

A Deep Approach for Multi-modal User Attribute Modeling

Xiu Huang, Zihao Yang, Yang Yang^(✉), Fumin Shen, Ning Xie,
and Heng Tao Shen

School of Computer Science and Technology and Center for Future Media,
University of Electronic Science and Technology of China, Chengdu, China
hxiu321@163.com, zivon396@163.com, dlyyang@gmail.com, fumin.shen@gmail.com,
seanxiening@gmail.com, shenhengtao@hotmail.com

Abstract. With the explosive growth of user-generated contents (e.g., texts, images and videos) on social networks, it is of great significance to analyze and extract people's interests from the massive social media data, thus providing more accurate personalized recommendations and services. In this paper, we propose a novel multimodal deep learning algorithm for user profiling, dubbed multi-modal User Attribute Model (mmUAM), which explores the intrinsic semantic correlations across different modalities. Our proposed model is based on Poisson Gamma Belief Network (PGBN), which is a deep learning topic model for count data in documents. By improving PGBN, we succeed in addressing the problem of learning a shared representation between texts and images in order to obtain textual and visual attributes for users. To evaluate the effectiveness of our proposed method, we collect a real dataset from Sina Weibo. Experimental results demonstrate that the proposed algorithm achieves encouraging performance compared with several state-of-the-art methods.

Keywords: User profiling · Deep learning · Multi-model · Social media

1 Introduction

With the rapid development of social networks, massive information (e.g., texts, images and videos) generated by users is emerging on various social media platforms. The activities people participating in and the contents people producing play a significant role of analyzing people's interests and preferences, which are of great importance to provide personalized recommendation and on-line retrieval for them. In particular, microblogging is now one of the most popular social media services, where people are keen on posting daily activities, sharing opinions and focusing on hot and interesting topics. For example, Sina Weibo¹, a commonly used social media platform in China, has attracted a great amount of users to participate in. Released by Sina Weibo Data Center, the number of

¹ <http://www.weibo.com>.

monthly active users approaches to 222 million up to October 2015. Besides, the social media applications involve multi-modal data, where the visual information is vital to strengthen the description of short texts.

In order to explore user attributes, prior works construct topic modeling from users' previous behaviors and preferences. For example, Latent Dirichlet Allocation (LDA) [7] is a widely used generative probabilistic model for text corpora. By modifying LDA, there are other traditional topic models to tackle the problem of short texts from social media data, such as author topic model [19] and twitter-user model [24]. Besides, previous works also focus on dynamic topic models to analyze the change of topics in data streams, such as Dynamic User Attribute Model (DUAM) [12] which models the dynamics using time windows, and dynamic User Clustering Topic model (UCT) [27] to capture the dynamics of users' interests by integrating the interests at previous time periods with newly collected data in text streams. In addition, there are topic models proposed to explore the correlations among different modalities. For instance, mm-LDA [2] and corr-LDA [6] are presented to learn the correspondence between textual and visual information. Cross-Media-LDA (CMLDA) [5] is also proposed to discover the intrinsic correlations among multiple media types for social event summarization. Some similar methods proposed by Bian et al. are demonstrated in [3, 4].

Recently, there is a great interest in deep learning, which succeeds in many applications. The Deep Belief Network (DBN) [11] and the Deep Boltzmann Machine (DBM) [20] are deep networks both designed to model binary observations, whose hidden units are also typically restricted to be binary. However, different from conventional deep networks, the Poisson Gamma Belief Network (PGBN) [29] is proposed to construct a deep networks architecture with non-negative real hidden units to automatically tune both the width of each layer and the depth of the network. Despite PGBN learns the representation of count observations, it is a unimodal network and not applicative to short texts of social data. To deal with multi-modal social data, we propose a novel multi-modal User Attribute Model (mmUAM). Different from traditional methods of constructing user interest model that only take account of one layer of topic modelling, our model is designed to capture the correlations among multiple modalities. To facilitate this study, we collect a real dataset from Sina Weibo, on which extensive experiments show the superiority over state-of-the-art methods.

The main contributions of our work are summarized as follows.

1. We propose a novel multi-modal deep learning approach, named multi-modal User Attribute Model (mmUAM), through which we manage to automatically infer user attributes.
2. The proposed mmUAM captures the semantic correlations between texts and images, which enables us to learn effective textual and visual representation for more comprehensive user profiling.
3. We construct a Sina Weibo microblog dataset with multi-modality information. The promising results on this dataset demonstrate the efficacy of our proposed approach.

2 Related Work

Text-based User Profiling. With the tremendous growth of social networks, how to provide more accurate services for users is tough challenging. Previous works have been studied to explore users interests through extracting users' characteristics and preferences from user-generated texts on social media platforms. Generative topic models, such as LDA [7], provide an explicit representation of a document. However, such topic models fail to tackle the sparsity problem of short texts. Many variations of LDA have been proposed. For example, He et al. [10] propose a modified topic model, named Bi-labeled LDA, which utilizes users' relationship information to learn interest tags. Rosen-Zvi et al. [19] extend LDA to propose the author-topic model, which models the content of documents, including the author information. While, Xu et al. [24] introduce a modified author-topic model, twitter-user mode, which outperforms LDA and author-topic model. Besides, some other studies also make attempts to exploit external knowledge to enrich the s of short texts. Abel et al. [1] analyze Twitter activities in semantic way by integrating Twitter posts with related news articles. Instead of introducing external knowledge, Cheng et al., [8] model the generation of word co-occurrence patterns for topic modeling in order to address the sparsity problem of short texts. However, the above topic models are mostly applied to text corpus.

Image-based User Profiling. Deep Convolutional Neural Networks (CNNs) [13] have recently achieved a great success in large scale image feature learning. Consequently, many researches focus on building user profiling by extracting visual information. For instance, Geng et al. [9] propose a deep learning strategy to learn visual features for user profiling on Pinterest² in fashion domain. A Socially Embedded Visual Representation learning (SEVIR) approach [15] has been proposed to capture the semantics and user intentions based on learning image representation, which tackles the sparsity and unreliability problems. Li et al. [14] construct a Gaussian relational topic model by utilizing user-shared images to infer users' interests. Moreover, a pinboard recommendation system for Twitter users is presented in [25], which combines two different social media platforms in order to recommend users for more relevant and interesting topics. Also, the way of exploiting user-tagged Web images for video indexing can be learned in [26]. Despite the visual information exploiting user interests is definitely significant, more works should take account of multiple modalities.

Multi-modal User Profiling. As more and more social media data is integrated with texts, images and videos, most of the works have shifted their focus to dealing with multi-modal data. In [6], the correspondence Latent Dirichlet Allocation (corr-LDA) is a three hierarchical probabilistic mixture model to describe the correlations between images and annotations. Similarly, multi-modal Latent Dirichlet Allocation (mm-LDA) [2] is proposed to learn the joint distribution of images and their associated texts, which is used for social relation mining.

² <http://www.pinterest.com>.

In addition, Bian et al. [5] present a novel probabilistic modal, named Cross-Media-LDA (CMLDA), which aims to explore intrinsic correlations between texts and images for multimedia microblog summarization.

Besides, some deep networks are constructed to learn features among multiple modalities. Ngiam et al. [17] propose a cross-modality deep learning methods based on Restricted Boltzmann Machines (RBMs). Subsequently, Srivastava et al. [23] propose a multi-modal Deep Boltzmann Machine (DBM) model for images and texts. They construct multi-modal DBM by building an image-specific two-layer DBM that uses Gaussian RBM and a text-specific two-layer DBM that utilizes Replicated Softmax model. Similarly, Pang et al. [18] use multi-modal DBN to learn joint representation of the visual, auditory and textual features for user-generated web videos. In addition, the Deep Belief Network (DBN) presented in [22] to create a joint representation for texts and images, is different from DBM in that DBN is a directed model.

Nevertheless, the hidden units of DBM and DBN are typically restricted to be binary. These multi-modal deep learning approaches are not successfully applied to a real dataset for social service. We work on learning features of multiple modalities input data from large-scale real dataset and construct multi-modal deep networks to tackle the sparsity of short texts to explore more relevant interests that meet users' demand. To achieve better inference of our proposed deep topic modal, we employ upward-downward Gibbs sampling.

3 The Proposed Model

3.1 Overview

We employ the conventional bag-of-words method to deal with the texts and images to automatically infer user attributes. To construct our model, we utilize Sina Weibo data with user-generated short texts and corresponding images. Firstly, we extract both texts and images raw features as bag of words and bag of visual words, respectively. Formally, each document is under two different topic distributions. Note that Θ , which is a shared latent variable between visual and textual modalities, is concatenated by textual hidden unit θ_{w-j} and visual hidden unit θ_{v-j} . Topics Φ_w are specific to textual modality and topics Φ_v are unique to visual modality. Then, we build our proposed mmUAM in deep networks with five layers. The performance of multi-modal fusion in five different layers is presented in Sect. 4.3. As we use probabilistic models, upward-downward Gibbs sampling [29] is adopted to infer various parameters.

As we all know, microblogs always consist of short texts and relevant images, in which each text is restricted to 140 characters. Thus, each document is a piece of microblog and is composed of textual content, visual content, or the mixing of textual and visual information. In particular, the observation of an image is represented as a multivariate vector of visual words, which is denoted as v_j in j th document. Similarly, the observation of a text is defined as a vector w_j in j th document. The correlations between the K_0 features of (v_1, v_2, \dots, v_J) can be represented by the columns of Φ_v . In the same way, the correlations between

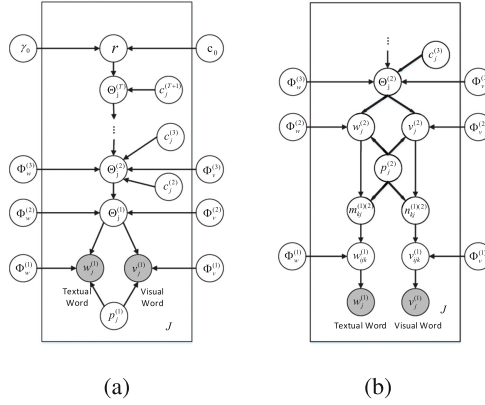


Fig. 1. The graphical illustration of the proposed mmUAM. (a) the mmUAM hierarchical model; (b) a presentation of layer $t = 1$ in the mmUAM.

the K_0 features of (w_1, w_2, \dots, w_J) are captured by the columns of Φ_w . Note that $\Theta_j \in R_+^{K_t}$ is the K_t hidden units of sample the j th document. We use the Poisson likelihood to connect the observed textual content $w_j^{(1)} \in Z^{K_0}$ (visual content $v_j^{(1)} \in Z^{K_0}$) to the product $\Phi_w^{(1)} \theta_{w-j}^{(1)}$ ($\Phi_v^{(1)} \theta_{v-j}^{(1)}$) at layer one as follows

$$w_j^{(1)} \sim Pois \left(\Phi_w^{(1)} \theta_{w-j}^{(1)} \right), \quad v_j^{(1)} \sim Pois \left(\Phi_v^{(1)} \theta_{v-j}^{(1)} \right).$$

3.2 Multi-modal User Attribute Model

Our proposed model is based on PGBN [29], a deep networks architecture that is designed only for text analysis. Then, Zhou et al. [30] propose augmentable gamma belief networks to learn multilayer deep representations for high-dimensional sparse count vectors and nonnegative real vectors. Nevertheless, the augmentable gamma belief networks are not adapted to social data. As a result, we propose a multi-modal user attribute model for multiple modalities data on social media. For microblogging document, we make an assumption that the generated topics are composed of two domains, including textual topics generated from microblog texts, and visual topics generated from posted images. In order to capture correlations of these two modalities, we learn a shared representation between textual and visual information. We use Θ_j to represent the shared gamma distribution between textual and visual information in the j th microblogging document. With T hidden layers, we give the example of our proposed mmUAM fusing in the first hidden layer as follows

$$\begin{aligned}
 \Theta_j^{(T)} &\sim \text{Gam}\left(\mathbf{r}, 1/c_j^{(T+1)}\right), \\
 \Theta_j^{(t)} &\sim \text{Gam}\left(\begin{matrix} \dots \\ \left[\begin{matrix} \Phi_w^{(t+1)} \theta_{w-j}^{(t+1)} \\ \Phi_v^{(t+1)} \theta_{v-j}^{(t+1)} \end{matrix}\right] \end{matrix}, 1/c_j^{(t+1)}\right), \\
 \Theta_j^{(1)} &\sim \text{Gam}\left(\begin{matrix} \dots \\ \left[\begin{matrix} \Phi_w^{(2)} \theta_{w-j}^{(2)} \\ \Phi_v^{(2)} \theta_{v-j}^{(2)} \end{matrix}\right] \end{matrix}, p_j^{(2)}/(1-p_j^{(2)})\right).
 \end{aligned} \tag{1}$$

The graphic representation of the mmUAM is depicted in Fig.1. For $t = 1, 2, \dots, T - 1$, the hidden units $\Theta_j^{(t)} \in R_+^{K_t}$ of layer t are under gamma distribution, which factorize the shape parameters into the concatenation of $\Phi_w^{(t+1)} \theta_{w-j}^{(t+1)}$ and $\Phi_v^{(t+1)} \theta_{v-j}^{(t+1)}$. With $c_j^{(2)} = (1 - p_j^{(2)})/p_j^{(2)}$, $p_j^{(2)}$ and $\{1/c^{(t)}\}_{3, T+1}$ are probability parameters and gamma scale parameters respectively. For the top layer, the gamma shape parameters of hidden units are vector $\mathbf{r} = (r_1, \dots, r_K^{(T)})'$. The columns of $\phi_w^{(t+1)}$ and $\phi_v^{(t+1)}$ decide the correlations between the K_t latent features of $(\Theta_1^{(t)}, \dots, \Theta_J^{(t)})$.

In order to simplify parameter inference, we impose the constraints on $\Phi_w^{(t)}$ and $\Phi_v^{(t)}$ that every column of $\Phi_w^{(t)}$ and $\Phi_v^{(t)}$ has a unit L_1 norm. Thus, for $t \in \{1, \dots, T - 1\}$, the hierarchical model is completed as follows

$$\begin{aligned}
 \phi_{w-k}^{(t)} &\sim \text{Dir}(\eta^t, \dots, \eta^t), \quad \phi_{v-k}^{(t)} \sim \text{Dir}(\xi^t, \dots, \xi^t), \\
 c_0 &\sim \text{Gam}(e_0, 1/f_0), \quad \gamma_0 \sim \text{Gam}(a_0, 1/b_0), \quad r_k \sim \text{Gam}(\gamma_0/K_T, 1/c_0).
 \end{aligned}$$

For $t \in \{3, \dots, T + 1\}$, we have

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

we divide T hidden layers into T related subproblems, thus every subproblem has the similar way of solution.

Lemma 1 (*augment-and-conquer the mmUAM*). With $p_j^{(1)} = 1 - e^{-1}$ and

$$p_j^{(t+1)} = -\ln(1 - p_j^{(t)}) / \left[c_j^{(t+1)} - \ln(1 - p_j^{(t)}) \right]. \tag{3}$$

For $t \in \{1, \dots, T\}$, we can define that the observed (if $t = 1$) or some latent (if $t \geq 2$) textual contents $w_j^t \in Z^{K_t-1}$ are under the Poisson distribution with the product $\Phi_w^t \theta_{w-j}^t$, and the observed (if $t = 1$) or some latent (if $t \geq 2$) visual word counts $v_j^t \in Z^{K_t-1}$ are under the Poisson distribution with the product $\Phi_v^t \theta_{v-j}^t$.

$$w_j^{(t)} \sim \text{Pois} \left[-\Phi_w^{(t)} \theta_{w-j}^{(t)} \ln \left(1 - p_j^{(t)} \right) \right], \tag{4}$$

$$v_j^{(t)} \sim \text{Pois} \left[-\Phi_v^{(t)} \theta_{v-j}^{(t)} \ln \left(1 - p_j^{(t)} \right) \right]. \tag{5}$$

Proof. The definition (4), (5) are absolutely true for layer one. Assume that (4), (5) are true for layer $t \geq 2$, then each textual count $w_{ij}^{(t)}$ and visual count $v_{ij}^{(t)}$ are separately augmented into the summation of K_t latent textual and visual counts. Thus, the summation of K_t two different latent counts is smaller than or equal to $w_{ij}^{(t)}$ and $v_{ij}^{(t)}$.

$$w_{ij}^{(t)} = \sum_{k=1}^{K_t} w_{ijk}^{(t)}, \quad w_{ijk}^{(t)} \sim \text{Pois} \left[-\phi_{w-ik}^{(t)} \theta_{w-kj}^{(t)} \ln(1 - p_j^{(t)}) \right],$$

$$v_{ij}^{(t)} = \sum_{k=1}^{K_t} v_{ijk}^{(t)}, \quad v_{ijk}^{(t)} \sim \text{Pois} \left[-\phi_{v-ik}^{(t)} \theta_{v-kj}^{(t)} \ln(1 - p_j^{(t)}) \right].$$

where $i \in \{1, \dots, K_{t-1}\}$. Then, we have

$$m_{kj}^{(t)(t+1)} = w_{.jk}^{(t)} = \sum_{i=1}^{K_{t-1}} w_{ijk}^{(t)}, \quad m_j^{(t)(t+1)} = \left(w_{.j1}^{(t)}, \dots, w_{.jK_t}^{(t)} \right)',$$

$$n_{kj}^{(t)(t+1)} = v_{.jk}^{(t)} = \sum_{i=1}^{K_{t-1}} v_{ijk}^{(t)}, \quad n_j^{(t)(t+1)} = \left(v_{.j1}^{(t)}, \dots, v_{.jK_t}^{(t)} \right)'.$$

$m_{kj}^{(t)(t+1)}$ denotes the counts in layer t that factor $k \in \{1, \dots, K_t\}$ appears in document j , and $v_{kj}^{(t)(t+1)}$ represents the counts in layer t that factor $k \in \{1, \dots, K_t\}$ appears in document j . On account of $\sum_{i=1}^{K_{t-1}} \phi_{w-ik}^{(t)} = 1$ and $\sum_{i=1}^{K_{t-1}} \phi_{v-ik}^{(t)} = 1$, we utilize the method in [31] to marginalize out Φ_w^t and Φ_v^t . As a result,

$$m_j^{(t)(t+1)} \sim \text{Pois} \left[-\theta_{w-j}^{(t)} \ln(1 - p_j^{(t)}) \right], \quad n_j^{(t)(t+1)} \sim \text{Pois} \left[-\theta_{v-j}^{(t)} \ln(1 - p_j^{(t)}) \right].$$

Then, by employing the above Poisson likelihood, we further marginalize out $\theta_{w-j}^{(t)}$ and $\theta_{v-j}^{(t)}$ that follows the gamma distribution.

$$m_j^{(t)(t+1)} \sim \text{NB} \left[\Phi_w^{(t+1)} \theta_{w-j}^{(t+1)}, p_j^{(t+1)} \right], \quad (6)$$

$$n_j^{(t)(t+1)} \sim \text{NB} \left[\Phi_v^{(t+1)} \theta_{v-j}^{(t+1)}, p_j^{(t+1)} \right]. \quad (7)$$

As demonstrated in [28], (6) and (9) can also be generated from their compound Poisson distribution as

$$m_{kj}^{(t)(t+1)} = \sum_{x=1}^{w_{kj}^{(t+1)}} u_x, \quad u_x \sim \text{Log}(p_j^{(t+1)}), \quad w_{kj}^{(t+1)} \sim \text{Pois} \left[\phi_{w-k}^{(t+1)} \theta_{w-j}^{(t+1)} \ln(1 - p_j^{(t+1)}) \right],$$

$$n_{kj}^{(t)(t+1)} = \sum_{y=1}^{v_{kj}^{(t+1)}} u_y, \quad u_y \sim \text{Log}(p_j^{(t+1)}), \quad v_{kj}^{(t+1)} \sim \text{Pois} \left[\phi_{v-k}^{(t+1)} \theta_{v-j}^{(t+1)} \ln(1 - p_j^{(t+1)}) \right].$$

Hence, if (4), (5) are true for layer t , they are also true for layer $t + 1$.

Inspired by the lemmas and theorems in [28,31], we propagate the latent textual counts $w_{ij}^{(t)}$ and visual counts $v_{ij}^{(t)}$ of layer t upward to layer $t+1$ as

$$\begin{aligned} & \left\{ \left(w_{ij1}^{(t)}, \dots, w_{ijK_t}^{(t)} \right) \mid w_{ij}^{(t)}, \phi_{w-i:}^{(t)}, \theta_{w-j}^{(t)} \right\} \\ \sim & \text{Mult} \left(w_{ij}^{(t)}, \frac{\phi_{w-i1}^{(t)} \theta_{w-1j}^{(t)}}{\sum_{k=1}^{K_t} \phi_{w-ik}^{(t)} \theta_{w-kj}^{(t)}}, \dots, \frac{\phi_{w-iK_t}^{(t)} \theta_{w-K_tj}^{(t)}}{\sum_{k=1}^{K_t} \phi_{w-ik}^{(t)} \theta_{w-kj}^{(t)}} \right), \end{aligned} \quad (8)$$

$$\left(w_{kj}^{(t+1)} \mid m_{kj}^{(t)(t+1)}, \phi_{w-k:}^{(t+1)}, \theta_{w-j}^{(t+1)} \right) \sim \text{CRT} \left(m_{kj}^{(t)(t+1)}, \phi_{w-k:}^{(t+1)}, \theta_{w-j}^{(t+1)} \right), \quad (9)$$

$$\begin{aligned} & \left\{ \left(v_{ij1}^{(t)}, \dots, v_{ijK_t}^{(t)} \right) \mid v_{ij}^{(t)}, \phi_{v-i:}^{(t)}, \theta_{v-j}^{(t)} \right\} \\ \sim & \text{Mult} \left(v_{ij}^{(t)}, \frac{\phi_{v-i1}^{(t)} \theta_{v-1j}^{(t)}}{\sum_{k=1}^{K_t} \phi_{v-ik}^{(t)} \theta_{v-kj}^{(t)}}, \dots, \frac{\phi_{v-iK_t}^{(t)} \theta_{v-K_tj}^{(t)}}{\sum_{k=1}^{K_t} \phi_{v-ik}^{(t)} \theta_{v-kj}^{(t)}} \right), \end{aligned} \quad (10)$$

$$\left(v_{kj}^{(t+1)} \mid n_{kj}^{(t)(t+1)}, \phi_{v-k:}^{(t+1)}, \theta_{v-j}^{(t+1)} \right) \sim \text{CRT} \left(n_{kj}^{(t)(t+1)}, \phi_{v-k:}^{(t+1)}, \theta_{v-j}^{(t+1)} \right). \quad (11)$$

3.3 Parameter Inference

In conventional topic models, variational inference and collapsed Gibbs sampling are often used for parameter inference. To estimate the latent variables under the multivariate observations, we utilize upward-downward Gibbs sampling [29] with the width of the first layer being restricted to K_{1max} . The sampling process of mmUAM is as below.

Sample $w_{ijk}^{(t)}$ and $v_{ijk}^{(t)}$. For all the layers, we can use (10) to sample $w_{ijk}^{(t)}$ and (12) to sample $v_{ijk}^{(t)}$. But for the first hidden layer, the observed counts $w_{ij}^{(1)}$ is considered as word tokens at the i th term in the j th document, where the size of textual vocabulary is denoted as $I = K_0$, and the observed counts $v_{ij}^{(1)}$ is treated as visual word tokens at the i th term (the size of visual vocabulary $I' = K'_0$) in the j th document. We define z_{w-j_s} and $z_{v-j_{s'}}$ as the topic index for i_{j_s} and $i_{j_{s'}}$ ($s \in \{1, \dots, w_j^{(1)}\}$, $s' \in \{1, \dots, v_j^{(1)}\}$).

$$P(z_{w-j_s} = k \mid -) \propto \frac{\eta^{(1)} + (w_{i_{j_s} \cdot k}^{(1)})_{-j_s}}{I\eta^{(1)} + (w_{\cdot k}^{(1)})_{-j_s}} \left((w_{\cdot jk}^{(1)})_{-j_s} + \phi_{w-k:}^{(2)} \theta_{w-j}^{(2)} \right), \quad (12)$$

$$P(z_{v-j_{s'}} = k \mid -) \propto \frac{\xi^{(1)} + (v_{i_{j_{s'}} \cdot k}^{(1)})_{-j_{s'}}}{I'\xi^{(1)} + (v_{\cdot k}^{(1)})_{-j_{s'}}} \left((v_{\cdot jk}^{(1)})_{-j_{s'}} + \phi_{v-k:}^{(2)} \theta_{v-j}^{(2)} \right). \quad (13)$$

where $k \in \{1, \dots, K_{1max}\}$. We let $w_{ijk}^{(1)} = \sum_s \delta(i_{j_s} = i, z_{w-j_s} = k)$ and $v_{ijk}^{(1)} = \sum_{s'} \delta(i_{j_{s'}} = i, z_{v-j_{s'}} = k)$. $w_{ijk}^{(1)}$ and $v_{ijk}^{(1)}$ represent the number of times when

term i is assigned to the topic k in document j . Besides, we use w_{-js} and v_{-js} to separately denote the count of w and v except when term i appears in document j . Especially when $T = 1$, we use Poisson Factor Analysis (PFA) with gamma-negative binomial process [28] to replace $\phi_{w-k}^{(2)} \cdot \theta_{w-j}^{(2)}$ and $\phi_{v-k}^{(2)} \cdot \theta_{v-j}^{(2)}$ with r_k . For simplification, if $T = 1$, we set K_{1max} factors and let $r_k \sim Gam(\gamma_0/K_{1max}, 1/c_0)$.

Sample $\phi_{w-k}^{(t)}$. We sample the textual topic $\phi_{w-k}^{(t)}$ as

$$\left(\phi_{w-k}^{(t)} | -\right) \sim Dir\left(\eta^{(t)} + w_{1 \cdot k}^{(t)}, \dots, \eta^{(t)} + w_{K_{t-1} \cdot k}^{(t)}\right).$$

Sample $\phi_{v-k}^{(t)}$. In the same way, we sample the visual topic $\phi_{v-k}^{(t)}$ as

$$\left(\phi_{v-k}^{(t)} | -\right) \sim Dir\left(\xi^{(t)} + v_{1 \cdot k}^{(t)}, \dots, \xi^{(t)} + v_{K_{t-1} \cdot k}^{(t)}\right).$$

Sample $w_{ij}^{(t+1)}$ and $v_{ij}^{(t+1)}$. We sample $w_j^{(t+1)}$, $v_j^{(t+1)}$ separately using (9), (11).

Sample r . c_0 and γ_0 are sampled using (3), whose detailed introduction is in [28].

$$\left(r_v | -\right) \sim Gam\left(\gamma_0/K_T + w_i^{(T+1)} + v_i^{(T+1)}, c_0 - \sum_j \ln\left(1 - p_j^{(T+1)}\right)^{-1}\right).$$

Sample $\Theta_j^{(t)}$. Using the latent counts propagated upward and the gamma-Poisson conjugacy, we downward sample the hidden units Θ_j as

$$\left(\Theta_j^{(t)} | -\right) \sim Gam\left(\left[\begin{array}{c} \Phi_w^{(t+1)} \theta_{w-j}^{(t+1)} \\ \Phi_v^{(t+1)} \theta_{v-j}^{(t+1)} \end{array}\right] + \left[\begin{array}{c} m_j^{(t)(t+1)} \\ n_j^{(t)(t+1)} \end{array}\right], c_j^{(t+1)} - \ln\left(1 - p_j^{(t)}\right)^{-1}\right).$$

Sample $p_j^{(2)}$ and $c_j^{(t)}$. We calculate $p_j^{(t)}$ ($t \geq 3$) and $c_j^{(2)}$ using (6), and sample $p_j^{(2)}$ and $c_j^{(t)}$, where $t \geq 3$ as

$$\begin{aligned} \left(p_j^{(2)} | -\right) &\sim Beta\left(a_0 + \left[\begin{array}{c} m_j^{(1)(2)} \\ n_j^{(1)(2)} \end{array}\right], b_0 + \Theta_j^{(2)}\right), \\ \left(c_j^{(t)} | -\right) &\sim Gam\left(e_0 + \Theta_j^{(t)}, \left[f_0 + \Theta_j^{(t-1)}\right]^{-1}\right). \end{aligned}$$

4 Experiments

4.1 Dataset Construction

As we know, Sina Weibo is one of the most popular social media platforms in China, where we collect a dataset and conduct our experiments on the real




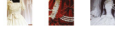


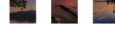



	layer 1	layer 2	layer 3	layer 4	layer 5
user1	 match, color, not bad, clothing brand, designing, skirt, photograph	 color, girl, brand, designing clothes, skirt, cute, clothing	 match, color, suit, designing brand, girl, skirt, clothes	 brand, beautiful, suit, girl model, photograph, clothes, color	 beautiful, girl, pretty, suit entirety, brand, photograph, effect
user2	 travel, photography, geography, shooting scenery, photo, composition, vision	 photography, shooting, scenery, city travel, camera, Shanghai, scene	 travel, shoot, frame, scene beautiful, mountain, photography, view	 photography, shoot, scenery, vision travel, camera, beautiful, scene	 shooting, scenery, beautiful, peaceful sight, environment, plant, scenery

Fig. 2. The generated visual words and textual keywords for the mmUAM-1 of five different layers.

data. We crawl 1349 users, the data including users’ basic information and their posted microblogs from January 2015 to December 2016, in which each microblog contains both texts and images. After filtering inactive users and separating each user’s microblogs into two documents according to posting time, we have 193798 documents. In order to comprehensively evaluate the generated user attributes, we utilize both the crawled tags and the posted microblogs to manually label the users’ interests.

For preprocessing the textual dataset, we firstly utilize jieba participle³ to segment the Chinese words, and then we eliminate the non-Chinese characters, stop words, and the low-frequency words that appear less than five times. For visual feature description, Scale-Invariant Feature Transform (SIFT) [16] is used to extract discriminant local features of images, thus generating 128-dimensional SIFT descriptors. To construct a codebook of visual words, we utilize k-means with each descriptor being a cluster center quantized into a visual word. As a result, each image is represented with the count of visual words in the codebook.

4.2 Evaluation Metrics and Compared Methods

The standard classification algorithm evaluation methods like precision, recall and F1-measure would not be sufficient to understand the performance of multi-label problems. Thus, we adopt the following evaluation measures proposed in [21], where we set $X = \{x_1, x_2, \dots, x_k\}$ as a set of output attributes and $Y = \{y_1, y_2, \dots, y_k\}$ as a set of ground-truth attributes.

Average Precision. We use average precision to calculate the mean value of ranking H of ground-truth attributes in predicted attributes. $|N|$ is the number of documents.

$$ap(H) = \frac{1}{N} \sum_{i=1}^N \frac{1}{|Y_i|} \sum_{y \in Y_i} \frac{\left| \left\{ y' \in Y_i \mid rank_f(x_i, y') \leq rank_f(x_i, y) \right\} \right|}{rank_f(x_i, y)}. \quad (14)$$

One Error. The measure evaluates how many times the top-ranked predicted attributes were not in the set of possible attributes Y . We express one-error of a hypothesis f as $one - err(f)$.

³ <https://github.com/fxsjy/jieba>.

$$oe(f) = \frac{1}{N} \sum_{i=1}^N \{[\text{argmax}_{y \in Y} f(x_i, y)] \notin Y_i\}. \quad (15)$$

Ranking Loss. This evaluation is to minimize the average fraction of crucial pairs which are misordered.

$$rl = \frac{1}{N} \sum_{i=1}^N \frac{1}{|Y_i| |\overline{Y}_i|} \left| \left\{ (y, y') \mid f(x_i, y) \leq f(x_i, y'), (y, y') \in Y_i \times \overline{Y}_i \right\} \right|. \quad (16)$$

Coverage. The coverage measures in a sequence queue, to go down the list of predicted attributes in order to cover all the possible attributes assigned to a document.

$$co(f) = \frac{1}{N} \sum_{i=1}^N \text{maxrank}_f(x_i, y) - 1. \quad (17)$$

To demonstrate the effectiveness of our proposed mmUAM, we compare our algorithm with the following methods.

Poisson Gamma Belief Network (PGBN). The PGBN is applied on our crawled Sina Weibo dataset for short texts.

Multi-modal User Attribute Model (mmUAM). We adapt the multi-modal fusion in five layers, separately. The ways of five different fusion are denoted as mmUAM-1, mmUAM-2, mmUAM-3, mmUAM-4, mmUAM-5. Specifically, mmUAM-1 is expressed as the shared representation for texts and images learned in the first layer, and other mmUAMs are denoted in the same way.

4.3 Experimental Results

In this paper, we employ the layer-wise training method for the mmUAMs, with which we set a fixed budget on the width of layer one $K_{1max} = 400$ and the depth of the network $T = 5$. Besides, we set hyper-parameters as $e_0 = f_0 = 1$, $a_0 = b_0 = 0.01$, and $\eta^{(t)} = \xi^{(t)} = 0.05$ for all the layers.

As the mmUAM learns the shared representation of textual and visual information, we do the qualitative analysis to show the effectiveness of our proposed model. We randomly selected two users to analyze the topics generated by the mmUAM-1 of five different layers. As an example, we choose 3 generated visual words and the top 6 textual words of the specific topic showed in Fig. 2. We can see that the visual representation and textual representation strengthen the description of user attributes. Obviously, the mmUAM can effectively model the semantic correlations of social data in multiple modalities.

We compare our mmUAM with PGBN and different ways of layer-fusion to evaluate the quality of our generating user attributes. Figure 3 displays the performance in term of average precision, one error, ranking loss, coverage. On the crawled Sina Weibo dataset, our proposed mmUAM with five ways of multi-modal fusion all performs better than the PGBN. The results also confirm the fact that multi-modal topic modeling works better than unimodal

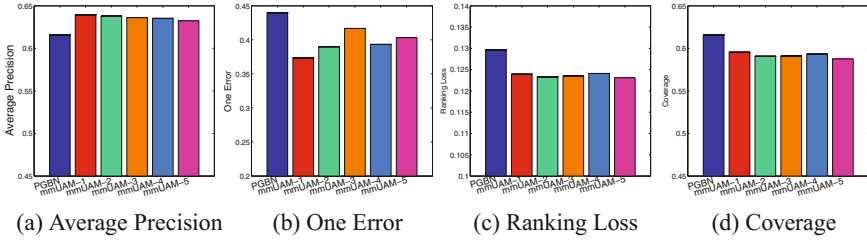


Fig. 3. The performance is evaluated by average precision, one error, ranking loss, coverage, respectively.

approaches. As for mmUAMs, there is slight difference among results on the four evaluation metrics. Especially, mmUAM-1 achieves the best result of the average precision and the one error. For the evaluation of the ranking loss and the coverage, mmUAM-5 achieves the lowest results that represent the best quality of the classification. As a result, the first-layer-fusion and the last-layer-fusion capture better correlations between texts and images than other layer-fusion.

5 Conclusion

In this paper, we proposed a novel multi-modal user attribute (mmUAM) model which automatically generated user interests from multi-modal social media data, by capturing correlations between textual and visual information. In particular, we improved the PGBN network to extract topics of interests better in line with users' characteristics. We conducted experiments on our crawled microblog dataset, where the results demonstrated the superiority of our mmUAM as compared to the state-of-the-art methods.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Project 61572108 and Project 61502081, the National Thousand-Young-Talents Program of China, and the Fundamental Research Funds for the Central Universities under Project ZYGX2014Z007 and Project ZYGX2015J055.

References

1. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Semantic enrichment of Twitter posts for user profile construction on the social web. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., Leenheer, P., Pan, J. (eds.) *ESWC 2011*. LNCS, vol. 6644, pp. 375–389. Springer, Heidelberg (2011). doi:[10.1007/978-3-642-21064-8_26](https://doi.org/10.1007/978-3-642-21064-8_26)
2. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. *JMLR* **3**, 1107–1135 (2003)
3. Bian, J., Yang, Y., Chua, T.S.: Multimedia summarization for trending topics in microblogs. In: *ACM CIKM*, pp. 1807–1812 (2013)

4. Bian, J., Yang, Y., Chua, T.S.: Predicting trending messages and diffusion participants in microblogging network. In: ACM SIGIR, pp. 537–546 (2014)
5. Bian, J., Yang, Y., Zhang, H., Chua, T.S.: Multimedia summarization for social events in microblog stream. *IEEE Trans. Multimedia* **17**(2), 216–228 (2015)
6. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: ACM SIGIR, pp. 127–134 (2003)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *JMLR* **3**, 993–1022 (2003)
8. Cheng, X., Yan, X., Lan, Y., Guo, J.: BTM: topic modeling over short texts. *IEEE TKDE* **26**(12), 2928–2941 (2014)
9. Geng, X., Zhang, H., Song, Z., Yang, Y., Luan, H., Chua, T.S.: One of a kind: User profiling by social curation. In: ACM MM, pp. 567–576 (2014)
10. He, W., Liu, H., He, J., Tang, S., Du, X.: Extracting interest tags for non-famous users in social network. In: ACM CIKM, pp. 861–870 (2015)
11. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. *Neural Comput.* **18**(7), 1527–1554 (2006)
12. Huang, X., Yang, Y., Hu, Y., Shen, F., Shao, J.: Dynamic user attribute discovery on social media. In: Li, F., Shim, K., Zheng, K., Liu, G. (eds.) APWeb 2016. LNCS, vol. 9931, pp. 256–267. Springer, Cham (2016). doi:[10.1007/978-3-319-45814-4_21](https://doi.org/10.1007/978-3-319-45814-4_21)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
14. Li, X., Cheung, M., She, J.: Connection discovery using shared images by gaussian relational topic model. [arxiv:1612.03639](https://arxiv.org/abs/1612.03639) (2016)
15. Liu, S., Cui, P., Zhu, W., Yang, S.: Learning socially embedded visual representation from scratch. In: ACM MM, pp. 109–118 (2015)
16. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV, vol. 2, pp. 1150–1157 (1999)
17. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML, pp. 689–696 (2011)
18. Pang, L., Ngo, C.W.: Multimodal learning with deep Boltzmann machine for emotion prediction in user generated videos. In: ICMR, pp. 619–622 (2015)
19. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: UAI, pp. 487–494 (2004)
20. Salakhutdinov, R., Hinton, G.E.: Deep boltzmann machines. In: AISTATS, vol. 1, p. 3 (2009)
21. Schapire, R.E., Singer, Y.: Boostexter: a boosting-based system for text categorization. *Mach. Learn.* **39**(2–3), 135–168 (2000)
22. Srivastava, N., Salakhutdinov, R.: Learning representations for multimodal data with deep belief nets. In: ICML Workshop (2012)
23. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep Boltzmann machines. In: NIPS, pp. 2222–2230 (2012)
24. Xu, Z., Ru, L., Xiang, L., Yang, Q.: Discovering user interest on twitter with a modified author-topic model. In: IEEE/WIC/ACM WI-IAT, vol. 1, pp. 422–429 (2011)
25. Yang, X., Li, Y., Luo, J.: Pinterest board recommendation for twitter users. In: ACM MM, pp. 963–966 (2015)
26. Yang, Y., Zha, Z.J., Gao, Y., Zhu, X., Chua, T.S.: Exploiting web images for semantic video indexing via robust sample-specific loss. *IEEE Trans. Multimedia* **16**(6), 1677–1689 (2014)
27. Zhao, Y., Liang, S., Ren, Z., Ma, J., Yilmaz, E., de Rijke, M.: Explainable user clustering in short text streams. In: ACM SIGIR, pp. 155–164 (2016)

28. Zhou, M., Carin, L.: Negative binomial process count and mixture modeling. *IEEE T PAMI* **37**(2), 307–320 (2015)
29. Zhou, M., Cong, Y., Chen, B.: The poisson gamma belief network. In: *NIPS*, pp. 3043–3051 (2015)
30. Zhou, M., Cong, Y., Chen, B.: Augmentable gamma belief networks. *JMLR* **17**(163), 1–44 (2016)
31. Zhou, M., Hannah, L., Dunson, D.B., Carin, L.: Beta-negative binomial process and poisson factor analysis. In: *AISTATS*, vol. 22, pp. 1462–1471 (2012)