Multimodal Poisson Gamma Belief Network

Chaojie Wang and **Bo Chen**[*]
National Laboratory of Radar Signal Processing
Collaborative Innovation Center of Information Sensing&Understanding
Xidian University, Xi'an, Shaanxi, China

Mingyuan Zhou[*]
McCombs School of Business
University of Texas at Austin
Austin, TX 78712, USA

Abstract

To learn a deep generative model of multimodal data, we propose a multimodal Poisson gamma belief network (mPGBN) that tightly couple the data of different modalities at multiple hidden layers. The mPGBN unsupervisedly extracts a nonnegative latent representation using an upward-downward Gibbs sampler. It imposes sparse connections between different layers, making it simple to visualize the generative process and the relationships between the latent features of different modalities. Our experimental results on bi-modal data consisting of images and tags show that the mPGBN can easily impute a missing modality and hence is useful for both image annotation and retrieval. We further demonstrate that the mPGBN achieves state-of-the-art results on unsupervisedly extracting latent features from multimodal data.

Introduction

Data in the real world come through multiple input channels, typically exhibiting multiple modalities that carry different types of information. Different data modalities often have distinct statistical properties. For example, natural images, which are often represented with pixels or image descriptors, can also be described with the associated text (e.g., user tags or subtitles) and audio (e.g., human voice or natural sound).

To exploit the connections between different data modalities, there has been significant recent interest in multimodal learning. One of the leading approaches is using latent Dirichlet allocation (LDA) of Blei, Ng, and Jordan (2003), or other more sophisticated topic models. For example, to discover the relationship between the images and their associated annotations, correspondence LDA (Corr-LDA) one-to-one maps the image and text topics (Blei and Jordan 2003). Multimodal LDA generalizes Corr-LDA by learning a regression module relating the topics from different modalities (Putthividhy, Attias, and Nagarajan 2010). Besides annotated tags, the embedding of class labels can also help improve the discriminative power of the learned joint representation (Mcauliffe and Blei 2008). One appealing feature of the topic modeling based approach is that the task of extracting latent representation from the data can be easily framed as a probabilistic inference problem, which can be solved with routine procedures.

Another common approach to multimodal representation learning is to build a deep neural network for each modality, and share the top hidden layer of the networks of all modalities. For example, a deep autoencoder is used to learn a joint representation for speech and vision, showing that using both modalities for representation learning outperforms using only one modality (Ngiam et al. 2011). To infer a joint representation for image-text pairs, the multimodal deep belief network (DBN) of Srivastava and Salakhutdinov (2012a) uses a DBN for each modality and combine both DBNs by sharing a restricted Boltzmann machine (RBM) as their top hidden layer. The multimodal DBN is further generalized to multimodal deep Boltzmann machine (DBM) by replacing the DBNs with DBMs (Srivastava and Salakhutdinov 2012b). Another successful example is the multimodal deep recurrent neural network (MDRNN) of Sohn, Shang, and Lee (2014), which uses a recurrent encoding function to predict the target modality given the input modality, achieving state-of-the-art performance on the MIR-Flicker (Huiskes and Lew 2008) after fine-tuning the whole network.

Inspired by the success of both approaches for multimodal representation learning, we propose a multimodal Poisson gamma belief network (PGBN) that generalizes the PGBN of Zhou, Cong, and Chen (2016) to infer a nonnegative latent representation of multimodal data in an unsupervised manner. The PGBN is a Bayesian deep model that combines the interpretability of a topic model and the nonlinear modeling capability of a deep neural network. It can be equivalently represented as deep LDA (Cong et al. 2017) and is a deep generative model whose latent multilayer network structure can be easily interpreted. Before going into the technical details, we show in Fig. 1 how an image-tags pair is represented under the proposed multimodal PGBN via a sparse set of non-negligibly weighted multimodal latent features, where the chosen image topics for generating image are highly correlated with the key words of the corresponding text topic benefiting from our special model structure. In addition to providing easily interpretable deep multimodal latent representations, we show the multimodal PGBN achieves state-of-the-art results in predicting a missing modality conditioning on the other observed ones.

[*]Correspondence to: Bo Chen (bchen@mail.xidian.edu.cn), Mingyuan Zhou (mingyuan.zhou@mccombs.utexas.edu).

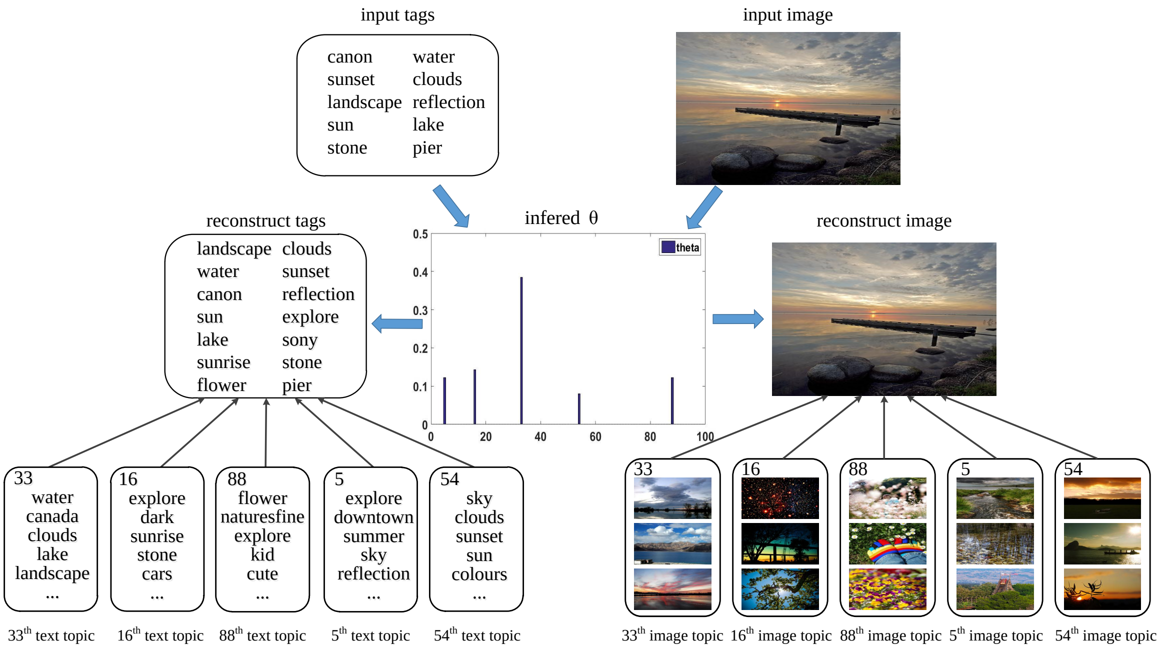

Figure 1: The generative process visualization of the input image-tags pair by visualizing the joint distribution and different modal topics learned from training data following the method proposed in section "Exploratory data analysis"

Preliminaries

In this section we briefly review the PGBN (Zhou, Cong, and Chen 2015; 2016), which serves as the building block for the proposed model for multimodal learning.

Denoting the j^{th} observed or K_0-dimensional count vectors as $x_j^{(1)} \in \mathbb{Z}^{K_0}$, where $\mathbb{Z} := \{0, 1, ...\}$ and the superscript indexes the layer, the generative model of the PGBN with T hidden layers, from top to bottom, is expressed as

$$\begin{aligned}
&\boldsymbol{\theta}_j^{(T)} \sim \mathrm{Gam}(\boldsymbol{r}, 1/c_j^{(T+1)}),\\
&\cdots\\
&\boldsymbol{\theta}_j^{(t)} \sim \mathrm{Gam}(\boldsymbol{\Phi}^{(t+1)}\boldsymbol{\theta}_j^{(t+1)}, 1/c_j^{(t+1)}),\\
&\cdots\\
&\boldsymbol{x}_j^{(1)} \sim \mathrm{Pois}(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}),\ \boldsymbol{\theta}_j^{(1)} \sim \mathrm{Gam}\left(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)}, \tfrac{p_j^{(2)}}{1-p_j^{(2)}}\right),
\end{aligned} \tag{1}$$

where the observed multivariate count vectors $\boldsymbol{x}_j^{(1)}$ are factorized under the Poisson likelihood. Defining the dimension of the t^{th} hidden layer as K_t, the shape parameters of the gamma distribution hidden units $\boldsymbol{\theta}_j^{(t)} \in \mathbb{R}_+^{K_t}$ of layer t , where $\mathbb{R}_+ = \{x : x \geq 0\}$, are factorized into the product of connection weight matrix $\boldsymbol{\Phi}^{(t+1)} \in \mathbb{R}_+^{K_t \times K_{t+1}}$ and the hidden units $\boldsymbol{\theta}_j^{(t+1)} \in \mathbb{R}_+^{K_{t+1}}$ of layer $t+1$, while the top layer's hidden units $\boldsymbol{\theta}_j^{(T)}$ share the same vector $\boldsymbol{r} = (r_1, ..., r_{K_T})'$ as their gamma shape parameters. The $p_j^{(2)}$ in PGBN are probability parameters and $\{1/c^{(t)}\}_{3,T+1}$ are gamma scale parameters, with $c_j^{(2)} := (1 - p_j^{(2)})/p_j^{(2)}$.

For scale identifiability and ease of inference, each column of $\boldsymbol{\Phi}^{(t)} \in \mathbb{R}_+^{K_{t-1} \times K_t}$ is restricted to have a unit L_1 norm and hence $0 \leq \boldsymbol{\Phi}^{(t)}(k', k) \leq 1$. To complete the hierarchical model, for $t \in \{1, ..., T-1\}$, we let

$$\boldsymbol{\phi}_k^{(t)} \sim \mathrm{Dir}(\eta^{(t)}, ..., \eta^{(t)}),\ r_k \sim \mathrm{Gam}(\gamma_0/K_T, 1/c_0), \tag{2}$$

where $\boldsymbol{\phi}_k^{(t)} \in \mathbb{R}_+^{K_{t-1}}$ is the kth column of $\boldsymbol{\Phi}^{(t)}$, we impose $c_0 \sim \mathrm{Gam}(e_0, 1/f_0)$ and $\gamma_0 \sim \mathrm{Gam}(a_0, 1/b_0)$, and for $t \in \{3, ..., T+1\}$, we let

$$p_j^{(2)} \sim \mathrm{Beta}(a_0, b_0),\ \ c_j^{(t)} \sim \mathrm{Gam}(e_0, 1/f_0). \tag{3}$$

In addition to fitting high-dimensional count data, Zhou, Cong, and Chen (2016) have introduced a set of link functions to extend the PGBN to model other types of data. If the observations are high-dimensional sparse binary vectors $\boldsymbol{b}_j^{(1)} \in \{0, 1\}^V$, they are factorized as

$$\boldsymbol{b}_j^{(1)} = 1(\boldsymbol{x}_j^{(1)} \geq 0),\ \boldsymbol{x}_j^{(1)} \sim \mathrm{Pois}(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}). \tag{4}$$

If the observations are high-dimensional nonnegative real-value vector $\boldsymbol{y}_j^{(1)} \in \mathbb{R}_+^V$, they are factorized as

$$\boldsymbol{y}_j^{(1)} \sim \mathrm{Gam}(\boldsymbol{x}_j^{(1)}, 1/a_j),\ \boldsymbol{x}_j^{(1)} \sim \mathrm{Pois}(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}). \tag{5}$$

Multimodal PGBN

Existing multimodal learning approaches often fall short of extracting interpretable multilayer hidden structures, which help visualize the connections between different modalities at different levels of abstraction. Building on the PGBN, we construct a novel multimodal PGBN (mPGBN) that well captures the correlations between different modalities at multiple levels of abstraction. We focus on analyzing image-text pairs and show that the mPGBN provides nicely coupled image and text topics at multiple different layers, and these coupled topics exhibit an increasing level of abstraction when moving towards a deeper hidden layer.

In analyzing image-text pairs, the proposed mPGBN can be considered as an integration of a text-specific PGBN and an image-specific one that share their latent representations at multiple layers. The text-specific PGBN can directly fit integer word count vectors or use the Bernoulli-Poisson link shown in (4) to model binary annotated tags, whereas the image-specific PGBN can fit positive image features such as pixel values using the Poisson randomized gamma link shown in (5) or model feature count vectors extracted from images. Below we explain the multimodal PGBN in detail, assuming the count vectors are input to both the image and text modalities.

Model Architecture

We first construct the mPGBN using two PGBNs that share all their multilayer hidden variables except for their connection weights between the visible layer and first hidden layer. From top to bottom, the generative model is expressed as

$$\begin{aligned} \boldsymbol{\theta}^{(T)}_{share\_j} &\sim \text{Gam}(\boldsymbol{r}_{share}, 1/c^{(T+1)}_{share\_j}), \\ &\cdots \\ \boldsymbol{\theta}^{(t)}_{share\_j} &\sim \text{Gam}(\boldsymbol{\Phi}^{(t+1)}_{share}\boldsymbol{\theta}^{(t+1)}_{share\_j}, 1/c^{(t+1)}_{share\_j}), \\ &\cdots \\ \boldsymbol{x}^{(1)}_{img\_j} \sim \text{Pois}(\boldsymbol{\Phi}^{(1)}_{img}\boldsymbol{\theta}^{(1)}_{share\_j}),\ & \boldsymbol{x}^{(1)}_{txt\_j} \sim \text{Pois}(\boldsymbol{\Phi}^{(1)}_{txt}\boldsymbol{\theta}^{(1)}_{share\_j}). \end{aligned} \tag{6}$$

The upward-downward Gibbs sampler of Zhou, Cong, and Chen (2016), each iteration of which upward samples Dirichlet distributed connection weight vectors starting from the first layer (bottom data layer), then downward samples gamma distributed hidden units starting from the top hidden layer, can be applied to train the hidden layers of the mPGBN, with the sampling update equation for the first hidden layer replaced as

$$\begin{aligned} (\theta^{(1)}_{share\_j} \,|\, -) \sim \text{Gam}(& m^{(1)(2)}_{img\_j} + m^{(1)(2)}_{tags\_j} + \Phi^{(2)}_{share}\theta^{(2)}_{share\_j}, \\ & [c^{(2)}_j - 2\ln(1 - p^{(1)}_j)]^{-1}), \end{aligned} \tag{7}$$

where $m^{(1)(2)}_{img\_j}$ and $m^{(1)(2)}_{tags\_j}$ are latent counts that are sampled separately from their corresponding modalities, but both directly influence the conditional posteriors of the hidden units of the the first hidden layer. Benefiting from training the whole network jointly, the mPGBN can not only capture the relationships between different layers from top to bottom, but also connect learned image themes and text topics tightly by coupling all their hidden layers. Note different from the multimodal DBN of Srivastava and Salakhutdinov (2012a) that is constructed by only sharing the top hidden layer, we share the whole network to tightly couple the image and text topics across all hidden layers.

The intuition behind our construction is that even though different data modalities may exhibit distinct statistical properties, there could be strong correlations between their latent representations at multiple levels of abstraction. In particular, the image and text in a pair can be considered as two different exhibitions of the same semantic meaning. For example, the image of a tiger and the word “tiger” share the semantic meaning at the same level, the image of a tiger and the word “big cat” share that at a higher abstraction level, and the image of a tiger and the word “carnivore” share that at an even higher abstraction level. It is our hope that our mPGBN could capture the shared latent structure at different levels of abstractions, which help better understand the semantic meanings of these multilayer latent representations. We show in Fig. 2 some example topics learned by the mPGBN, which clearly help understand how topics at different layers are related, understand the general and specific aspects of the image-text pairs used for training, and understand how the same level of abstraction is reflected in both the text and image modalities.

In comparison with conventional multimodal topic models that only relate different modalities at the single hidden layer, the mPGBN clearly provides much more expressive latent structure. With extensive experiments in text and image analysis, below we will further show that the mPGBN with two or more hidden layers clearly outperforms a shallow one in unsupervisedly extracting latent features for classification.

Adaptive Normalization

A potential issue that the mPGBN model in (6) faces is that the input to different modalities may be at very different scales. To address that potential issue, we propose to modify the mPGBN model as

$$\theta^{(1)}_{img\_j} = k_{img\_j}\theta^{(1)}_{share\_j},\ \theta^{(1)}_{txt\_j} = k_{txt\_j}\theta^{(1)}_{share\_j},$$

$$x^{(1)}_{img\_j} \sim \text{Pois}(\Phi^{(1)}_{img}\theta^{(1)}_{img\_j}),\ x^{(1)}_{txt\_j} \sim \text{Pois}(\Phi^{(1)}_{txt}\theta^{(1)}_{txt\_j}),$$

which means that the first hidden layers of both modalities only share their gamma shape parameters in the prior but have adaptive scale parameters to suit different input scales.

Related Work

Two key challenges in multimodal learning are learning a shared representation across modalities and predicting missing data (e.g., by synthesis or retrieval) in one modality conditional on the other ones. To learn a good representation from multimodal data, a naive approach is to concatenate the data descriptors from different sources of input. It results in a single high-dimensional multimodal feature vector for each observation, which often clearly helps improve classification accuracy (Huiskes and Lew 2008; Guillaumin, Verbeek, and Schmid 2010). However, that naive approach is not able to deal with missing modalities and often leads to a clear increase in computation for classification due to the increase of the feature dimension.

Some popular deep learning based approaches (Srivastava and Salakhutdinov 2012a; 2012b; Ngiam et al. 2011) may help address these issues, but there is no distinct association between different data modalities in these models and how to learn a good association between multiple data modalities remains a challenging question. Sohn, Shang, and Lee (2014) solve this problem by introducing the Variation of Information theory, but similar to conventional deep neural network structures trained with backpropagation, it is

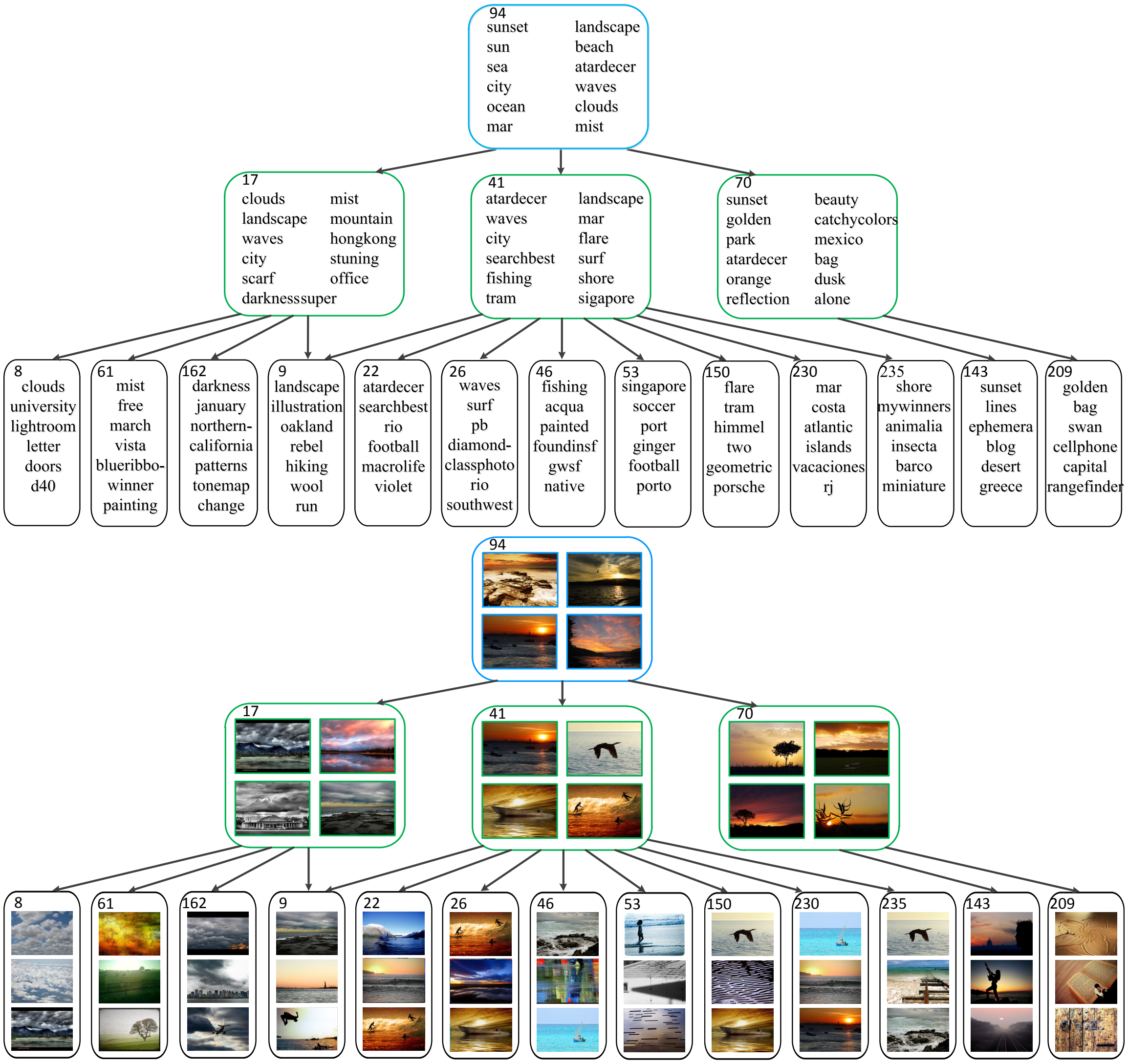

Figure 2: Two [13, 3, 1] modality-specific trees that include all the lower-layer nodes (directly or indirectly) linked with non-negligible weights to the 96^{th} node of the top layer, taken from the full $[500, 200, 100]$ network inferred by the mPGBN on 1995 image-text pairs selected from MIR-Flicker 25k whose annotated words are more than 10. A line from node k at layer t to node k' at layer $t-1$ indicates that $\Phi^{(t)}(k', k) > 10/K_{t-1}$. For each node on the text tree, 12 words of the corresponding topic are displayed inside the text box at layers three and two and 6 words at layer one. As for the image tree, the top-k nearest images are displayed inside the image box evaluated using the cosine distances between the inferred image features and the features of the images from MIR-Flicker 25k.

often difficult for a conventional deep learning approach to express, let along visualize, the relationships of its hidden layers in a multimodal learning setting.

In contrast to conventional deep networks, the mPGBN has an excellent ability in exploratory data analysis, as illustrated in Fig. 2, where we visualize various aspects of the data and how they are related to each other, by following the paths of a tree extracted from the learned deep network. Below we provide further experiments to demonstrate that the mPGBN can be used to impute missing modalities, and extract excellent latent features for additional downstream analysis.

Experimental Results

Dataset and Feature Extraction

We use in our experiments the MIR-Flicker data set (Huiskes and Lew 2008), which consists of 1 million images along with their user assigned tags that are retrieved from the social photography website Flicker. Among these images, 25,000 have been annotated for 24 concepts including object categories such as bird, tree, and people, and scene categories such as indoor, sky, and night. For 14 of them, a stricter labeling was done in which an image was assigned an annotation only if the corresponding category was salient

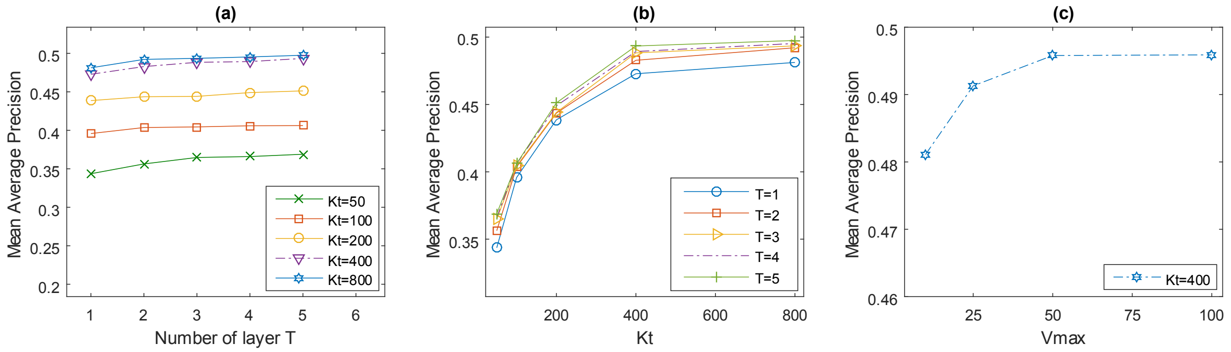

Figure 3: Mean Average Precision of the mPGBN for MIR-Flicker 25k classification (a) as a function of the depth T with various $K_t \in \{50, 100, 200, 400, 800\}$ and (b) as a function of K_t with various depths $T \in \{1, 2, 3, 4, 5\}$. (c) Comparison of Mean Average Precision among various $V_{max} \in \{10, 25, 50, 100\}$ in a same network architecture with $T = 3$ and $K_t = 400$.

in the image. This leads to a total of 38 classes where each image may belong to several different classes.

To compare with the results of multimodal DBM, we use the same text and image features used in Srivastava and Salakhutdinov (2012b). Each text input is represented using a vocabulary consisting of the 2000 most frequent tags. Each image is represented by a 3857-dimensional feature vector consisting of Pyramid Histogram of Words (PHOW) (Bosch, Zisserman, and Munoz 2007), Gist (Oliva and Torralba 2001), and MPEG-7 descriptors including EHD, HTD, CSD, CLD, and SCD (Manjunath et al. 2001). Publicly available code (Vedaldi and Fulkerson 2010; Bastan et al. 2010) could be used to extract these features. To match our model, each dimension could be discretized to $[0, V_{\max}]$ to produce count input or subtracted by the minimum of each dimension to provide nonnegative real input. Here we use the count input, and our sensitive analysis below shows that it is simple to find an appropriate $V_{\max}$.

Model Architecture and Learning

We first focus on understanding the influence of the network depth and the upper-bound imposed on the network width. We test the mPGBN for unsupervisedly extracting latent features that are to be used for classification on MIR-Flicker 25k. For hyper-parameters, we set $\eta^t = 0.05$ for all t, $a_0 = b_0 = 0.01$, and $e_0 = f_0 = 1$. We use 15k image-text pairs randomly selected from MIR-Flicker 25k to infer a set of networks with $T \in \{1, 2, 3, 4, 5\}$ and $K_t \in \{50, 100, 200, 400, 800\}$, and apply the upward-downward Gibbs sampler to collect 200 MCMC samples after 200 burn-in to estimate the posterior mean of the latent representation of each test data sample. Using the extracted first-hidden-layer representations as the feature vectors, we perform 1-vs-all classification with logistic regression. Mean Average Precision (MAP) is used as the performance metric in our experiments and the results in Figs. 3 (a) and (b) show a clear trend of improvement in MAP by increasing the depth with the layer widths fixed, or by increasing the widths of the hidden layers with the depth fixed.

Another factor that may affect the performance of the mPGBN is the selection of the $V_{\max}$ value. Hence we test $V_{\max} \in \{10, 25, 50, 100\}$ in a fixed network architecture with $T = 3$ and $K_t = 400$. As shown in Fig. 3 (c). Although increasing $V_{\max}$ in general improves the performance of the mPGBN, the performance gain quickly diminishes once $V_{\max}$ becomes sufficiently large. To achieve a compromise between the performance and computation, we set $V_{\max} = 25$ in all following experiments.

Generative Task

In our second set of experiments, we qualitatively evaluate the generative ability of the mPGBN. Fig. 4 shows the tags generated conditioning on their corresponding images, which are from MIR-Flicker and cover a variety of different categories. From the results, it is clear that the mPGBN can successfully impute the missing text given the image. For example, given the third image of the first row in Fig. 4, the mPGBN not only captures scene level features like "snow" and "winter," as the main part of the image is white, but also captures more subtle information such as "people" and "tree" that also appear in the image.

We have also examined the images that are retrieved based on the image features generated from the proposed model conditioned on the text, as shown in Fig. 5. More specifically, we generate image features conditioned on the text shown in the left part of Fig. 5, and then retrieve from MIR-Flicker 25k the top 5 images, whose features are closest to the generated image features measured by the cosine distance.

Exploratory data analysis

Our intuition in this third part is visualizing the topics of different layers to understand the general and specific aspects of the image-text pairs used to train our models, and further illustrate how the topics of different layers are related to each other and reveal the relationships between image themes and text topics, via their projections to the bottom data layer.

To verify this intuition, we consider constructing trees to visualize the mPGBN learned from subsets of MIR-Flicker, setting a network structure as $[K_1, K_2, K_3] = [500, 200, 100]$. Pick a node at top layer as the root of a tree

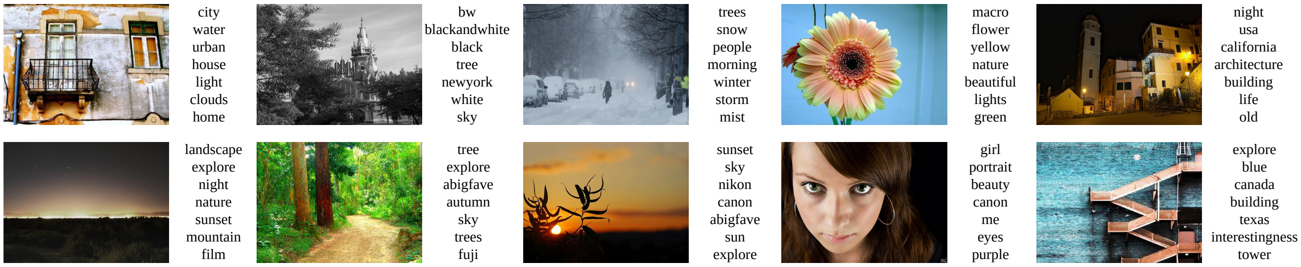

Figure 4: Examples of the tags generated by the multimodal PGBN conditioned on the images.

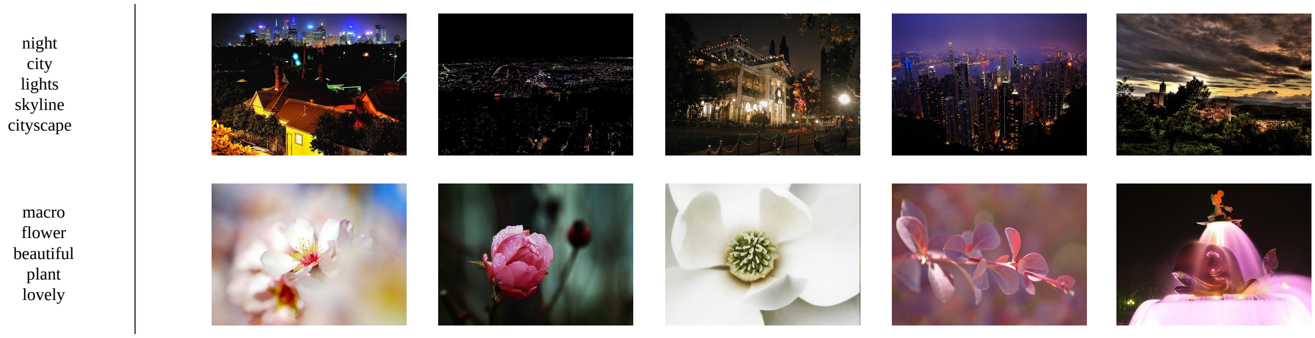

Figure 5: Top-5 nearest images retrieved using the features generated by the multimodal PGBN conditioning on the tags.

and grow the tree downward by drawing a line from node k at layer t, the root or a leaf node of the tree, to node k' at layer $t-1$ for all k' in the set $\{k' : \Phi^{(t)}(k', k) > \tau_t/K_{t-1}\}$, and use τ_t to adjust the complexity of this tree. In general, increasing τ_t would discard more weak connections and hence make the tree sparser and easier to visualize.

We set $\tau_t = 10$ for all t to visualize the three-layer tree rooted at the 94^{th} node of the top hidden layer, as shown in Fig. 2. Following the branches of each tree shown in Fig. 2, it is clear that the text topics become more and more specific when moving along the tree from the top to bottom. The root node on "sunset landscape waves clouds mountains" splits into three nodes when moving from layer three to two, and the three nodes located at the second layer are mainly about "clouds landscape mountains," "landscape waves," and "sunset golden clouds."

The image tree can also be visualized in a similar way as mentioned above. Since the low-level features used in the paper cannot be directly visualized, the "key words" of a node in different layers are expressed with the top 4 or 3 nearest images that are retrieved using the topic feature of that node as shown in the bottom of Fig. 2. Comparing the text and image trees shown in Fig. 2, it is clear that the top retrieved images, which reveal the inferred features of an image topic, are highly correlated with the key words of the corresponding text topic in terms of semantic meanings. Taking the 70^{th} node of layer two as an example, the corresponding text topic is mainly about "sunset golden," while the corresponding image-topic are characterized by images related to "golden sunset." When moving form layer two to layer one, the 70^{th} text-node on "sunset golden" split into node 143 on "sunset" and node 209 on "golden," which are also the key elements appearing in the retrieved images of the corresponding nodes in the image tree.

Discriminative Task

To further evaluate the mPGBN and make comparison to previously proposed multimodal learning algorithms, we use the mPGBN to unsupervisedly extract latent features from the labeled 25k image-text pairs of the MIR-Flicker dataset (Huiskes and Lew 2008), where 15k image-text pairs are used for training and the remaining 10k pairs for testing. Following Srivastava and Salakhutdinov (2012a), we use the same 1857 dimensional image features and 2000 dimensional text features. We choose a two-hidden-layer mPGBN, with 1024 hidden units in both hidden layers. We use 1000 Gibbs sampling iterations to train the mPGBN on the 15k training image-text pairs, and retain the inferred network (global variables) of the last sample. For each test image-text pair, we collect 500 MCMC samples after 500 burn-in iterations to infer its latent representation (local variables) under the network retained after training.

With the extracted latent features, we perform 1-vs-all classification using logistic regression. Mean Average Precision (MAP) and Percsion@50 are used for evaluation and the results are averaged over 5 random training/testing partitions. Table 1 shows the comparison of MAP between the mPGBN and the multimodal learning models listed in Srivastava and Salakhutdinov (2012b). In addition, for the propose of showing the benefit of having a deep model, we include for comparison a single-hidden-layer PGBN, which reduces to the gamma-negative binomial process Poisson

Table 1: Comparison of AP scores and Precision@50 of various multimodal models on the MIR-Flicker 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	0.654	0.528	0.621
SVM	0.531	0.200	0.165	0.443	0.520	0.339	0.434	0.685	0.434	0.607
DBN	0.498	0.129	0.134	0.184	0.255	0.309	0.354	0.759	0.691	0.342
DBM	0.511	0.139	0.145	0.190	0.253	0.319	0.368	**0.768**	**0.723**	0.351
mPFA	0.603	0.260	0.297	0.487	0.531	0.332	0.496	0.643	0.509	0.601
mPGBN	**0.615**	**0.288**	**0.320**	**0.515**	**0.552**	**0.357**	**0.502**	0.657	0.554	**0.609**
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	0.414	0.335
DBN	0.376	0.540	0.478	0.593	0.679	0.447	0.750	0.262	0.503	0.406
DBM	0.385	0.535	0.493	0.604	0.668	0.462	**0.759**	**0.277**	**0.505**	**0.424**
mPFA	0.650	0.519	0.468	0.605	0.714	0.562	0.678	0.262	0.477	0.382
mPGBN	0.656	**0.551**	**0.497**	**0.614**	**0.736**	**0.579**	0.692	0.268	0.488	0.399
LABELS	NIGHT	NIGHT*	PEOPLE	PEOPLE*	PLANTLIFE	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	0.655	0.483	0.800	0.730	0.791	0.642	0.635	0.263	0.110	**0.586**
DBM	**0.666**	**0.505**	**0.802**	**0.742**	**0.794**	**0.651**	**0.665**	0.274	0.110	0.582
mPFA	0.599	0.373	0.768	0.692	0.744	0.522	0.516	0.299	0.118	0.524
mPGBN	0.625	0.407	0.781	0.719	0.759	0.547	0.541	0.301	0.121	0.533
LABELS	SEA*	SKY	STRUCTURES	SUNSET	TRANSPORT	TREE	TREE*	WATER	**MAP**	**Prec@50**
RANDOM	0.009	0.316	0.400	0.085	0.116	0.187	0.027	0.133	0.124	0.124
LDA	0.197	0.800	0.709	0.528	0.411	0.515	0.342	0.575	0.492	0.754
SVM	0.201	0.823	0.695	0.613	0.369	0.559	0.321	0.527	0.475	0.758
DBN	0.259	0.873	0.787	0.648	0.406	0.660	0.483	0.629	0.503	-
DBM	0.260	**0.883**	**0.796**	**0.659**	0.423	**0.668**	**0.492**	0.628	0.513	0.791
mPFA	0.280	0.798	0.748	0.510	0.445	0.520	0.360	0.622	0.515	0.834
mPGBN	**0.343**	0.809	0.764	0.516	**0.455**	0.539	0.377	**0.630**	**0.532**	**0.844**

factor analysis of Zhou and Carin (2015). We refer to this single-layer model to as multimodal Poisson factor analysis (mPFA).

As shown in Table 1, the shallow mPFA already clearly outperforms other models including SVM, LDA, and DBN by achieving a MAP of 0.515, and is comparable to DBM that achieves a MAP of 0.513. With two hidden layers, the mPGBN achieves the best MAP of 0.532 and outperforms mPFA in every single category, showing that introducing a deep structure certainly benefits the performance of joint learning of multiple modalities, a phenomenon that has also been reported in Salakhutdinov, Tenenbaum, and Torralba (2013). In term of Precision@50, the mPGBN also outperforms mPFA, which performs better than the other multimodal approaches.

Conclusion

We propose a multimodal Poisson gamma belief network (mPGBN) that couples the latent representations of different modalities at multiple hidden layers, extracting the latent features of different modalities at multiple levels of abstraction. The mPGBN infers highly interpretable latent network structure from a collection of image-text pairs, and shows its power in missing modality imputation by both successfully inferring highly relevant tags given an image, and retrieving closely related images given the tags. Quantitative results on a widely used benchmark dataset further demonstrate that the mPGBN achieves state-of-the-art performance on unsupervisedly extracting latent features from multimodal data.

Acknowledgments

Bo Chen thanks the support of the Thousand Young Talent Program of China, NSFC (61771361), and NDPR-9140A07010115DZ01019.

References

Bastan, M.; Cam, H.; Gudukbay, U.; and Ulusoy, O. 2010. Bilvideo-7: an mpeg-7-compatible video indexing and retrieval system. *IEEE MultiMedia* 17(3).

Blei, D. M., and Jordan, M. I. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 127–134. ACM.

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

Bosch, A.; Zisserman, A.; and Munoz, X. 2007. Image classification using random forests and ferns. In *IEEE International Conference on Computer Vision*, 1–8.

Cong, Y.; Chen, B.; Liu, H.; and Zhou, M. 2017. Deep latent dirichlet allocation with topic-layer-adaptive stochastic gradient riemannian mcmc. In *Proceedings of the 34th international conference on machine learning*, 864–873.

Guillaumin, M.; Verbeek, J.; and Schmid, C. 2010. Multimodal semi-supervised learning for image classification. In *Computer Vision and Pattern Recognition*, 902–909.

Huiskes, M. J., and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM interna-*

tional conference on Multimedia information retrieval, 39–43. ACM.

Manjunath, B. S.; Ohm, J.-R.; Vasudevan, V. V.; and Yamada, A. 2001. Color and texture descriptors. *IEEE Transactions on circuits and systems for video technology* 11(6):703–715.

Mcauliffe, J. D., and Blei, D. M. 2008. Supervised topic models. *Advances in Neural Information Processing Systems* 121–128.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning*, 689–696.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision* 42(3):145–175.

Putthividhy, D.; Attias, H. T.; and Nagarajan, S. S. 2010. Topic regression multi-modal latent dirichlet allocation for image annotation. In *Computer Vision and Pattern Recognition*, 3408–3415.

Salakhutdinov, R.; Tenenbaum, J. B.; and Torralba, A. 2013. Learning with hierarchical-deep models. *IEEE transactions on pattern analysis and machine intelligence* 35(8):1958–1971.

Sohn, K.; Shang, W.; and Lee, H. 2014. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, 2141–2149.

Srivastava, N., and Salakhutdinov, R. 2012a. Learning representations for multimodal data with deep belief nets. In *International conference on machine learning workshop*.

Srivastava, N., and Salakhutdinov, R. 2012b. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, 2222–2230.

Vedaldi, A., and Fulkerson, B. 2010. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM international conference on Multimedia*, 1469–1472. ACM.

Zhou, M., and Carin, L. 2015. Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* 37(2):307–320.

Zhou, M.; Cong, Y.; and Chen, B. 2015. The poisson gamma belief network. In *Advances in Neural Information Processing Systems*, 3043–3051.

Zhou, M.; Cong, Y.; and Chen, B. 2016. Augmentable gamma belief networks. *Journal of Machine Learning Research* 17(163):1–44.