Multimodal Weibull Variational Autoencoder for Jointly Modeling Image-Text Data

Chaojie Wang, Bo Chen, *Senior Member, IEEE*, Sucheng Xiao, Zhengjue Wang, Hao Zhang, Penghui Wang, Ning Han, and Mingyuan Zhou

***Abstract*—For multimodal representation learning, traditional black-box approaches often fall short of extracting interpretable multilayer hidden structures, which contribute to visualize the connections between different modalities at multiple semantic levels. To extract interpretable multimodal latent representations and visualize the hierarchial semantic relationships between different modalities, based on deep topic models, we develop a novel multimodal Poisson gamma belief network (mPGBN) that tightly couples the observations of different modalities via imposing sparse connections between their modality-specific hidden layers. To alleviate the time-consuming Gibbs sampler adopted by traditional topic models in the testing stage, we construct a Weibull-based variational inference network (encoder) to directly map the observations to their latent representations, and further combine it with the mPGBN (decoder), resulting in a novel multimodal Weibull variational autoencoder (MWVAE), which is fast in out-of-sample prediction and can handle large-scale multimodal datasets. Qualitative evaluations on bimodal data consisting of image-text pairs show that the developed MWVAE can successfully extract expressive multimodal latent representations for downstream tasks like missing modality imputation and multimodal retrieval. Further extensive quantitative results demonstrate that both MWVAE and its supervised extension sMWVAE achieve state-of-the-art performance on various multimodal benchmarks.**

***Index Terms*—Bayesian inference, deep topic model, multimodal representation learning, variational autoencoder (VAE).**

Manuscript received June 14, 2020; revised January 11, 2021 and March 15, 2021; accepted March 23, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61771361 and Grant 61701379; in part by the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) under Grant B18039; and in part by the Thousand Young Talent Program of China. This article was recommended by Associate Editor Y. Jin. *(Corresponding authors: Penghui Wang; Ning Han; Bo Chen.)*

Chaojie Wang, Bo Chen, Zhengjue Wang, and Penghui Wang are with the National Lab of Radar Signal Processing, Collaborative Innovation Center of Information Sensing and Understanding, Xidian University, Xi'an 710071, China (e-mail: xd_silly@163.com; bchen@mail.xidian.edu.cn; zhengjuewang@163.com; wangpenghui@mail.xidian.du.cn).

Sucheng Xiao is with the Security Department, ByteDance, Guangzhou 518000, China (e-mail: xiaosc9502@163.com).

Hao Zhang is with the Department of Population Health Sciences, Weill Cornell Medicine, NY 14853 USA (e-mail: zhanghao_xidian@163.com).

Ning Han is with Research Room 7, Institute of Mechanical Technology, Xi'an 710071, China (e-mail: haning1103@163.com).

Mingyuan Zhou is with McCombs School of Business, The University of Texas at Austin, TX 78712 USA (e-mail: mingyuan.zhou@mccombs.utexas.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3070881.

Digital Object Identifier 10.1109/TCYB.2021.3070881

I. Introduction

DATA in the real world usually come through various input channels, typically exhibiting multiple modalities that carry different formulations of information. Although each modality is often characterized with extremely distinct statistical properties, the semantic content of any modality is unlikely to be independent of the others. For instance, images are often associated with annotations or captions (e.g., user tags or subtitles), and videos contain both visual and audio signals. Aiming at combing the information of co-occurrence modalities, multimodal data modeling is increasingly attracting attention in many fields especially computer version (CV) [1]–[3], neural language processing (NLP), and computer–human interaction [4], [5].

To exploit the connections between different data modalities, significant progress has been made in the field of multimodal representation learning. One of the leading approaches could be based on probabilistic topic models, specifically latent Dirichlet allocation (LDA) [6] and other more sophisticated variations [7]–[10]. Through constructing a probabilistic model over integer bag-of-word (BoW) representations, topic models provide a meaningful semantic latent representation for each document via inferring its document-topic proportions, which can be naturally applied to model images represented as visual-word vectors [11]. Benefitting from the flexibility of LDA, several multimodal variants, such as correspondence LDA (corr-LDA) [12] and multimodal LDA [13], have been proposed recently to model the joint generative process of multimodal data, through heuristically constructing the relationships between modality-specific topics. Moreover, the class label can also be regarded as an additional modality and embedded into LDA, which can significantly improve the discriminative power of the inferred latent representations [11], [14]. One appealing characteristic of these approaches based on topic modeling is that the task of extracting multimodal latent representations can be easily formulated as a probabilistic inference problem, which can be further solved with routine procedures [15]. Although achieving appealing performance, these multimodal topic models are still limited by their shallow structures that can only explore the connections between different modalities at a shallow semantic level.

Another popular approach of multimodal representation learning is based on the distributed representations modeled by artificial neurons. A common strategy is to construct

a deep neural network (DNN) for each data modality, and then share some specific semantic hidden layers of these modality-specific networks. For instance, based on the restricted Boltzmann machine (RBM), several multimodal approaches [16]–[18] are typically developed via sharing an RBM as their top hidden layers and have achieved great success in jointly modeling pairs of images and text annotations. Recently, with the improvement of calculations, deep learning methods, including convolutional neural networks (CNNS) [19], recurrent neural networks (RNNs) [20], and self-attention networks (SANs) [21] have been widely used for modality-specific tasks and also achieved promising results. However, there is still a challenge for these deep learning approaches to interpret or even visualize the relationships between their modality-specific hidden layers, under conventional multimodal representation learning settings.

To alleviate this issue and take advantage of both aforementioned approaches, we consider constructing a novel multimodal probabilistic model based on deep topic models [8], [22]–[24], which can not only provide multilayer latent document representations in an unsupervised manner but also be easily interpreted for their intuitive top-down network structures. Due to the fact that these deep topic models are still constrained by sophisticated inference algorithms and require a large number of iterations to infer the document-topic proportions, there is a recent trend to construct an inference network (encoder) directly mapping the observations to their latent representations, which can be jointly optimized with the probabilistic generative model (decoder) via minimizing the negative evidence lower bound (ELBO), resulting in a variational autoencoder (VAE) [25]. Moreover, there is also a trend to extend VAE-based methods to multimodal fields, such as multimodal factorization model [26] for exploring intramodal and cross-modal interactions for prediction, and multimodal VAE [27] for solving the multimodal inference problem. However, considering that most existing VAEs [28], [29] still heavily rely on Gaussian latent variables, which often fail to well approximate the posteriors of the skewed, sparse, and non-negative latent representations, how to construct an effective VAE framework for the proposed probabilistic multimodal model remains to be carefully investigated.

In this article, we first develop a novel multimodal Poisson gamma belief network (mPGBN) based on a deep topic model, specifically PGBN [8], and then construct a non-Gaussian multimodal VAE, making our model both scalable and fast in out-of-sample prediction. The contributions of this article are as follows.

1) A novel multimodal deep topic model named mPGBN is proposed, which tightly couples the image-text topics across all hidden layers and provides easily interpretable hierarchical semantic representations.
2) To make the mPGBN more flexible, we propose adaptive normalization to handle variable input scales and extend different link functions to fit multiple modalities characterized with distinct statistical properties.
3) Moving beyond Gaussian reparameterization, we construct a Weibull-based multimodal inference network (encoder) to approximate the analytic posteriors provided by the mPGBN (decoder), resulting in a novel multimodal Weibull VAE (MWVAE).
4) Benefitting from the extensibility of MWVAE, side information, like the label of the image-text pair, can be interpreted to generalize MWVAE to fit various downstream multimodal tasks and further improve the performance.
5) A hybrid MCMC/VAE inference method is developed for MWVAE to handle large-scale datasets, and experimental results demonstrate that our models can achieve state-of-the-art performance on various benchmarks.

Note that the mPGBN presented here first appeared in Wang *et al.* [30] and we have unified related materials in our conference publication. Moving beyond the mPGBN, we develop a novel Weibull-based multimodal VAE extension, equipped with a hybrid MCMC/VAE inference method in this version, and further investigate how to integrate global image features into our models. To obtain more discriminate multimodal representations, a supervised variant sMWVAE is developed, which balances the generative and discriminative aspects in the loss function via introducing a regularization hyperparameter.

The remainder of this article is organized as follows. Section II overviews some related works to demonstrate the differences and advantages of our models. Section III introduces the preliminary of PGBN [8] and explains the structure of mPGBN equipped with extension techniques. Section IV illustrates the VAE framework based on mPGBN, specifically MWVAE, and the corresponding inference details. Section V reports a series of experimental results on both qualitative and quantitative aspects to evaluate our models.

II. Related Work

Learning a multimodal representation across modalities and predicting missing modality conditioned on the others are two key challenges in the field of multimodal representation learning, where a naive approach could be directly concatenating the data descriptors of different modalities, resulting in a raw high-dimensional feature vector. Although significantly improving the performance of downstream tasks like multimodal classification [3], [31] and retrieval [32], [33], this naive approach has difficulty in accomplishing missing modality imputation and often increases the calculation burden for these tasks due to the explosion of the feature dimension.

As previously mentioned, one appealing approach of modeling multimodal data is to explore the multimodal extensions of topic models. For instance, aiming at discovering the relationships between the images and their corresponding annotated tags, corr-LDA is developed via constructing one-to-one mapping between modality-specific topics [12]. Similarly, multimodal LDA directly relates the topics of different modalities with a regression module [13]. Besides annotated tags, supervised LDA (sLDA) [11], [14] introduces the label embedding into LDA and further improves the discriminative power of the inferred multimodal representations. Although there have been a lot of principled solutions, such as variational inference (VI) or MCMC sampling, to solve the inference

problem of these topic models, the inference procedure is still trivial and computationally expensive. Seriously, in the testing stage, these topic modeling-based multimodal approaches often rely on a large number of iterations to infer their latent representations, making them unfriendly to the request of real-time processing. Moreover, the basic LDA can only capture the shallow semantic information, leading to these variants based on LDA having difficulties in exploring high-level semantics, which motivates us to construct a hierarchical multimodal probabilistic topic model.

Thanks to the stochastic gradient (SG) optimization, some popular deep learning approaches are developed recently and may contribute to address these aforementioned issues. For instance, a variant of deep autoencoder (DAE) is applied to learn multimodal representations for both vision and speech modalities, exhibiting that cross-modality feature learning outperforms only using a single modality [34]. To jointly model image-text pairs, the multimodal deep belief network (mDBN) [16] constructs modality-specific DBNs for different modalities and then combines them via sharing an RBM as their top hidden layers. Further, mDBN is extended to a multimodal deep Boltzmann machine (mDBM) [17] via directly replacing the modality-specific DBNs with DBMs. Recently, a neural autoregressive topic model called a document neural autoregressive distribution estimator (DocNADE) [35] is applied to deal with multimodal data in computer vision, via incorporating the spatial visual-word information with a novel structure. To make the joint representations of image-text pairs more discriminative, a supervised variant (SupDocNADE) [36] is developed and confirms the importance of balancing the generative and discriminative aspects in the loss function. However, limited by nonlinear black-box neural networks, these deep-learning-based multimodal approaches still have difficulty in visualizing network structures and there is no distinct association between different modalities.

In contrast to aforementioned conventional deep-learning-based approaches, the proposed MWVAE has an excellent ability in exploratory multimodal data analysis, benefitting from the integration of the mPGBN, which plays an important decoder role in the MWVAE. Before going into technical details, we intuitively exhibit how an image-text pair is factorized under the developed MWVAE as shown in Fig. 1, where the chosen image topics for generating the image are highly correlated with the keywords of the corresponding text topics.

III. Multimodal Poisson Gamma Belief Network

Based on the deep probabilistic topic models, we construct a novel mPGBN that tightly couples different modalities at multiple levels of abstraction and can easily visualize the generative process of multimodal input. Below we first briefly review the preliminaries of PGBN, which plays a building block role in our models and then explain the novel mPGBN equipped with corresponding extension techniques in detail.

A. Poisson Gamma Belief Network

In this part, we first briefly review the PGBN [8], which can infer multilayer latent representations from a group of discrete observations. Specifically, representing a set of N documents $X = \{x_n\}_{n=1}^N$ as BoW vectors, each sample could be a high-dimensional sparse count vector $x_n \in \mathbb{Z}^{K^{(0)}}$, where $\mathbb{Z} = \{0, 1, \dots, \}$ and $K^{(0)}$ denotes the vocabulary length. Factorizing the observed multivariate count vectors x_n under the Poisson likelihood, the generative model of the PGBN with L hidden layers can be formulated as

$$\begin{aligned}
&\boldsymbol{\theta}_n^{(L)} \sim \mathrm{Gam}\Big(\boldsymbol{r},\, 1/c_n^{(L+1)}\Big)\\
&\cdots\\
&\boldsymbol{\theta}_n^{(l)} \sim \mathrm{Gam}\Big(\boldsymbol{\Phi}^{(l+1)}\boldsymbol{\theta}_n^{(l+1)},\, 1/c_n^{(l+1)}\Big)\\
&\cdots\\
&\boldsymbol{x_n} \sim \mathrm{Pois}\Big(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_n^{(1)}\Big),\, \boldsymbol{\theta}_n^{(1)} \sim \mathrm{Gam}\Big(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_n^{(2)},\, 1/c_n^{(2)}\Big) \quad (1)
\end{aligned}$$

where the superscript denotes the layer index. For each hidden layer l, the latent representation $\boldsymbol{\theta}_n^{(l)} \in \mathbb{R}_+^{K^{(l)}}$ can be factorized into the product of the factor loading $\boldsymbol{\Phi}^{(l+1)} \in \mathbb{R}_+^{K^{(l)} \times K^{(l+1)}}$ and the latent representation $\boldsymbol{\theta}_n^{(l+1)} \in \mathbb{R}_+^{K^{(l+1)}}$ in the next layer under the gamma distribution, where $\mathbb{R}_+ = \{x : x \geq 0\}$. Specifically, the top layer's hidden units $\boldsymbol{\theta}_n^{(L)}$ share the same gamma shape parameters $\boldsymbol{r} = (r_1, \dots, r_{K^{(L)}})'$ and $\{\boldsymbol{\theta}_n^{(l)}\}_{l=1}^L$ at different layers can be also referred as topic proportions.

To make the PGBN both scale identifiable and inference convenient, each column of $\boldsymbol{\Phi}^{(l)} \in \mathbb{R}_+^{K^{(l-1)} \times K^{(l)}}$ is restricted to have the L_1 norm through introducing a Dirichlet prior as

$$\boldsymbol{\phi}_k^{(l)} \sim \mathrm{Dir}\Big(\eta^{(l)}, \dots, \eta^{(l)}\Big) \quad (2)$$

where $\boldsymbol{\phi}_k^{(l)} \in \mathbb{R}_+^{K^{(l-1)}}$ denotes the kth column of $\boldsymbol{\Phi}^{(l)}$. And the gamma scale parameters $\{1/c_n^{(l)}\}_{2,L+1}$ satisfy

$$c_n^{(l)} \sim \mathrm{Gam}(e_0, 1/f_0) \quad (3)$$

for $l \in \{2, \dots, L+1\}$. Note that the single-layer version of PGBN reduces to Poisson factor analysis (PFA) [7].

1) Hierarchical Semantic Topic: One of the most attractive properties is that the PGBN can provide a principled probabilistic interpretation for the extracted hierarchical semantic topics, denoted as $\{\boldsymbol{\Phi}^{(l)}\}_{l=1}^L$ in (1). For the bottom layer, each document $\boldsymbol{x}_n$ can be seen as a random mixture over $K^{(1)}$ topics like LDA [6], since $\mathbb{E}[\boldsymbol{x}_n|\boldsymbol{\Phi}^{(1)}, \boldsymbol{\theta}_n^{(1)}] = \boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_n^{(1)}$. For higher layers, successive factorization under the gamma distribution results in

$$\mathbb{E}\Big[\boldsymbol{x}_n|\boldsymbol{\theta}_n^{(l)}, \Big\{\boldsymbol{\Phi}^{(t)}, c_n^{(t)}\Big\}_{t=1}^{l}\Big] = \Big[\prod_{t=1}^{l}\boldsymbol{\Phi}^{(t)}\Big]\frac{\boldsymbol{\theta}_n^{(l)}}{\prod_{t=2}^{l} c_n^{(t)}} \quad (4)$$

which makes it easy to examine the nodes of hidden layers via projecting them to the bottom layer. In other words, the semantic topics at layer l can be visualized according to their projections calculated as $\{[\prod_{t=1}^{l-1}\boldsymbol{\Phi}^{(t)}]\boldsymbol{\phi}_k^{(l)}\}_{k=1}^{K^{(l)}}$, and thus each document can also be seen as a random mixture over $K^{(l)}$ topics with $\boldsymbol{\theta}_n^{(l)}$ being the corresponding topic proportions at layer l. Moreover, the topics learned by PGBN tend to be more specific at lower layers and those at higher layers are

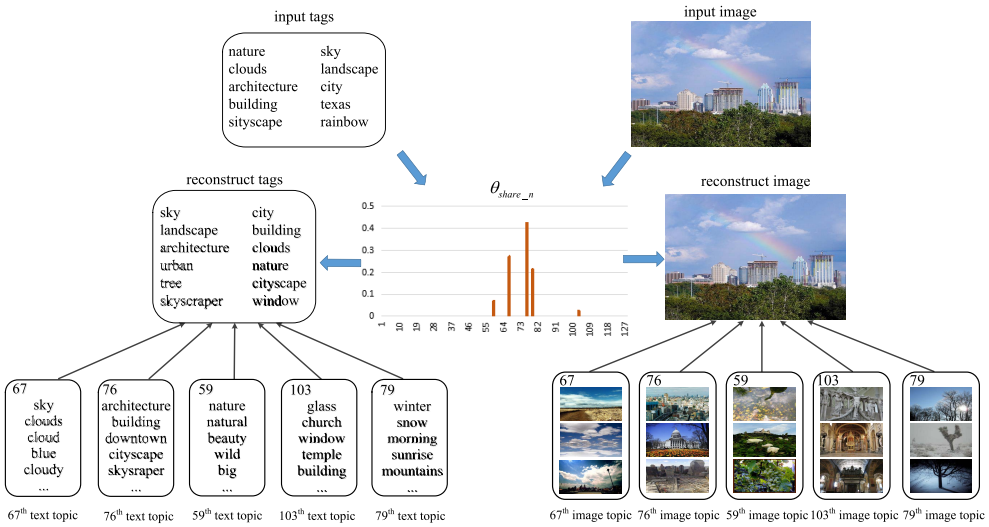


Fig. 1. Generate process visualization of input image-text pair learned by MWVAE trained on the MIR-Flickr dataset. Taking the image-text pair as input, MWVAE can extract a shared latent multimodal representation $\boldsymbol{\theta}_{\text{share}\_n}$ with the inference network (encoder) and then provide reconstructions with mPGBN (decoder), where the exhibited modality-specific topics are top-k largest factors and the detailed visualization technique has been described in Section V-C.

more general as shown in Fig. 2, which is quite similar to the underlying thought of deep learning.

B. Multimodal Poisson Gamma Belief Network

Below we explain the technical details of the developed mPGBN, taking that both the image and text inputs are count vectors as our basic case.

1) Model Architecture: Assuming that both image and text modality-specific descriptors are BoW count vectors, we first construct the mPGBN via sharing the multilayer hidden variables of two modality-specific PGBNs, except for their connection weights between the visible layer and first hidden layer. Then the generative model of the mPGBN can be expressed as

$$\begin{aligned}
&\boldsymbol{\theta}^{(L)}_{\text{share}_n} \sim \text{Gam}\Big(\boldsymbol{r}_{\text{share}}, 1/c^{(L+1)}_{\text{share}_n}\Big)\\
&\cdots\\
&\boldsymbol{\theta}^{(l)}_{\text{share}_n} \sim \text{Gam}\Big(\boldsymbol{\Phi}^{(l+1)}_{\text{share}}\boldsymbol{\theta}^{(l+1)}_{\text{share}_n}, 1/c^{(l+1)}_{\text{share}_n}\Big)\\
&\cdots\\
&\boldsymbol{x}_{\text{img}_n} \sim \text{Pois}\Big(\boldsymbol{\Phi}^{(1)}_{\text{img}}\boldsymbol{\theta}^{(1)}_{\text{share}_n}\Big), \boldsymbol{x}_{\text{txt}_n} \sim \text{Pois}\Big(\boldsymbol{\Phi}^{(1)}_{\text{txt}}\boldsymbol{\theta}^{(1)}_{\text{share}_n}\Big)
\end{aligned} \tag{5}$$

where the subscript indicates the abbreviation of each specific modality (or shared by both). Note that both modality-specific input vectors $\boldsymbol{x}_{\text{img}\_n}$ and $\boldsymbol{x}_{\text{txt}\_n}$ are first projected into a common semantic representation $\boldsymbol{\theta}^{(1)}_{\text{share}\_n} \in \mathbb{R}^{K^{(1)}}_{+}$, and then each latent representation $\boldsymbol{\theta}^{(l)}_{\text{share}\_n} \in \mathbb{R}^{K^{(l)}}_{+}$ of layer l is further successively factorized into the product of the factor loading $\boldsymbol{\Phi}^{(l+1)}_{\text{share}\_n} \in \mathbb{R}^{K^{(l)}\times K^{(l+1)}}_{+}$ and the latent representation $\boldsymbol{\theta}^{(l+1)}_{\text{share}\_n} \in \mathbb{R}^{K^{(l+1)}}_{+}$ of the next layer under the gamma distribution. Moreover, the same as $\{\boldsymbol{\Phi}^{(t)}_{\text{share}}\}^{L}_{t=2}$, each column of $\boldsymbol{\Phi}^{(1)}_{\text{img}}$ or $\boldsymbol{\Phi}^{(1)}_{\text{txt}}$ is restricted to have a simplex constrain.

The underlying intuition behind sharing modality-specific representations at multiple layers is that even though different modalities tend to exhibit distinct statistical properties, there could be still strong correlations between their latent representations at multiple semantic levels. Specifically, the image and text descriptors collected from the same image-text pair can be seen as two different exhibitions of the same semantic meaning. For instance, the image of a tiger shares relative shallow semantic meanings with the word "tiger," high-level semantics with the word "big cat," and even a higher abstraction level with the word "carnivore." Similar network structures have been proven efficient in DeepDocNADE [36], but there are still a lot of differences between the proposed mPGBN and these deep-learning-based methods. As shown in Fig. 2, the multimodal generative model mPGBN, which plays a decoder role in the following proposed MWVAE, could capture an interpretable latent structure at different semantic levels, contributing to understanding both the semantic meanings of multilayer latent representations and the correlations between different modalities reflected at the same level of abstraction.

2) Upward–Downward Gibbs Sampler: Thanks to the data augmentation technique [37], the proposed mPGBN provides analytic posteriors and can be further trained with an upward–downward Gibbs sampler [8]. For each iteration, the Gibbs sampler first upward samples factor loadings $\{\boldsymbol{\Phi}^{(l)}\}^{L}_{l=1}$ starting from the bottom visible layer, then downward samples topic proportions $\{\boldsymbol{\theta}^{(l)}_{n}\}^{L}_{l=1}$ starting from the top hidden layer. Distinct from the PGBN, the developed mPGBN adopts multisource data augmentations at the first hidden layer and the corresponding update equation of $\boldsymbol{\theta}^{(1)}_{\text{share}\_n}$ can be formulated as

$$\begin{aligned}
\Big(\boldsymbol{\theta}^{(1)}_{\text{share}_n}|-\Big) \sim \text{Gam}\Big(&\boldsymbol{m}^{(1)(2)}_{\text{img}_n} + \boldsymbol{m}^{(1)(2)}_{\text{txt}_n}\\
&+ \boldsymbol{\Phi}^{(2)}_{\text{share}}\boldsymbol{\theta}^{(2)}_{\text{share}_n} 1/\Big[2 + c^{(2)}_{n}\Big]\Big)
\end{aligned} \tag{6}$$

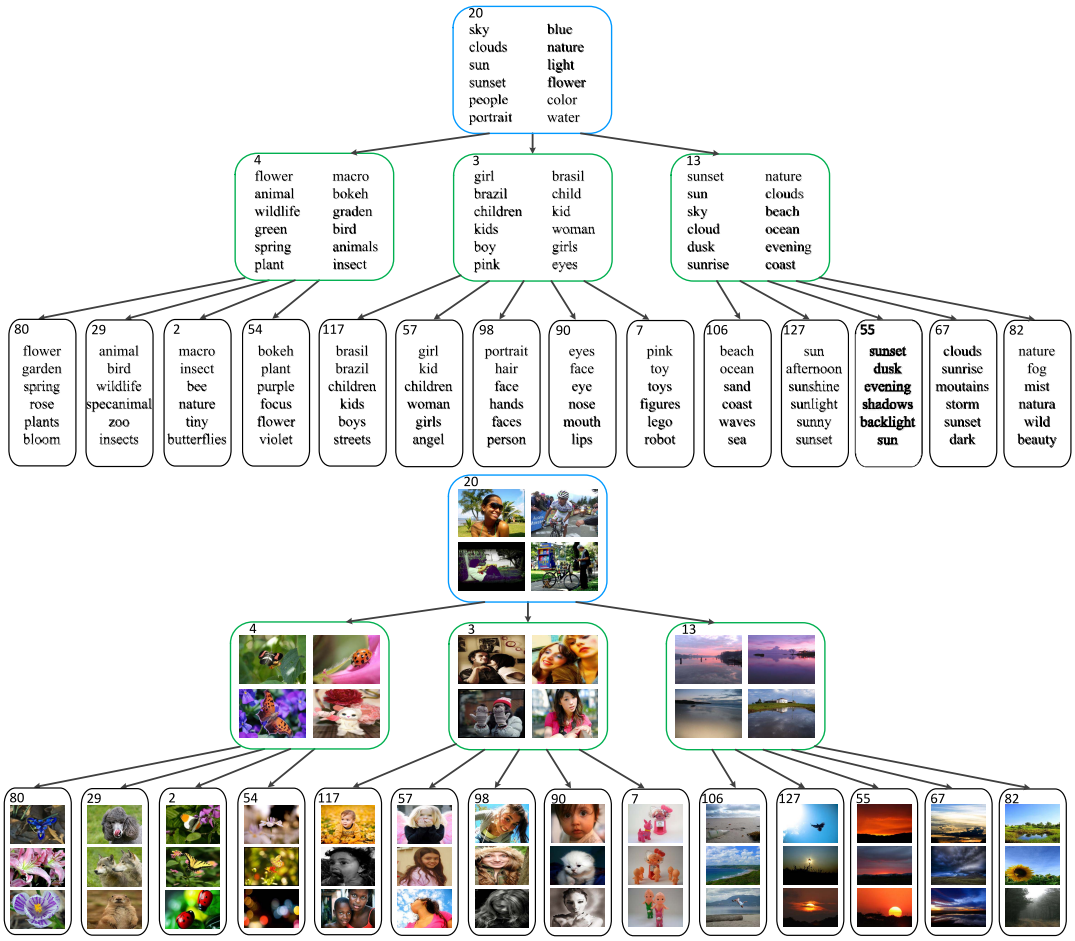


Fig. 2. Two [14, 3, 1] modality-specific trees taken from the MWVAE trained on the MIR-Flickr dataset. Note each modality-specific tree has included all the lower layer nodes (directly or indirectly) linked with non-negligible weights starting from the 20th node at the top hidden layer, where the connection from node k at layer l to node k' at layer $l-1$ satisfies the constraint of $\Phi_{k'k}^{(l)} > 10/K^{(l-1)}$. For the text-specific tree, six keywords of the corresponding topic are displayed inside the text box at layer one and 12 keywords at other layers. As for the image-specific tree, top-k relevant images retrieved from MIR-Flickr are exhibited inside the corresponding image box, via measuring their Euclidean distances to the learned image theme features.

where $\boldsymbol{m}_{\text{img}\_n}^{(1)(2)}$ and $\boldsymbol{m}_{\text{txt}\_n}^{(1)(2)}$ are both augmented count matrix, which are sampled from their corresponding modalities, but will both directly influence the gamma shape parameter of the first layer's latent representation, specifically $\boldsymbol{\theta}_{\text{share}\_n}^{(1)}$.

Note that different from the layerwise training adopted in multimodal DBN [16], the upward–downward Gibbs sampler can train the multilayer mPGBN as an integration and provide the corresponding iterative upward–downward information propagation. Further, benefit from training the whole network jointly, the mPGBN can not only tightly couple image themes and text topics learned from multimodal input but also explore the relationships of semantic meanings between different modality-specific hidden layers.

C. Model Extension Techniques

Aiming at making the proposed mPGBN adaptive to both variable input scales and multiple modalities with distinct statistical properties, we propose flexible model extension techniques as follows.

1) Link Functions: In addition to directly fitting high-dimensional count observations with the Poisson likelihood, we equip the proposed mPGBN with a set of link functions [37] to fit other types of modality-specific inputs.

Supposing the observations are high-dimensional binary vectors $\boldsymbol{b}_n \in \{0,1\}^{K^{(0)}}$, we can adopt the Bernoulli-Poisson link formulated as

$$\boldsymbol{b}_n = 1(\boldsymbol{x}_n \geq 0),\ \boldsymbol{x}_n \sim \text{Pois}\left(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_n^{(1)}\right). \tag{7}$$

If the observations are high-dimensional non-negative real-value vectors $\boldsymbol{y}_n \in \mathbb{R}_+^{K^{(0)}}$, we can factorize them with the Poisson randomized gamma link specifically as

$$\boldsymbol{y}_n \sim \text{Gam}(\boldsymbol{x}_n, 1/a_n), \boldsymbol{x}_n \sim \text{Pois}\left(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_n^{(1)}\right). \tag{8}$$

Taking advantage of link functions mentioned above, the text-specific PGBN can directly fit integer BoW vectors or model binary annotated tags with the Bernoulli-Poisson link formulated in (7), whereas the image-specific PGBN can use the Poisson randomized gamma link shown in (8) to fit positive image features or model handcraft count vectors like visual words extracted from images.

2) Adaptive Normalization: The original mPGBN formulated in (5) has a potential issue that the different modality-specific inputs may have various data scales. To this end, we

propose a novel adaptive normalization technique to modify the mPGBN as

$$\boldsymbol{\theta}^{(1)}_{\text{img\_}n} = \delta_{\text{img\_}n}\boldsymbol{\theta}^{(1)}_{\text{share\_}n}, \boldsymbol{\theta}^{(1)}_{\text{txt\_}n} = \delta_{\text{txt\_}n}\boldsymbol{\theta}^{(1)}_{\text{share\_}n}$$
$$\boldsymbol{x}_{\text{img\_}n} \sim \text{Pois}\Big(\boldsymbol{\Phi}^{(1)}_{\text{img}}\boldsymbol{\theta}^{(1)}_{\text{img\_}n}\Big), \boldsymbol{x}_{\text{txt\_}n} \sim \text{Pois}\Big(\boldsymbol{\Phi}^{(1)}_{\text{txt}}\boldsymbol{\theta}^{(1)}_{\text{txt\_}n}\Big). \quad (9)$$

Benefit the adaptive normalization technique, the first latent representations of both modalities, specifically $\boldsymbol{\theta}^{(1)}_{\text{img\_}n}$ and $\boldsymbol{\theta}^{(1)}_{\text{txt\_}n}$, only share their gamma shape parameters but have their own adaptive scale parameters to suit different input scales.

Equipped with these useful extension techniques, the proposed mPGBN can improve the expressivity of hierarchial latent representations, outperforming conventional multimodal topic models that only construct connections between different modalities at a single semantic level.

IV. Multimodal Weibull Variational Autoencoder

Although the analytic posteriors of mPGBN provide efficient inference with Gibbs sampler that can be further accelerated with GPU [38], mPGBN is still limited by the following disadvantages.

1) Characterized by a top-down generative structure, it relies on time-consuming batch sampling when inferring the latent representations.
2) Restricted by the Gibbs sampler, it is not easy to plug in extra side information to extend mPGBN, such as image labels.
3) To handle the increasing amount and complexity of data, a scalable inference algorithm is required for mPGBN.

To alleviate these issues, we combine the mPGBN (decoder) with a Weibull-based VI network (encoder), resulting in a novel MWVAE. Then we develop a corresponding hybrid stochastic-gradient-MCMC/autoencoding VI algorithm for MWVAE, which learns the global parameters of the mPGBN jointly with those of the inference network.

A. Weibull Variational Posterior

The most important issue in constructing a VAE-like model is the choice of latent distributions, and most existing latent Gaussian-based VAEs [28], [29] have achieved a great success benefiting from the characteristics of the Gaussian reparameterization. However, these Gaussian distributed latent variables have difficulty in modeling gamma distributed ones, which are often sparse, non-negative, and skewed. Moving beyond Gaussian-based VAEs, we choose the Weibull reparameterization to approximate the gamma distributed conditional posteriors of $\{\boldsymbol{\theta}^{(l)}_{\text{share\_}n}\}^{L}_{l=1}$ considering the following advantages.

1) Similar PDF With Gamma Distribution: The Weibull distribution owns similar probability density functions (PDFs) with a gamma one, which makes it flexible to model sparse and non-negative latent representations

$$\text{Weibull PDF: } P(x|k,\lambda) = \frac{k}{\lambda^k}x^{k-1}e^{(x/\lambda)^k}$$
$$\text{Gamma PDF: } P(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}. \quad (10)$$

2) Easily Reparameterization: The latent variable $x \sim \text{Weibull}(k,\lambda)$ can be easily reparameterized as

$$x = \lambda(-\ln(1-\varepsilon))^{1/k}, \quad \varepsilon \sim \text{Uniform}(0,1) \quad (11)$$

leading to a similar gradient calculation with the Gaussian reparameterization.

3) Analytic KL-Divergence: Moverover, the KL-divergence between the Weibull and gamma distributions has an analytic expression formulated as

$$\text{KL}(\text{Weibull}(k,\lambda)||\text{Gamma}(\alpha,\beta)) = -\alpha\ln\lambda + \frac{\gamma\alpha}{k} + \ln k + \beta\lambda\Gamma\Big(1+\frac{1}{k}\Big) - \gamma - 1 - \alpha\ln\beta + \ln\Gamma(\alpha). \quad (12)$$

B. Multimodal Weibull Variational Autoencoder

Taking advantage of the Weibull distribution, we construct a novel multimodal VAE framework based on the proposed mPGBN, so-called mutlimodal Weibull VAE (MWVAE), where the mPGBN plays a role of the generative model (decoder) equipped with a corresponding Weibull-based inference network as shown in Fig. 3. Distinct from the common inference network of a usual VAE [25], which adopts a pure bottom-up structure ignoring the impact of the prior and only interacts with the generative model via the ELBO, MWVAE constructs an upward–downward structure inspired by the upward–downward information propagation in Gibbs sampler of mPGBN. Note that this upward–downward structure has not only an upward information propagation through hierarchical semantic latent representations $\{\boldsymbol{h}^{(l)}_{\text{share\_}n}\}^{L}_{l=1}$ but also a downward one $\{\boldsymbol{\Phi}^{(l+1)}_{\text{share}}\boldsymbol{\theta}^{(l+1)}_{\text{share\_}n}\}^{L-1}_{l=1}$ acting as the prior from the higher layer, which naturally meets the posteriors of $\{\boldsymbol{\theta}^{(l)}_{\text{share\_}n}\}^{L}_{l=1}$ as shown in (6). Specifically, the inference network of MWVAE can be formulated as

$$\prod_{l=1}^{L-1} q\Big(\boldsymbol{\theta}^{(l)}_{\text{share\_}n}|\boldsymbol{\Phi}^{(l+1)}_{\text{share}}, \boldsymbol{\theta}^{(l+1)}_{\text{share\_}n}, \boldsymbol{h}^{(l)}_{\text{share\_}n}\Big)q\Big(\boldsymbol{\theta}^{(L)}_{\text{share\_}n}|-\Big) \quad (13)$$

where the Weibull distribution is introduced to approximate the gamma distributed conditional posterior of each latent representation $\boldsymbol{\theta}^{(l)}_{\text{share\_}n}$ formulated as

$$q\Big(\boldsymbol{\theta}^{(l)}_{\text{share\_}n}|-\Big) = \text{Weibull}\Big(\boldsymbol{\Phi}^{(l+1)}_{\text{share}}\boldsymbol{\theta}^{(l+1)}_{\text{share\_}n} + \boldsymbol{k}^{(l)}_{\text{share\_}n}, \boldsymbol{\lambda}^{(l)}_{\text{share\_}n}\Big). \quad (14)$$

Note that $\boldsymbol{k}^{(l)}_{\text{share\_}n} \in \mathbb{R}^{K^{(l)}}_{+}$ and $\boldsymbol{\lambda}^{(l)}_{\text{share\_}n} \in \mathbb{R}^{K^{(l)}}_{+}$ are both Weibull parameters of the lth shared hidden layer to reparameterize $\boldsymbol{\theta}^{(l)}_{\text{share\_}n}$ with (11), and deterministically transformed from the observed image-text pairs using the neural networks expressed as

$$\boldsymbol{h}^{(l)}_{\text{share\_}n} = \ln\Big(1+\exp\Big(\boldsymbol{W}^{(l)}_{h}\boldsymbol{h}^{(l-1)}_{\text{share\_}n} + \boldsymbol{b}^{(l)}_{h}\Big)\Big)$$
$$\boldsymbol{k}^{(l)}_{\text{share\_}n} = \ln\Big(1+\exp\Big(\boldsymbol{W}^{(l)}_{k}\boldsymbol{h}^{(l)}_{\text{share\_}n} + \boldsymbol{b}^{(l)}_{k}\Big)\Big)$$
$$\boldsymbol{\lambda}^{(l)}_{\text{share\_}n} = \ln\Big(1+\exp\Big(\boldsymbol{W}^{(l)}_{\lambda}\boldsymbol{h}^{(l)}_{\text{share\_}n} + \boldsymbol{b}^{(l)}_{\lambda}\Big)\Big) \quad (15)$$

where $\boldsymbol{W}^{(l)}_{h} \in \mathbb{R}^{K^{(l-1)}\times K^{(l)}}$, $\boldsymbol{W}^{(l)}_{k}$, $\boldsymbol{W}^{(l)}_{\lambda} \in \mathbb{R}^{K^{(l)}\times K^{(l)}}$, $\boldsymbol{b}^{(l)}_{h}, \boldsymbol{b}^{(l)}_{k}, \boldsymbol{b}^{(l)}_{\lambda} \in \mathbb{R}^{K^{(l)}}$. For the first hidden layer, we take the

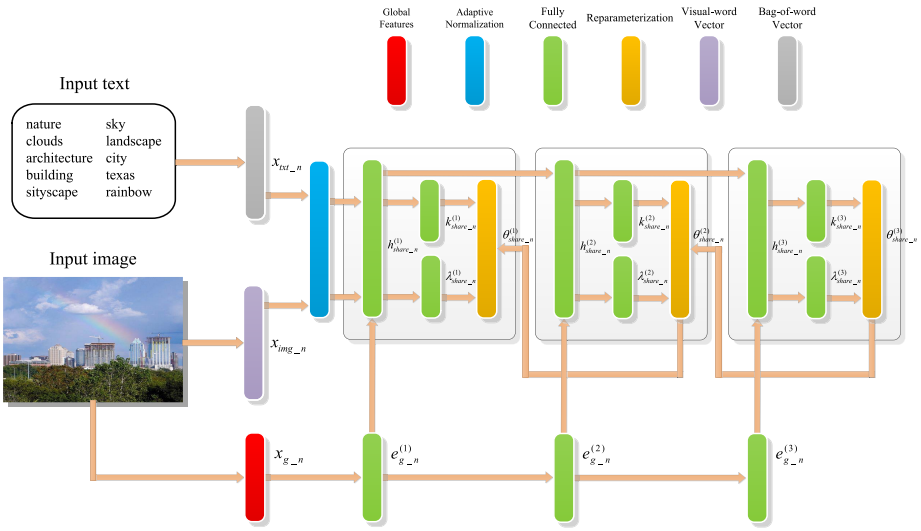


Fig. 3. Detailed structure of a three-layer multimodal Weibull-based inference network (enocder) for the proposed MWVAE, where different colored blocks indicate different network components.

normalized image-text pair as input and specifically denote $\boldsymbol{h}^{(1)}_{\text{share\_}n} = \text{softplus}(\boldsymbol{W}^{(1)}_{\text{img}} \log(1+\boldsymbol{x}_{\text{img\_}n}) + \boldsymbol{W}^{(1)}_{\text{txt}} \log(1+\boldsymbol{x}_{\text{txt\_}n}) + \boldsymbol{b}^{(1)}_h)$ with $\boldsymbol{b}^{(l)}_h \in \mathbb{R}^{K^{(1)}}$, $\boldsymbol{W}^{(1)}_{\text{img}} \in \mathbb{R}^{K^{(0)}_{\text{img}} \times K^{(1)}}$, $\boldsymbol{W}^{(1)}_{\text{txt}} \in \mathbb{R}^{K^{(0)}_{\text{txt}} \times K^{(1)}}$.

Moving beyond the traditional VAE-like models giving point estimations for their latent representations, MWVAE adopts the Weibull reparameterization in the inference network to introduce stochastics into these latent representations and further improves the model performance. Moreover, the additional semantic hidden layers $\{\boldsymbol{h}^{(l)}_{\text{share\_}n}\}^L_{l=1}$ tightly couple different modalities in a hierarchical fashion and have the equivalent of the augmented vectors $\{\boldsymbol{m}^{(l)(l+1)}_{\text{share\_}n}\}^L_{l=1}$ in Gibbs sampler. From another perspective, it is also flexible to incorporate side information into $\{\boldsymbol{h}^{(l)}_{\text{share\_}n}\}^L_{l=1}$, and then propagate it to the latent representations $\{\boldsymbol{\theta}^{(l)}_{\text{share\_}n}\}^L_{l=1}$.

1) Exploiting Global Image Features: Although equipped with flexible link functions described in Section III-C1, the proposed mPGBN still has difficulty in dealing with real-valued global modality-specific embeddings, which have been proven efficient to improve the expressivity of the inferred multimodal latent representations [17]. Moving beyond the constraint of Gibbs sampler, MWVAE provides a potential solution to integrate the global features into latent representations $\{\boldsymbol{\theta}^{(l)}_{\text{share\_}n}\}^L_{l=1}$, benefiting from the flexible VAE-like framework. Then we will describe how to incorporate global image features into the MWVAE, which can be also easily extended to other modalities.

Distinct from SupDocNADE [36], which introduces global image features $\boldsymbol{x}_{g\_n} \in \mathbb{R}^{K^{(0)}_g}$ via directly concatenating it with the visual-word vector $\boldsymbol{x}_{\text{img\_}n}$, we project $\boldsymbol{x}_{g\_n}$ into multilayer representations $\{\boldsymbol{e}^{(l)}_{g\_n}\}^L_{l=1}$ and then integrate them with $\{\boldsymbol{h}^{(l)}_{\text{share\_}n}\}^L_{l=1}$ in a hierarchical fashion. Specifically, the latent representation of global image features can be obtained as

$$\boldsymbol{e}^{(l)}_{g\_n} = \text{softplus}\left(\boldsymbol{W}^{(l)}_e \boldsymbol{e}^{(l-1)}_{g\_n} + \boldsymbol{b}^{(l)}_e\right) \tag{16}$$

where $\boldsymbol{e}^{(0)}_{g\_n} = \boldsymbol{x}_{g\_n}$. Then the multilayer global image representations are interpreted via a linear transformation as

$$\boldsymbol{h}^{(l)}_{\text{share\_}n} = \text{softplus}\left(\boldsymbol{W}^{(l)}_h \boldsymbol{h}^{(l-1)}_{\text{share\_}n} + \boldsymbol{W}^{(l)}_c \boldsymbol{e}^{(l)}_{g\_n} + \boldsymbol{b}^{(l)}_h\right) \tag{17}$$

specifically defining $\boldsymbol{h}^{(1)}_{\text{share\_}n} = \text{softplus}(\boldsymbol{W}^{(1)}_{\text{img}} \log(1+\boldsymbol{x}_{\text{img\_}n}) + \boldsymbol{W}^{(1)}_{\text{txt}} \log(1+\boldsymbol{x}_{\text{txt\_}n}) + \boldsymbol{W}^{(1)}_c \boldsymbol{e}^{(1)}_{g\_n} + \boldsymbol{b}^{(1)}_h)$. The underlying intuition is that the global image features also exhibit variable statistical characteristics at different semantic levels. In the following experiments, both handcraft global image features and deep-learning-based ones like encoding the original image with CNN [19] are taken into consideration to extend our models.

C. Hybrid MCMC/VAE Inference Method

Although having closed-form update equations, the Gibbs sampler for the global parameters of the mPGBN, specifically $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}^{(1)}_{\text{img}}, \boldsymbol{\Phi}^{(1)}_{\text{txt}}, \{\boldsymbol{\Phi}^{(l)}_{\text{share}}\}^L_{l=2}\}$, still requires processing all image-text pairs in each iteration. To alleviate this issue, we adopt TLASGR-MCMC [39] based on the Fisher information matrix (FIM) to update these global parameters, via increasing the sampling efficiency. Specifically, suppose $\boldsymbol{\phi}^{(l)}_k$ is the kth topic in the lth layer of mPGBN with the prior $\boldsymbol{\phi}^{(l)}_k \sim \text{Dirichlet}(\boldsymbol{\eta}^{(l)}_k)$, the sampling of it can be realized as

$$\boldsymbol{\phi}^{(l)new}_k = \left\{ \boldsymbol{\phi}^{(l)}_k + \frac{\varepsilon^{(l)}_i}{\boldsymbol{M}^{(l)}_k} \left[\left(\rho \tilde{\boldsymbol{x}}^{(l)}_{:k\cdot} + \boldsymbol{\eta}^{(l)}\right) - \left(\rho \tilde{\boldsymbol{x}}^{(l)}_{:k\cdot} + \boldsymbol{\eta}^{(l)} K^{(l-1)}\right) \boldsymbol{\phi}^{(l)}_k \right] + N\left(0, \frac{2\varepsilon^{(l)}_i}{\boldsymbol{M}^{(l)}_k} \text{diag}\left(\boldsymbol{\phi}^{(l)}_k\right)\right) \right\}_{\angle} \tag{18}$$

where i denotes the number of mini-batches processed so far. Here, the symbol $\cdot$ in the subscript denotes summing over the corresponding dimension of the minibatch, and the definitions of ρ, $\varepsilon^{(l)}_i$, $\boldsymbol{M}^{(l)}_k$ and $\{\cdot\}_{\angle}$ are analogous to these in TLASGR-MCMC [39] and omitted here for brevity.

Then we combine the mentioned TLASGR-MCMC and the proposed MWVAE into a hybrid MCMC/VAE inference algorithm as shown in Algorithm 1, which updates the inference

Algorithm 1 Hybrid Stochastic-Gradient MCMC and VAE Inference for MWVAE

Set the mini-batch size m and the number of layer L;
Initialize the inference network parameters (encoder) $\boldsymbol{\Omega} = \{\boldsymbol{W}_{\text{img}}^{(1)}, \boldsymbol{W}_{\text{txt}}^{(1)}, \{\boldsymbol{W}_h^{(l)}\}_{l=2}^{L}, \{\boldsymbol{b}_h^{(l)}, \boldsymbol{W}_k^{(l)}, \boldsymbol{b}_k^{(l)}, \boldsymbol{W}_\lambda^{(l)}, \boldsymbol{b}_\lambda^{(l)}\}_{l=1}^{L}\}$ and mPGBN parameters $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}_{\text{img}}^{(1)}, \boldsymbol{\Phi}_{\text{txt}}^{(1)}, \{\boldsymbol{\Phi}_{\text{share}}^{(l)}\}_{l=1}^{L}\}$
for *iteration* $= 1, 2, \cdots$ **do**
  Randomly select a mini-batch of m image-text pairs to form a subset $\boldsymbol{X} = \{\boldsymbol{x}_{\text{img\_}i}, \boldsymbol{x}_{\text{txt\_}i}\}_{i=1}^{m}$;
  Estimate local parameters $\{\delta_{\text{img\_}i}, \delta_{\text{txt\_}i}\}_{i=1}^{m}$;
  Draw random noise $\boldsymbol{\varepsilon} = \{\varepsilon_i^{(l)}\}_{i=1,l=1}^{m,L}$ from uniform distribution;
  Calculate $\nabla_{\boldsymbol{\Omega}} L(\boldsymbol{\Omega}, \boldsymbol{\Phi}; \boldsymbol{X}, \boldsymbol{\varepsilon})$ according to (19) and update $\boldsymbol{\Omega}$;
  Sample $\{\boldsymbol{\theta}_{\text{share\_}i}^{(l)}\}_{i=1,l=1}^{m,L}$ from (14) via $\boldsymbol{\Omega}$ and update $\boldsymbol{\Phi}$ according to (18);
end for

network parameters $\boldsymbol{\Omega}$, and infers the global parameters $\boldsymbol{\Phi}$ of the generative model jointly. Note that the whole proposed structure above can be optimized to maximize the ELBO in unsupervised manner, which can be formulated as

$$\begin{aligned} L_g = &\sum_{n=1}^{N} \mathbb{E}\Big[\ln p\Big(\boldsymbol{x}_{\text{img\_}n} | \boldsymbol{\Phi}_{\text{img}}^{(1)}, \boldsymbol{\theta}_{\text{share\_}n}^{(1)}, \delta_{\text{img\_}n}\Big)\Big] \\ &+ \sum_{n=1}^{N} \mathbb{E}\Big[\ln p\Big(\boldsymbol{x}_{\text{txt\_}n} | \boldsymbol{\Phi}_{\text{txt}}^{(1)}, \boldsymbol{\theta}_{\text{share\_}n}^{(1)}, \delta_{\text{txt\_}n}\Big)\Big] \\ &- \sum_{n=1}^{N} \sum_{l=1}^{L} \mathbb{E}\left[\ln \frac{q\Big(\boldsymbol{\theta}_{\text{share\_}n}^{(l)}\Big)}{p\Big(\boldsymbol{\theta}_{\text{share\_}n}^{(l)} | \boldsymbol{\Phi}_{\text{share}}^{(l+1)}, \boldsymbol{\theta}_{\text{share\_}n}^{(l+1)}\Big)}\right] \end{aligned} \tag{19}$$

where the third term is analytic as shown in (12). Moreover, benefit from the simple reparameterization of the Weibull distribution, the gradient of the first and second terms of the ELBO can be directly calculated [40].

To obtain more discriminative multimodal latent representations, we consider to extend the unsupervised MWVAE into a supervised version, referred as sMWVAE, by introducing label information into the latent representations. Moving beyond directly mapping the top layer's latent representation to label probabilities like SupDocNADE [36], we add a softmax classifier on the concatenation of these latent representations $\{\boldsymbol{\theta}_{\text{share\_}n}^{(l)}, \boldsymbol{h}_{\text{share\_}n}^{(l)}\}_{l=1}^{L}$, aiming at broadcasting label information across all hidden layers. Specifically defining the concatenated vector as $\boldsymbol{\Theta}_n$, the predicted label probabilities $\hat{\boldsymbol{y}}_n \in \mathbb{R}_+^C$, where C denotes the number of classes, can be obtained as $\hat{\boldsymbol{y}}_n = \text{softmax}(\boldsymbol{W}_y \boldsymbol{\Theta}_n + \boldsymbol{b}_y)$. Then the whole loss function of sMWVAE can be formulated as

$$L_s = L_g + \lambda \cdot L_c\big(\boldsymbol{y_n}, \hat{\boldsymbol{y}}_n\big) \tag{20}$$

where λ is treated as a regularization hyperparameter [36] to balance the generative loss L_g defined in (19) and the cross-entropy loss L_c.

V. Experiments and Results

In the experimental section, we investigate the proposed models on both qualitative and quantitative aspects. Specifically, we first evaluate the characteristics of the MWVAE with a series of qualitative tasks, showing that the MWVAE can successfully accomplish both missing modality imputation and multimodal retrieval tasks. Then we test both the performance of MWVAE and sMWVAE on three widely used multimodal datasets to demonstrate that our models can achieve state-of-the-art performance compared to other popular approaches for multimodal representation learning.

A. Datasets and Feature Extraction

Three popular real-world multimodal datasets, including: 1) LabelMe [41]; 2) UIUC-Sports [42]; and 3) MIR-Flickr [43], are used in the following experiments.

1) LabelMe: LabelMe dataset [41] is constructed via using online tool [11] to collect images from the following eight categories, including: 1) street; 2) coast; 3) forest; 4) mountain; 5) highway; 6) tall building; 7) inside city; and 8) open country. Specifically, 200 images are selected for each class, resulting in a total of 1600 images, and split evenly into the training and testing sets.

2) UIUC-Sports: UIUC-Sports dataset [42] contains 1792 images, covering eight classes: 1) badminton; 2) croquet; 3) rowing; 4) rockclimbing; 5) snowboarding; 6) sailing; 7) polo; and 8) bocce. After resizing, the images of each class are also split evenly into the training and testing sets, the same as the previous work [11].

3) MIR-Flickr: MIR-Flickr dataset [43] contains 1 million images equipped with annotated tags, which are retrieved from the social photography website Flickr. Among these retrieved image-text pairs, 25 000 pairs are annotated for 24 concepts and a stricter labeling is done for 14 of these concepts, resulting in a total of 38 classes, where each image may belong to several classes. In the following experiments, 15 000 labeled image-text pairs are used for training and other 10 000 ones used for testing.

4) Feature Extraction: To make a fair comparison, the same 128-D dense SIFT features [11] are used to extract the visual words from both LabelMe and UIUC-Sports datasets. Following the previous work [11], these SIFT features extracted from the training images are quantized into 240 clusters with K-means clustering algorithm, constructing the visual-word vocabulary. For annotated tags, we discard the words occurring less than three times to construct BoW vectors.

For the MIR-Flickr dataset, we adopt the same text and image features used in the experiments of multimodal DBM [17]. A text vocabulary consisting of the 2000 most frequent words is constructed and each text input is represented as a BoW vector. Then the images are represented as visual-word vectors using a 2000-D visual-word vocabulary, which is constructed via clustering SIFT features from the unlabeled images. Distinct from LabelMe and UIUC-Sports

TABLE I
SUMMARY STATISTICS FOR THE DATASETS AFTER FEATURE EXTRACTION

Dataset	N	C	$K_{img}^{(0)}$	$K_{txt}^{(0)}$	$K_g^{(0)}$
LabelMe [41]	1,600	8	240	240	–
UIUC-Sports [42]	1,792	8	240	240	–
MIR-Flickr [43]	2,5000	38	2,000	2,000	1,857

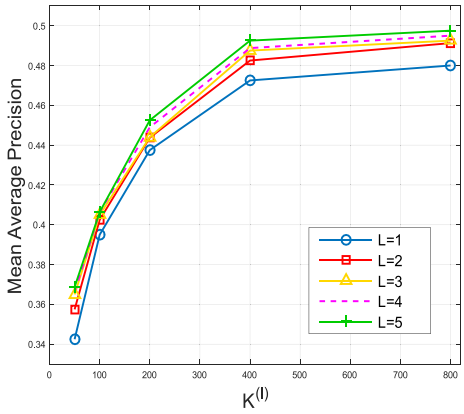


Fig. 4. MAP of the MWVAE for MIR-Flickr 25k as a function of $K^{(l)}$ with various depths $L \in \{1, 2, 3, 4, 5\}$.

datasets, MIR-Flickr provides additional 1857-D global features, consisting of multiple handcraft descriptors, which can be integrated into our models as described in Section IV-B.

For an intuitive insight, the summary statistics of all benchmarks are listed in Table I (N: dataset size, C: number of target classes, $K_{\text{img}}^{(0)}$: visual-word vocabulary size, $K_{\text{txt}}^{(0)}$: text-word vocabulary size, and $K_{\text{img}}^{(0)}$: global image feature length).

B. Model Architecture Learning

We first focus on investigating the influence of the network structure of MWVAE, including both the aspects of network width and depth. We construct MWVAEs with various network structures for unsupervisedly extracting latent representations from MIR-Flickr and the mean average precision (MAP) over all classes is used to measure the performance. Specifically, 15 000 randomly selected image-text pairs are used to train a set of MWVAEs with $K^{(l)} \in \{50, 100, 200, 400, 800\}$ and $L \in \{1, 2, 3, 4, 5\}$, setting the hyperparametersset $\eta^{(l)} = 0.05$ for all l, $\{r_k\}_{k=1}^{K^{(L)}} = 0.1$, and $\{c_n^{(l)}\}_{n=1,l=1}^{N,L} = 1$. Each MWVAE is trained with the hybrid MCMC/VAE inference method as described in Algorithm 1, setting minibatch $m = 200$, and the standard Adam [44] with learning rate 0.001 is used for optimazation. After the model converges, we estimate the posteriors of latent representations for the remaining 10 000 test samples, and perform 1-versus-all classification with the logistic regression on the first hidden layers $\{\boldsymbol{\theta}_{\text{share\_}n}^{(1)}\}_{n=1}^{N}$ to get MAP scores. The results are shown in Figs. 4 and 5 exhibit a clear trend of performance improvement by increasing the width of the hidden layers with a fixed network depth, or by increasing the depth with a fixed layer width.

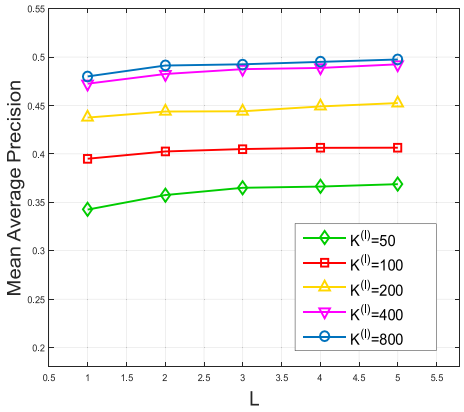


Fig. 5. MAP of the MWVAE for MIR-Flickr 25k as a function of the depth L with various $K^{(l)} \in \{50, 100, 200, 400, 800\}$.

C. Qualitative Tasks

In this part, we evaluate the characteristics of the MWVAE with qualitative tasks, including multimodal retrieval, modality generation, and semantic topic visualization. We construct a 3-layer MWVAE for MIR-Flickr, setting the network structure as $[K^{(1)}, K^{(2)}, K^{(3)}] = [500, 200, 100]$, and other training details are the same as described in Section V-B.

1) Multimodal Retrieval: We first perform multimodal retrieval tasks, taking image-text pairs as input, to evaluate the expressivity of the multimodal latent representations inferred by MWVAE. Given a query image-text pair, the task is to retrieve other similar pairs from a collection with the inferred multimodal latent representations. Specifically, for each query pair randomly selected from MIR-Flickr, we calculate the cosine similarity between the latent representations of the query sample and the other remaining ones to measure their relevance.

As shown in Fig. 6, for each query sample, we exhibit the corresponding retrieved top-5 most relevant image-text pairs in the same row. From the results, the retrieved image-text pairs are highly relevant to the query sample listed in the leftmost column, showing that the MWVAE can provide expressive semantic latent representations for multimodal inputs.

2) Modality Generation: Second, we tend to evaluate the generative ability of the MWVAE, which can be formulated as fixing the missing modality given another observed one. Following bimodal DAE [34], we mask either image or text input to infer the single-modality latent representation $\{\boldsymbol{\theta}_{\text{share\_}n}^{(l)}\}_{n=1,l=1}^{N,L}$ via directly projection, and then fix the missing modality using the generative model mPGBN (decoder).

As shown in Fig. 7, we randomly select several images covering various categories of MIR-Flickr and display the corresponding generated tags on the right side. Benefit from tightly coupling different modalities in a hierarchial fashion, the MWVAE can successfully accomplish the missing text imputation given the image and these generated tags are highly correlated with the corresponding image at multiple semantic levels. Taking the third image of the first row as an example, the MWVAE not only generates scene-level words, such as "winter," "storm" and "snow," as the main part of the image

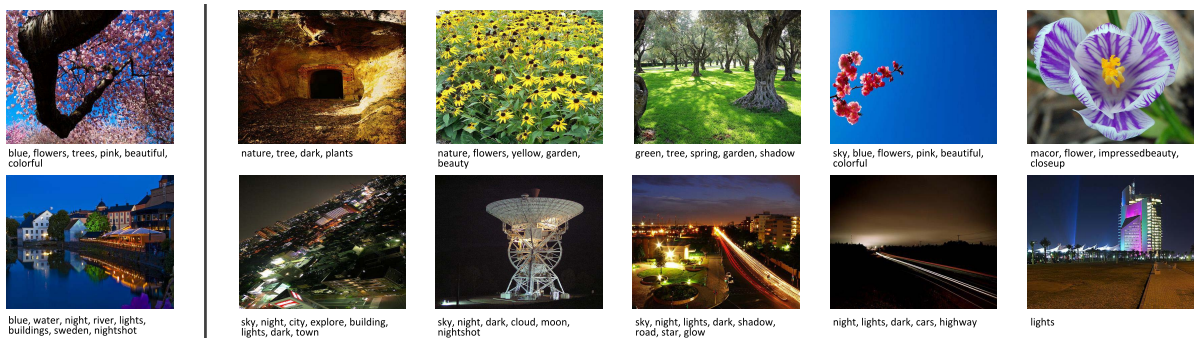


Fig. 6. Top-5 relevant image-text pairs retrieved with the multimodal latent representations inferred by the MWVAE. For each row, the leftmost image-text pair indicates the query sample, while the others on the right side are retrieved samples with the highest cosine similarities in the database.

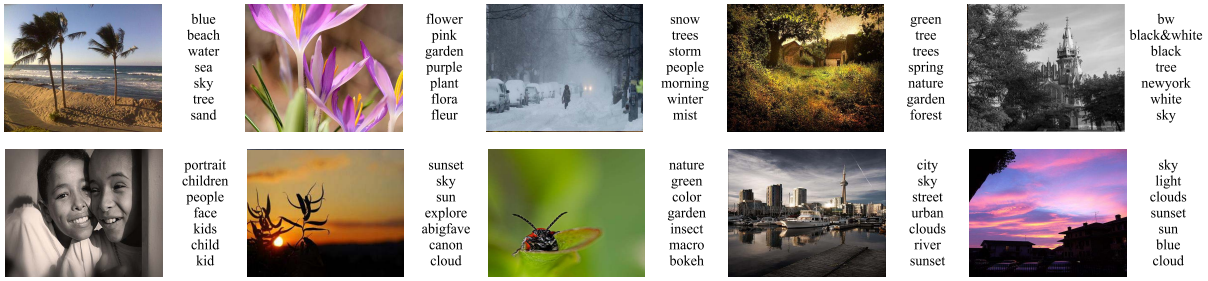


Fig. 7. Examples of annotated tags (right column) generated by the MWVAE conditioned on the observed images (left column).

is white but also captures more subtle information like "tree" and "people" that also occur in the corresponding image.

In addition, we also evaluate the quality of the generated image features of the MWVAE, taking only text as input. To visualize these image features intuitively, we retrieve the top-5 relevant images from MIR-Flickr for each generated image feature, via measuring their Euclidean distances. From the results shown in Fig. 8, it is obvious that the MWVAE can also impute missing image modality, which provides a more impressive explanation for text input.

3) Semantic Topic Visualization: To understand the specific and general aspects of the image-text pairs for training, two hierarchical modality-specific trees, which pick the 20th node of the top hidden layer as their root nodes, are constructed to visualize the topics at different semantic levels, specifically $\{\mathbf{\Phi}_{\text{img}}^{(l)}, \mathbf{\Phi}_{\text{txt}}^{(l)}\}_{l=1}^{L}$ of MWVAE learned from MIR-Flickr. For each modality-specific tree, we only retain the node whose connection from node k at layer l to node k' at layer $l-1$ satisfies the constraint of $\Phi_{k'k}^{(l)} > \tau^{(l)}/K^{(l-1)}$, where $\tau^{(l)}$ is a hyperparameter to adjust the complexity of a tree, and we set $\tau^{(l)} = 10$ for all l. Then we can explore the connections between the image themes and text topics, through projecting them to the corresponding visible layers as described in Section III-A.

As shown in Fig. 2, following the branches of the text-specific tree, it is obvious that when moving along the tree from top to bottom, these text topics become more and more specific. According to the keywords displayed inside text boxes, the root node on "sky clouds sunset flower people" is split into three nodes on behalf of distinct semantic meanings when moving from the layer three to the layer two. Specifically, the three nodes of the second layer are mainly about "flower animal garden," "girl boy child," and "sunset clouds sky" and all split into more specific topics of the first hidden layer. For the image-specific tree, we express the "key words" of a image node at different semantic levels with the top three or four relevant images that are retrieved using the image feature of the corresponding node, considering the low-level features cannot be directly visualized.

Comparing both the text and image trees shown in Fig. 2, we can find that the retrieved images, which reveal the inferred theme of a image node, have a highly correlated semantic meaning with the keywords of the corresponding text topic. Taking the fourth node at the second layer as an example, the keywords of this text node are mainly about "flower animal," while the corresponding image theme characterized by retrieved images is highly related to both "flower" and "animal" aspects. Then the fourth text node of the layer two on "flower animal" is split into several nodes of the layer one, including node 80 on "flower" and node 29 on "animal," which are also the semantic meanings of the retrieved images displayed in the corresponding image nodes.

D. Quantitative Tasks

To further evaluate our models, we make quantitative comparisons with other popular multimodal learning approaches in this part, showing that our models achieve state-of-the-art performance over strong baselines.

1) Image Annotation and Classification Tasks: In the first part, we measure the performance of single-layer MWVAE/sMWVAE on LabelMe and UIUC-Sports datasets, which are popular benchmarks for image annotation and

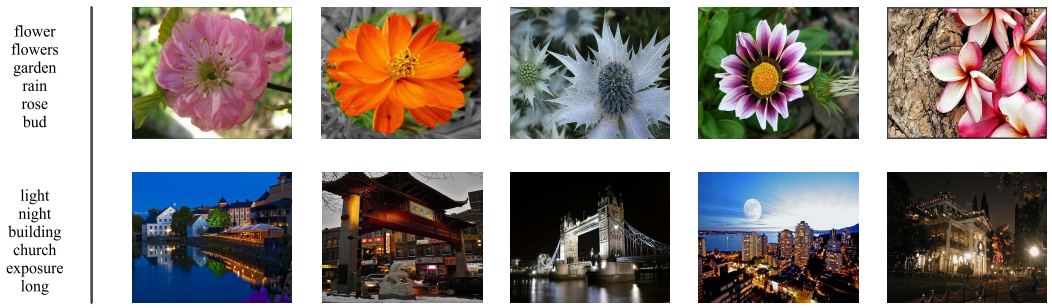


Fig. 8. Top-5 relevant images retrieved using the features generated by the MWVAE conditioned on the given tags at the leftmost column.

classification. To make extensive comparisons, the approaches included in our comparisons are demonstrated as follows.

1) *SPM [45]:* Spatial pyramid matching (SPM) is a computationally efficient extension of orderless BoW image representations, providing a significant performance improvement on challenging scene categorization tasks.
2) *MMLDA [46]:* Max-margin LDA (MMLDA) is a max-margin variant of supervised topic model and two different versions of MMLDA are developed to integrate either discriminative classification or image annotation with generative topic models.
3) *sLDA [11]:* sLDA embeds image tags into a probabilistic model, resulting in a supervised topic model. Thus sLDA combines the terms of images, tags and class labels, performing classification and annotation at the same latent semantic topic space.
4) *DocNDADE/SupDocNADE [36]:* DocNADE is developed to directly model the joint distribution of the words in a document and a supervised variant (SupDocNADE) is further proposed to model the joint distribution over an image's visual words, tags and class label.

To make a fair comparison, the same network structures with 200 hidden topic units are used for our models and the tradeoff hyperparameter λ in sMWVAE is chose based on cross validation. Following DocNADE/supDocNADE [36], we mask the text-modality input as zeros and only use image visual-word vectors to predict labels and annotations in the testing stage. For evaluation metrics, classification accuracy and the average *F*-measure of the top-5 predicted annotations are used to evaluate the performance of image classification and annotation, respectively, following previous works [11].

From the classification results (Accuracy%) listed in Table II, SPM [45], which separates feature extraction and classification stages, achieves a lower classification accuracy on both datasets. Benefit from combining both generative and discriminative aspects, sLDA [11] provides more expressive latent representations than a purely generative approach and slightly outperforms MMLDA [46]. However, sLDA still has difficulty in balancing the generative and discriminative aspects in the loss function, which has been proven quite important for supervised topic modeling [36], leading to a worse performance in the following image annotation task. Through introducing regularization hyperparameter, supDocNADE and sMWVAE can calibrate the generative and discriminative aspects of the loss function, outperforming their unsupervised visions and other mentioned methods. Compared to supDocNADE, sMWVAE provides a distribution estimation rather than a point estimation for multimodal latent representations, which effectively alleviate overfitting and further improve the performance, achieving state-of-art results 84.72% and 78.02% on LabelMe and UIUC-Sports, respectively. Similar conclusions can be found in the image annotation task (F-measure%) in Table II. Here, we emphasize that sMWVAE achieves comparable results to MMLDA on LabelMe, which performs classification and annotation separately, and state-of-art performance 47.46% on UIUC-sports.

TABLE II
PERFORMANCE COMPARISONS OF MWVAE AND SMWVAE WITH DIFFERENT METHODS ON LABELME AND UIUC-SPORTS DATASETS

Method	LabelMe		UIUC-Sports	
	Accuracy	F-measure	Accuracy	F-measure
SPM [45]	80.88	43.68	72.33	41.78
MMLDA [46]	81.47	**46.64**	74.65	44.51
sLDA [11]	81.87	38.7	76.87	35.00
DocNADE [36]	81.97	43.32	74.23	46.38
SupDocNADE [36]	84.43	43.87	77.29	46.95
MWVAE	82.51	45.97	75.18	46.93
sMWVAE	**84.72**	46.44	**78.02**	**47.46**

2) Multimodal Classification Tasks: Then we evaluate the performance of deep extensions of MWVAE/sMWVAE on MIR-Flickr, which is a challenging large-scale multimodal benchmark. To make a comprehensive investigation, we list both unsupervised and supervised models included in our comparisons as follows.

1) *Bimodal DAE [34]:* Bimodal DAE is first initialed with the parameters of a pretrained bimodal DBN [16] and then finetuned to reconstruct both modalities given either modality-specific input.
2) *Multimodal DBN/DBM [16], [17]:* mDBN constructs modality-specific DBNs for different modalities and then combine them via sharing an RBM as their top layers. Further, the mDBM is developed by replacing the DBNs of mDBN with corresponding modality-specific DBMs.
3) *DeepDocNADE [36]:* DeepDocNADE is the deep extension of DocNADE via extending the neural network in DocNADE into a multilayer version.
4) *Multiple Kernel Learning SVMs [31]:* Multiple Kernel learning SVMs is a semisupervised learning approach, leveraging the information of annotated tags equipped with unlabeled images in a two-step process.

TABLE III
PERFORMANCE AND TESTING TIME COMPARISONS OF UNSUPERVISED METHODS ON THE MIR-FLICKER DATASET

Method	MAP	Time
RANDOM	0.124	0.21s
TF-IDF	0.384	0.31s
SVM	0.475	0.32s
LDA [6]	0.492	10.73s
Bimodal Deep Autoencoder [34]	0.510	0.45s
Multimodal DBN [16]	0.503	0.61s
Multimodal DBM [17]	0.513	0.60s
Multimodal PGBN (1 layer)	0.515±0.004	11.88s
Multimodal PGBN (2 layers)	0.524±0.004	26.45s
Multimodal PGBN (3 layers)	0.532±0.003	40.32s
DeepDocNADE (1 layer) [36]	0.510±0.004	0.33s
DeepDocNADE (2 layers) [36]	0.517±0.004	0.38s
DeepDocNADE (3 layers) [36]	0.521±0.003	0.45s
DeepDocNADE-hc (1 layer) [36]	0.528±0.005	0.41s
DeepDocNADE-hc (2 layers) [36]	0.535±0.004	0.45s
DeepDocNADE-hc (3 layers) [36]	0.539±0.004	0.51s
MWVAE (1 layer)	0.515±0.003	0.38s
MWVAE (2 layers)	0.524±0.004	0.43s
MWVAE (3 layers)	0.530±0.004	0.50s
MWVAE-hc (1 layer)	0.570±0.005	0.42s
MWVAE-hc (2 layers)	0.581±0.005	0.48s
MWVAE-hc-no-both (3 layer)	0.578±0.003	0.50s
MWVAE-hc-no-norm (3 layer)	0.582±0.004	0.52s
MWVAE-hc-no-link (3 layer)	0.585±0.003	0.52s
MWVAE-hc (3 layers)	**0.587**±0.004	0.54s

5) *TagProp [47]:* A weighted nearest neighbor model that predicts the term relevance of images via taking a weighted sum of the annotations from the visually most relevant images within a collection of image-text pairs.
6) *Supervised Multimodal DBM [48]:* Supervised multimodal DBM is constructed via incorporating tree-based priors into the discriminatively trained neural networks.
7) *MDRNN [18]:* Multimodal deep RNN (MDRNN) adopts a recurrent encoding function to predict the target modality given another modality input via minimizing the variation of information.
8) *SupDeepDocNADE [36]:* SupDeepDocNADE is a supervised variant of DeepDocNADE that introduces the class label modality to extract more discriminative multimodal latent representations.

Note that the top three items are unsupervised methods while the remaining others are supervised ones. Then we construct three different MWVAEs/sMWVAEs with $T \in \{1, 2, 3\}$ and set the same network structures as pervious methods [16], [17] with $K^{(1)} = K^{(2)} = K^{(3)} = 2048$. To further evaluate the advantage of introducing global image features, we consider constructing MWVAEs/sMWVAEs coupled with either traditional handcraft or deep-learning-based image features. For handcraft features, we use the 1857-D global image features provided by DeepDocNADE [36], which are the same as those provided by mDBM [17]. For deep-learning-based features, VGG-16 [49] is applied to extract 2048-D global image features after resizing each image into the same size 224×224. The details of integration have been described in Section IV, and we add the suffix $-hc$ for these models coupled with traditional handcraft features and $-cnn$ for others coupled with CNN-based features. MAP is used for evaluation and we report the average performance of the five independent splits on MIR-Flickr, where the training/testing/validation partitions are the same as other methods.

TABLE IV
PERFORMANCE AND TESTING TIME COMPARISONS OF SUPERVISED METHODS ON THE MIR-FLICKER DATASET

Method	MAP	Time
Multiple Kernel Learning SVMs [31]	0.623	0.53s
TagProp [47]	0.640	0.48s
Supervised Multimodal DBM [48]	0.651	0.61s
MDRNN [18]	0.686	4.52s
SupDeepDocNADE-hc (1 layer) [36]	0.639±0.004	0.41s
SupDeepDocNADE-hc (2 layers) [36]	0.648±0.006	0.44s
SupDeepDocNADE-hc (3 layers) [36]	0.654±0.005	0.51s
SupDeepDocNADE-cnn (1 layer) [36]	0.735±0.003	0.41s
SupDeepDocNADE-cnn (2 layers) [36]	0.738±0.003	0.45s
SupDeepDocNADE-cnn (3 layers) [36]	0.741±0.004	0.52s
sMWVAE-hc (1 layer)	0.650 ±0.003	0.43s
sMWVAE-hc (2 layers)	0.657 ±0.003	0.48s
sMWVAE-hc (3 layers)	0.663 ±0.004	0.55s
sMWVAE-cnn (1 layer)	0.746±0.002	0.42s
sMWVAE-cnn (2 layers)	0.749±0.002	0.48s
sMWVAE-cnn (3 layers)	**0.751**±0.003	0.54s

As the unsupervised comparisons in Table III, we first provide the results of traditional handcraft features as baselines, including RANDOM, TF-IDF, and original BoW features, in the first group. LDA achieves a better performance than these handcraft features, showing the effectiveness of topic models in multimodal representation learning, but there is still a gap between the shallow LDA and other deep models. For RBM-based methods, mDBN separates the training stages of modality-specific DBNs and the shared top-layer RBM, resulting in a slightly worse performance than mDBM, which trains the whole deep network jointly. However, these RBM-based methods, including Bimodal DAE, are limited by only sharing the top hidden layer and have difficulty in capturing the connections between different modalities at multiple semantic levels. Benefit from tightly coupling multilayer latent representations, mPGBNs, DeepDocNADEs and MWVAEs outperform other methods, and there is a clear trend of performance improvement with the increase of the network depths. Specifically, taking advantage of providing analytic posteriors, mPGBNs outperform MWVAEs under the same network structure settings, but still have difficulty in plugging in extra side information, limited by the Gibbs sampler. Through integrating handcraft global image features into latent representations, DeepDocNADE-*hc* and MWVAE-*hc* have a great improvement in the multimodal classification. Moreover, MWVAEs tend to outperform DeepDocNADEs under the same network settings, proving the superiority of the distribution estimation for latent representations and the sparsity provided by the Weibull reparameterization. Here, we highlight that the developed MWVAE-*hc* with three hidden layers achieves the state-of-the-art performance, MAP of 0.587, on the MIR-Flickr in an unsupervised manner. To investigate the effectiveness of link function and adaptive normalization, we add an ablation study for the 3-layer MWVAE-*hc* at the bottom of Table III. From the results, we can find that either the link function or adaptive normalization can bring the performance improvement, and the bare 3-layer MWVAE-*hc*-no-both can still outperform the 3-layer DeepDocNADE-*hc*.

Table IV presents the comparisons between sMWVAEs and other outstanding supervised baselines. From the results, Multiple kernel learning SVMs and TagProp only consider

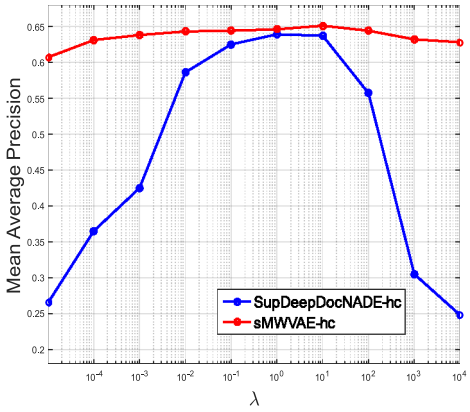


Fig. 9. MAP of the SupDeepDocNADE-*hc* and sMWVAE-*hc*, which incorporate handcraft global image features, for MIR-Flickr 25k as a function of the regularization hyperparameter λ.

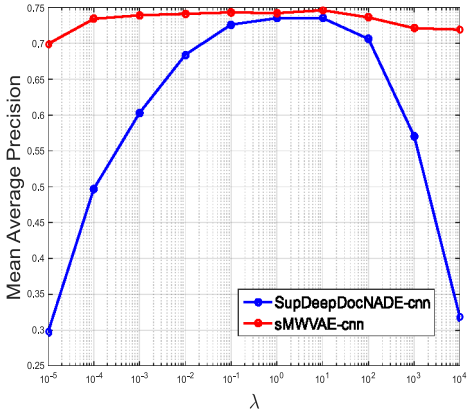


Fig. 10. MAP of the SupDeepDocNADE-*cnn* and sMWVAE-*cnn*, which incorporate CNN-based global image features, for MIR-Flickr 25k as a function of the regularization hyperparameter λ.

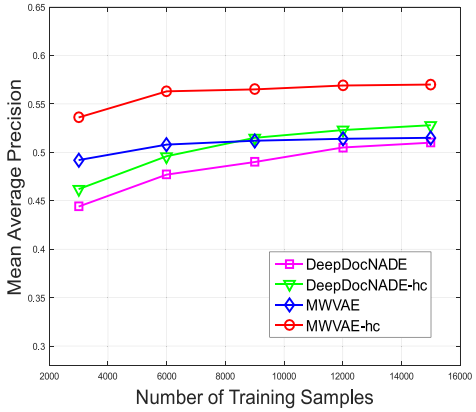


Fig. 11. MAP of unsupervised methods with different number of training samples on MIR-Flickr 25k.

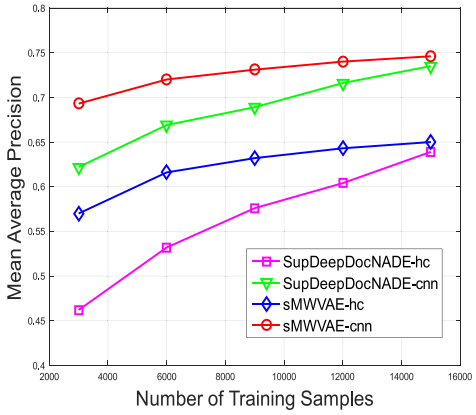


Fig. 12. MAP of supervised methods with different number of training samples on MIR-Flickr 25k.

the discriminate aspects in their loss functions, and achieve MAP 0.623 and 0.640, respectively. Moving beyond maximizing the likelihood, MDRNN is trained to minimize the variation of information and performs a recurrent encoding structure in prediction tasks, outperforming supervised multimodal DBM at the cost of increasing the network complexity. We emphasize that Multiple kernel learning SVMs, supervised multimodal DBM and MDRNN are required to be pretrained on the unlabeled 975 000 image-text pairs from MIR-Flickr. Without any pretraining stage, we evaluate the performance of SupDeepDocNADEs and sMWVAEs coupled with either handcraft or CNN-based global image features. As shown in the last two groups, these methods coupled with CNN-based image features outperform others coupled with handcraft ones, proving that CNNs can extract more discriminative image features. Moreover, sMWVAEs outperform SupDeepDocNADEs under the same network settings and sMWVAE-*cnn* with three hidden layers achieves state-of-art supervised multimodal classification performance with MAP of 0.751.

From the aspect of model efficiency, we exhibit the comparisons of testing time in the rightmost columns of both Tables III and IV, for unsupervised and supervised methods, respectively. Limited by the Gibbs sampler, LDA [6] and mPGBN require batch-level sampling iterations to obtain the multimodal latent representations, leading to a lot of time cost in the testing phase. Among the other methods based on feedforward projection, there is a clear trend that the testing time cost will increase with the complexity of the network structure (or the depth of the same method), where MDRNN [18] could be the most time-consuming method, constrained by the recurrent structure of RNN in prediction. Thanks to the Weibull-based inference network, the developed MWVAEs/sMWVAEs can directly infer the stochastic latent representations with feedforward projection, achieving a comparable efficiency to other deep-learning-based methods.

3) Balance Between the Generation and Classification: To investigate the impact of the regularization hyperparameter λ in sMWVAE, we perform additional experiments with various values of λ on MIR-Flickr. Two shallow sMWVAEs with 2048 hidden units are constructed, which incorporate either handcraft (-*hc*) or CNN-based (-*cnn*) global image features, and other details are the same as described above. As shown in Fig. 9, the sMWVAE-*hc* outperforms SupDeepDocNADE-*hc* in the range of λ from 10^{-5} to 10^{4}, and achieves the best performance at $\lambda = 10^{1}$, which indicates that the importance of balancing the generative and discriminative aspects in the loss function. Moreover, the performance of

SupDeepDocNADE-*hc* is extremely sensitive to the varies of λ, while sMWVAE-*hc* is relatively stable. The underlying reason could be that the sMWVAE provides a distribution estimation for the multimodal latent representations, which effectively alleviates the impact of the discriminate aspect, contributing to the flexibility of hyper parameter selection for other downstream tasks. Similar conclusions can be obtained by comparing the performance of SupDeepDocNADE-*cnn* and sMWVAE-*cnn* as shown in Fig. 10, which further confirm the superiority of our methods.

4) Robustness to the Smaller Training Set: To further investigate the impact of the training set size, we train the aforementioned models with variable number of training samples $N \in \{3000, 6000, 9000, 12\,000, 15\,000\}$ selected from MIR-Flickr. For an intuitive comparison, we divide these methods into two categories according to either unsupervised or supervised training stage, and report the average performance achieved by five independent runs as shown in Figs. 11 and 12, respectively.

As the unsupervised comparisons exhibited in Fig. 11, there is a clear trend that the MAP scores of all methods will improve as the number of samples increases, equipped with a gradually slowing down rate of performance improvement. Specifically, the performance of DeepDocNADE and DeepDocNADE-*hc* drop sharply as the number of samples decreases, while the MWVAE and MWVAE-*hc* tend to be more robust with a smaller dataset size. Notably, the MWVAE without global image features even outperforms DeepDocNADE-*hc* when the amount of training samples is less than 6000. For the supervised comparisons in Fig. 12, the sMWVAE coupled with handcraft (-*hc*) or CNN-based (-*cnn*) global image features is more robust to the reduction of training set, which is consistent with the unsupervised situation. We attribute the robustness of our methods to the following reasons.

1) The deep probabilistic topic model mPGBN (decoder) for data generation, which increases the diversity of multimodal observations [50].
2) The distribution estimation for multimodal latent representations, which introduces the stochastics into latent representations and alleviates the overfitting when the training samples are insufficient.

VI. Conclusion

In this article, we construct a novel multimodal probabilistic topic model named mPGBN that can tightly couple hierarchical latent representations of different modalities, providing an interpretable network structure to illustrate the connections between these modalities at multiple semantic levels. For both efficient inference and easy to plug in side information, MWVAE is developed to couple the mPGBN (decoder) and a Weibull-based multimodal inference network (encoder), equipped with a corresponding hybrid MCMC/VAE inference method. Based on MWVAE, we improve the expressivity of the inferred multimodal latent representations, via incorporating global image features, either handcraft or CNN-based. Then we evaluate the effectiveness of MWVAE with extensive qualitative experiments, showing that the MWVAE can successfully accomplish both missing modality imputation and multimodal retrieval tasks. Further quantitative analysis on various popular benchmark datasets demonstrate that our proposed models can achieve state-of-the-art performance on extracting multimodal latent representations.

References

[1] L. Zhang, Y. Gao, C. Hong, Y. Feng, J. Zhu, and D. Cai, "Feature correlation hypergraph: Exploiting high-order potentials for multimodal recognition," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1408–1419, Aug. 2014.

[2] L. Zhu, J. Shen, H. Jin, R. Zheng, and L. Xie, "Content-based visual landmark search via multimodal hypergraph learning," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2756–2769, Dec. 2015.

[3] X. Zhang, S. Wang, Z. Li, and S. Ma, "Landmark image retrieval by jointing feature refinement and multimodal classifier learning," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1682–1695, Jun. 2018.

[4] A. Humm, J. Hennebert, and R. Ingold, "Combined handwriting and speech modalities for user authentication," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 39, no. 1, pp. 25–35, Jan. 2009.

[5] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, "Fusing visual and behavioral cues for modeling user experience in games," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1519–1531, Dec. 2013.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[7] M. Zhou, L. Hannah, D. B. Dunson, and L. Carin, "Beta-negative binomial process and Poisson factor analysis," in *Proc. AISTATS*, 2012, pp. 1462–1471.

[8] M. Zhou, Y. Cong, and B. Chen, "The Poisson gamma belief network," in *Proc. NeurIPS*, 2015, pp. 3043–3051.

[9] C. Wang, H. Zhang, B. Chen, D. Wang, Z. Wang, and M. Zhou, "Deep relational topic modeling via graph Poisson gamma belief network," in *Proc. NeurIPS*, 2020, pp. 488–500.

[10] W. Chen, C. Wang, B. Chen, Y. Liu, H. Zhang, and M. Zhou, "Bidirectional convolutional Poisson gamma dynamical systems," in *Proc. NeurIPS*, 2020, pp. 3673–3685.

[11] W. Chong, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *Proc. IEEE CVPR*, 2009, pp. 1903–1910.

[12] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proc. ACM SIGIR*, 2003, pp. 127–134.

[13] D. Putthividhy, H. T. Attias, and S. S. Nagarajan, "Topic regression multi-modal latent dirichlet allocation for image annotation," in *Proc. IEEE CVPR*, 2010, pp. 3408–3415.

[14] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Proc. NeurIPS*, 2008, pp. 121–128.

[15] J. Ferreira, J. Lobo, P. Bessiere, M. Castelo-Branco, and J. Dias, "A Bayesian framework for active artificial perception," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 699–711, Apr. 2013.

[16] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *Proc. ICML Workshop*, vol. 79, 2012, pp. 1–8.

[17] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. NeurIPS*, 2012, pp. 2222–2230.

[18] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," in *Proc. NeurIPS*, 2014, pp. 2141–2149.

[19] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.

[22] N. Srivastava, R. R. Salakhutdinov, and G. E. Hinton, "Modeling documents with deep Boltzmann machines," in *Proc. UAI*, 2013, p. 110.

[23] R. Ranganath, L. Tang, L. Charlin, and D. Blei, "Deep exponential families," in *Proc. AISTATS*, 2015, pp. 762–771.

[24] Z. Gan, C. Chen, R. Henao, D. Carlson, and L. Carin, "Scalable deep Poisson factor analysis for topic modeling," in *Proc. ICML*, 2015, pp. 1823–1832.

[25] D. P. Kingma and M. Welling, "Stochastic gradient VB and the variational auto-encoder," in *Proc. ICLR*, 2014, pp. 1–6.

[26] Y. H. Tsai, P. P. Liang, A. Zadeh, L. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. ICLR*, 2019, pp. 1–6.

[27] M. Wu and N. D. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *Proc. NeurIPS*, 2018, pp. 5580–5590.

[28] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *Proc. NeurIPS*, 2016, pp. 3738–3746.

[29] Z. Dai, A. Damianou, J. González, and N. Lawrence, "Variational auto-encoded deep Gaussian processes," in *Proc. ICLR*, 2016, pp. 1–9.

[30] C. Wang, B. Chen, and M. Zhou, "Multimodal Poisson gamma belief network," in *Proc. AAAI*, 2018, pp. 2492–2499.

[31] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE CVPR*, 2010, pp. 902–909.

[32] Y. Zhen, Y. Gao, D.-Y. Yeung, H. Zha, and X. Li, "Spectral multimodal hashing and its application to multimedia retrieval," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 27–38, Jan. 2016.

[33] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

[34] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011, pp. 689–696.

[35] H. Larochelle and S. Lauly, "A neural autoregressive topic model," in *Proc. NeurIPS*, 2012, pp. 2708–2716.

[36] Y. Zheng, Y.-J. Zhang, and H. Larochelle, "A deep and autoregressive approach for topic modeling of multimodal data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1056–1069, Jun. 2016.

[37] M. Zhou, Y. Cong, and B. Chen, "Augmentable gamma belief networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 5656–5699, 2016.

[38] C. Wang, B. Chen, S. Xiao, and M. Zhou, "Convolutional Poisson gamma belief network," in *Proc. ICML*, vol. 97, 2019, pp. 6515–6525.

[39] Y. Cong, B. Chen, H. Liu, and M. Zhou, "Deep latent dirichlet allocation with topic-layer-adaptive stochastic gradient riemannian MCMC," in *Proc. ICML*, vol. 70, 2017, pp. 864–873.

[40] H. Zhang, B. Chen, D. Guo, and M. Zhou, "WHAI: Weibull hybrid autoencoding inference for deep topic modeling," in *Proc. ICLR*, 2018, pp. 1–6.

[41] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and Web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.

[42] L.-J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proc. IEEE ICCV*, 2007, pp. 1–8.

[43] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. ACM MIR*, 2008, pp. 39–43.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–6.

[45] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE CVPR*, vol. 2, 2006, pp. 2169–2178.

[46] Y. Wang and G. Mori, "Max-margin latent dirichlet allocation for image classification and annotation," in *Proc. BMVC*, vol. 2, 2011, p. 7.

[47] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid, "Image annotation with tagprop on the MIRFLICKR set," in *Proc. MIR*, 2010, pp. 537–546.

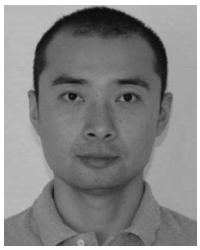
[48] N. Srivastava and R. R. Salakhutdinov, "Discriminative transfer learning with tree-based priors," in *Proc. NeurIPS*, 2013, pp. 2094–2102.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2014, pp. 1–6.

[50] Q. Yang, W.-N. Chen, Y. Li, C. P. Chen, X.-M. Xu, and J. Zhang, "Multimodal estimation of distribution algorithms," *IEEE Trans. Cybern.*, vol. 47, no. 3, pp. 636–650, Mar. 2017.

Chaojie Wang received the B.S. degree in electronic engineering from Xidian University, Xi'an, China, in 2016, where he is currently pursuing the Ph.D. degree in signal processing.

His research interests focus on statistical machine learning and its combinations with real-world applications, including multimodal learning, natural language processing, and knowledge graphs.



Bo Chen (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic engineering from Xidian University, Xi'an, China, in 2003, 2006, and 2008, respectively.

He became a Postdoctoral Fellow, a Research Scientist, and a Senior Research Scientist with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA, from 2008 to 2012. Since 2013, he has been a Professor with the National Laboratory for Radar Signal Processing, Xidian University. His current research interests include statistical machine learning, statistical signal processing, and radar automatic target detection and recognition.

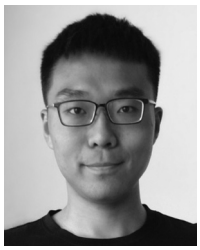
Dr. Chen was a recipient of the Honorable Mention for 2010 National Excellent Doctoral Dissertation Award and is selected into Overseas Talent by Chinese Central Government in 2014.

Suchen Xiao received the B.S. and M.S. degrees in electronic engineering from Xidian University, Xi'an, China, in 2017 and 2020, respectively.

He is currently working as an Algorithm Engineer with ByteDance, Beijing, China.

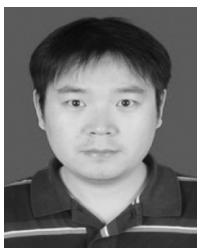
Zhengjue Wang received the B.S., M.S., and Ph.D. degrees in electronic engineering from Xidian University, Xi'an, China, in 2013, 2016, and 2019, respectively.

Her research interests include probabilistic model and deep learning, and their applications in image super-resolution, hyperspectral image fusion, and natural language processing.



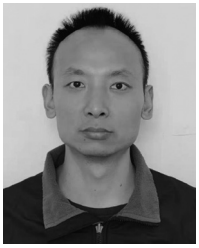
Hao Zhang received the B.S. and Ph.D. degrees in electronic engineering from Xidian University, Xi'an, China, in 2012 and 2019, respectively.

He worked as a Postdoctoral Researcher of Electrical and Computer Engineering with Duke University, Durham, NC, USA, in 2019. He is currently working as a Postdoctoral Researcher with the Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. His research interests include statistical machine learning and its combinations with deep learning, and the neural language processing.



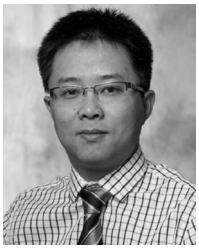
Penghui Wang received the B.S. degree in communication engineering from the National University of Defense Technology, Changsha, China, in 2005, and the Ph.D. degree in signal processing from Xidian University, Xi'an, China, in 2012.

He is currently an Associate Professor with the National Laboratory of Radar Signal Processing, Xidian University. His research interests include radar signal processing and automatic target recognition.



Ning Han received the B.S., M.S., and Ph.D. degrees from Ordnance Engineering College, Shijiazhuang, China, in 2006, 2009, and 2012, respectively.

He is an Engineer with the Institute of Mechanical Technology, Xi'an, China. His current research interests include sparse signal processing, statistical signal processing, and radar high-resolution imaging.



Mingyuan Zhou received the B.S. degree in acoustics from the Department of Electronic Science and Engineering, Nanjing University, Nanjing, China, in 2005, the M.S. degree in signal and information processing from the Chinese Academy of Sciences, Beijing, China, in 2008, and the Ph.D. degree in electrical and computer engineering from Duke University, Durham, NC, USA, in 2013.

He is an Associate Professor of Statistics with the McCombs School of Business, University of Texas at Austin, Austin, TX, USA. His research interest lies at the intersection of Bayesian statistics and machine learning, covering a diverse set of research topics in statistical theory and methods, hierarchical models, Bayesian nonparametrics, statistical inference for big data, and deep learning. He is currently focused on advancing both statistical inference with deep learning and deep learning with probabilistic methods.

Dr. Zhou has served as the 2018–2019 Treasurer of the Bayesian Nonparametrics Section, International Society for Bayesian Analysis, and as the Area Chair for leading machine learning conferences, including NeurIPS 2017–2019, ICLR 2019, and AAAI 2020.