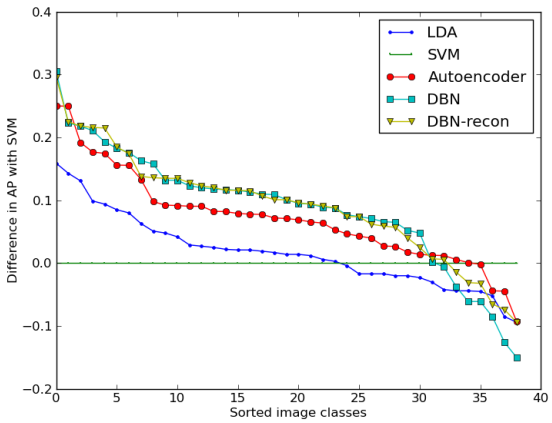

Figure 4. Visual comparison of LDA, SVM, Autoencoder, DBN and DBN-Recon models from Table 1

place of Gibbs Sampling to reduce noise. Table 1 shows that on average, the DBN-Recon model slightly outperforms the DBN model, achieving an average MAP of **0.566** compared to DBN's 0.563. Our best models give significant improvements over SVMs and LDA for almost all classes. For some classes they outperform them by a very large margin e.g., class *sea** goes from 0.201 (SVM) to 0.419 (DBN-Recon), *tree** from 0.321 to 0.546 and *clouds** from 0.434 to 0.739). Figure 4 shows the difference in AP scores of all the models in Table 1 with respect to the SVM model. The DBN and DBN-Recon curves outperform other models over majority of classes.

4.4. Multimodal Aspect

While the above experiments showed that DBNs outperform other multimodal methods, it is not obvious that learning multimodal features helps over using only one input modality. In this set of experiments, we focus on evaluating the ability of our model to learn multimodal features that are better for discriminative tasks than unimodal features. In Table 2 we compare our model with an SVM over Image features alone (**Image-SVM**) (Huiskes et al., 2010), a DBN over image features alone (**Image-DBN**) and a DBN over text features alone (**Text-DBN**). The unimodal DBNs were constructed by adding one extra layer to the unimodal pathways used for the multimodal DBNs, so that the number of hidden layers in all the DBNs is the same. The best multimodal DBN (**DBN-Recon**) clearly achieves far better overall performance. However, one may not find this to be very impressive given that the multimodal model had more data available to it at test time than any of the other models which used either image or text features only.

Therefore, to make a fair comparison, we conducted the following experiment. We take a multimodal DBN model that was pretrained and finetuned with both image and text features. However, at test time only image features are provided as input and the text input is replaced by zeros. This model is shown as **DBN-**

Table 2. Evaluation of the multimodal aspect of the model. Multimodal DBNs outperform unimodal models even when only one modality is given at test time.

Labels	animals	baby	baby*	bird	bird*	car	car*	clouds	clouds*	dog
Image-SVM	0.278	0.084	0.088	0.128	0.129	0.179	0.227	0.651	0.511	0.155
Image-DBN	0.348	**0.343**	**0.245**	0.424	0.384	**0.486**	0.407	0.601	0.403	0.106
Text-DBN	**0.650**	0.044	0.017	**0.512**	**0.598**	0.322	**0.463**	0.543	0.382	**0.615**
DBN-NoText	0.372	0.130	0.117	0.146	0.222	0.293	0.437	0.770	0.707	0.228
DBN-NoText-Recon	0.400	0.101	0.089	0.115	0.175	0.271	0.453	0.768	0.713	0.281
DBN-Recon	0.632	0.135	0.190	0.412	0.506	0.346	0.440	**0.796**	**0.730**	0.513
Labels	dog*	female	female*	flower	flower*	food	indoor	lake	male	male*
Image-SVM	0.156	0.461	0.389	0.469	0.519	0.293	0.605	0.188	0.407	0.294
Image-DBN	0.301	0.351	**0.625**	0.595	0.590	0.364	0.617	0.225	0.470	0.334
Text-DBN	**0.651**	0.531	0.476	0.576	0.662	**0.488**	0.672	0.234	0.474	0.378
DBN-NoText	0.280	0.551	0.509	0.487	0.621	0.437	0.716	0.264	0.494	0.397
DBN-NoText-Recon	0.311	0.560	0.527	0.524	0.636	0.433	0.720	0.245	0.493	0.396
DBN-Recon	0.567	**0.588**	0.545	**0.616**	**0.757**	0.482	**0.757**	**0.266**	**0.529**	**0.442**
Labels	night	night*	people	people*	plant_life	portrait	portrait*	river	river*	sea
Image-SVM	0.554	0.390	0.631	0.558	0.687	0.493	0.493	0.179	0.102	0.366
Image-DBN	0.337	0.240	0.420	0.389	0.481	0.415	0.609	**0.372**	0.116	0.318
Text-DBN	0.425	0.316	0.769	0.691	0.672	0.485	0.481	0.273	0.042	0.460
DBN-NoText	0.647	0.463	0.769	0.707	0.782	0.638	0.639	0.235	0.104	0.533
DBN-NoText-Recon	0.665	0.489	0.776	0.730	0.795	0.652	0.655	0.206	**0.131**	0.577
DBN-Recon	**0.684**	**0.585**	**0.836**	**0.780**	**0.819**	**0.696**	**0.693**	0.296	0.077	**0.644**
Labels	sea*	sky	structures	sunset	transport	tree	tree*	water	**MEAN**	
Image-SVM	0.126	0.775	0.626	0.588	0.298	0.514	0.205	0.448	0.375	
Image-DBN	0.363	0.622	0.586	0.579	0.352	0.600	0.218	0.457	0.413	
Text-DBN	0.147	0.726	0.759	0.480	**0.475**	0.480	0.299	0.612	0.471	
DBN-NoText	0.258	0.863	0.745	0.656	0.410	0.666	0.537	0.567	0.484	
DBN-NoText-Recon	0.300	0.877	0.760	**0.673**	0.394	0.675	0.542	0.579	0.492	
DBN-Recon	**0.419**	**0.885**	**0.811**	0.670	0.443	**0.679**	**0.546**	**0.712**	**0.566**	

NoText in Table 2. Observe that the DBN-NoText model performs significantly better than both SVM and DBN image only models. This result suggests that *learning multimodal features helps even when some modalities are absent at test time.* Having multiple modalities regularizes the model and makes it learn much better features. Moreover, this means that we do not need to learn separate models to handle each possible combination of missing data modalities. One joint model can be deployed at test time and used for any situation that may arise.

We can further improve performance if missing text input is inferred using the generative model and provided as input to the discriminative model at test time. This model is shown as **DBN-NoText-Recon**. Figure 5 shows the difference in AP scores of all the models in Table 2 with respect to an Image-SVM. The DBN-Recon curve outperforms other models over all classes. The DBNs that use only unimodal inputs (DBN-NoText and DBN-NoText-Recon) do better than other unimodal models.

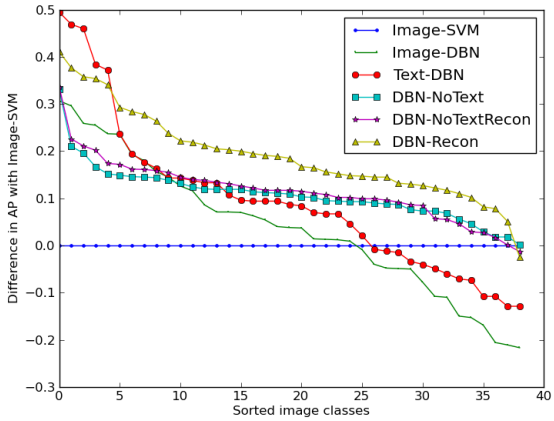

Figure 5. Visual comparison of models in Table 2

4.5. Generative Aspect

To evaluate the generative aspect of our model qualitatively, we look at samples of text generated from the multimodal DBN by conditioning on images taken from the test set. The images were chosen so as to cover a large number of the 38 categories. They are shown along with generated text in Figure 6. The model is extremely good at inferring text for images belonging to scene level categories such as *clouds*, *night**, *sea**, and *water*. Looking at the AP scores in Table 1 and comparing DBN-Recon with DBN, we see that for these classes significant gains in AP scores were made, e.g., *sea** goes from 0.359 to 0.419 (a relative improvement of 16%). For finer categories like *food* and *transport* it does not help improve classification accuracy.

We also look at images that were retrieved based on features generated from the model conditioned on text. Figure 2 shows some results where we retrieve images from a subset of the test set consisting of 4000 randomly chosen images. We start with a manually chosen piece of text and infer image features conditioned on it. Then we find the nearest neighbors to these features and retrieve the corresponding images. We used the L2 distance between the feature vectors to find nearest neighbors where all features were normalized to have zero mean and unit variance.

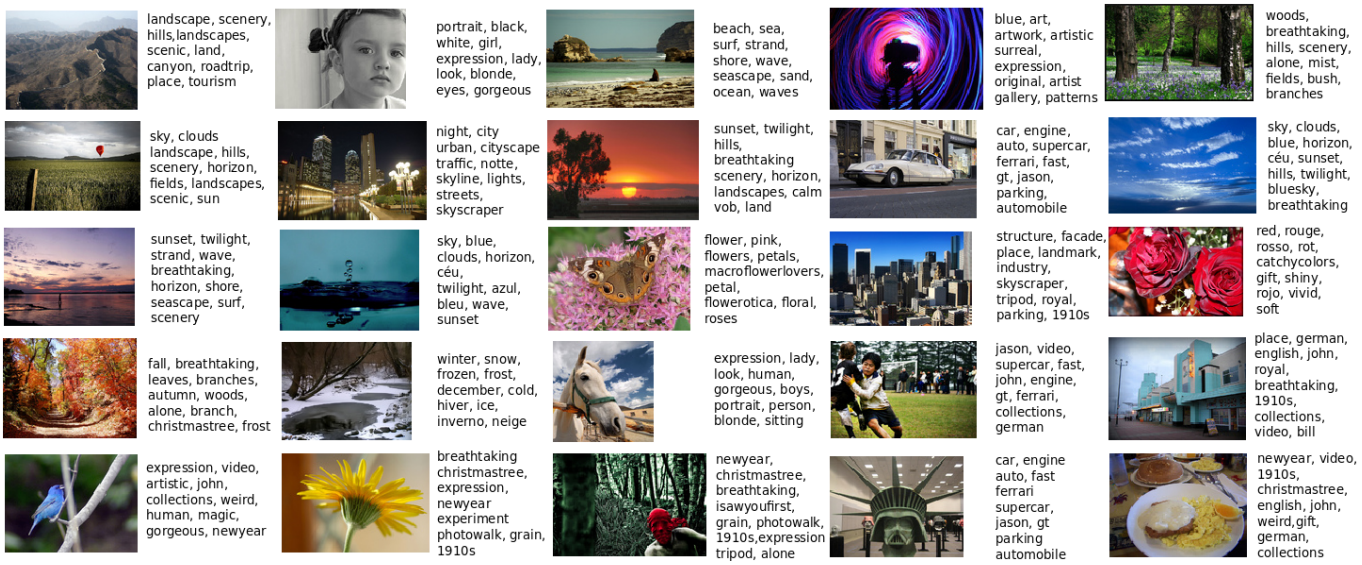

Figure 6. Examples of text generated by the DBN conditioned on images

5. Conclusions and Future Work

We proposed a Deep Belief Network architecture for learning multimodal data representations. The model fuses multiple data modalities into a joint hidden representation. The model defines a joint density model over multimodal input space that can be used for filling in missing inputs. It also performs well on discriminative tasks. When only one data modality is present at test time, it fills in the missing data and performs better than unimodal models which were trained on one modality alone. Qualitative evaluation of the model for image annotation and retrieval suggests that it learns meaningful conditional distributions. Large amounts of unlabeled data can be effectively utilized by the model. Pathways for each modality can be trained independently and "plugged in" together when learning multimodal features.

Our method benefits from the fact that the statistical properties of the final hidden representations across all pathways are similar. However, we did not explicitly impose any explicit objective to achieve this. It would be interesting to explore how this method can be improved by having an explicit penalty or constraint on certain properties of the hidden representations such as sparsity and entropy.

References

Bosch, A, Zisserman, Andrew, and Munoz, X. Image classification using random forests and ferns. *IEEE 11th International Conference on Computer Vision (2007)*, 23:1–8, 2007.

Guillaumin, M., Verbeek, J., and Schmid, C. Multimodal semi-supervised learning for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 902 –909, june 2010.

Hinton, Geoffrey and Salakhutdinov, Ruslan. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507, 2006.

Hinton, Geoffrey E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14 (8):1711–1800, 2002.

Huiskes, Mark J. and Lew, Michael S. The MIR Flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

Huiskes, Mark J., Thomee, Bart, and Lew, Michael S. New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative. In *Multimedia Information Retrieval*, pp. 527–536, 2010.

Manjunath, B.S., Ohm, J.-R., Vasudevan, V.V., and Yamada, A. Color and texture descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):703 –715, 2001.

Ngiam, Jiquan, Khosla, Aditya, Kim, Mingyu, Nam, Juhan, Lee, Honglak, and Ng, Andrew Y. Multimodal deep learning. In *International Conference on Machine Learning (ICML)*, Bellevue, USA, June 2011.

Oliva, Aude and Torralba, Antonio. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42: 145–175, 2001.

Salakhutdinov, Ruslan and Hinton, Geoffrey E. Replicated softmax: an undirected topic model. In *NIPS*, pp. 1607–1614. Curran Associates, Inc., 2009.

Xing, Eric P., Yan, Rong, and Hauptmann, Alexander G. Mining associated text and images with dual-wing harmoniums. In *UAI*, pp. 633–641. AUAI Press, 2005.