Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Mining User Interests from Personal Photos

Pengtao Xie, Yulong Pei, Yuan Xie and Eric Xing
{pengtaox,epxing}@cs.cmu.edu,{yulongp, yxie1}@andrew.cmu.edu
School of Computer Science, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213

Abstract

Personal photos are enjoying explosive growth with the popularity of photo-taking devices and social media. The vast amount of online photos largely exhibit users' interests, emotion and opinions. Mining user interests from personal photos can boost a number of utilities, such as advertising, interest based community detection and photo recommendation. In this paper, we study the problem of user interests mining from personal photos. We propose a User Image Latent Space Model to jointly model user interests and image contents. User interests are modeled as latent factors and each user is assumed to have a distribution over them. By inferring the latent factors and users' distributions, we can discover what the users are interested in. We model image contents with a four-level hierarchical structure where the layers correspond to themes, semantic regions, visual words and pixels respectively. Users' latent interests are embedded in the theme layer. Given image contents, users' interests can be discovered by doing posterior inference. We use variational inference to approximate the posteriors of latent variables and learn model parameters. Experiments on 180K Flickr photos demonstrate the effectiveness of our model.

Introduction

With the prosperity of photo-taking devices such as digital cameras and smart phones, people habitually take photos to record interesting stuff and memorable events in their daily life. Everyday, millions of photos are uploaded to photo sharing social networks, like Flickr, Pinterest and Instagram. Personal photos reveal people's interests explicitly or implicitly. For instance, people loving pets tend to shoot a lot of dog and cat images and share them on social media. People enjoying food frequently populate their online albums with various food images. Figure 1(a) shows photos of four Flickr users. Browsing these photos, we can easily figure out that the first user likes cars, the second user is fond of flowers, the third user loves football and the fourth user enjoys food. A picture is worth a thousand words. Compared with texts, images are more natural to express users' interests and emotion. Mining users' interests from their personal photos

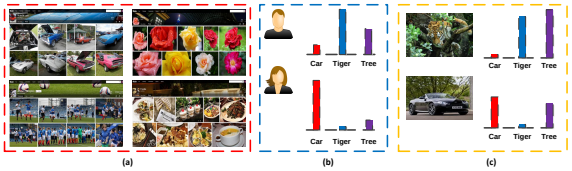

Figure 1: (a) Personal photos of four Flickr users. These photos clearly exhibit what these users show interest in. For example, from these photos, we can conjecture that the first user loves cars and the second user is fond of flowers. (b) Each user has a distribution over latent user interests. By identifying these distributions, we can discover users' interests. For instance, from the learned distributions, we can see that the man likes tiger and tree, but shows little interest in car. The woman loves car and is indifferent to tiger and tree. (c) We assume image contents are generated from latent interests. For example, the first image is more likely to be generated from tiger and tree interests while the second image is likely to be generated from car and tree.

can boost a number of utilities, such as advertising, interest based community detection, photo recommendation, to name a few. Taking Figure 1(a) as an example, if we can discover the personal interests of the four users from their albums, we can display car ads to the first user, recommend users who also like flowers to the second user, recommend sports news to the third user and recommend popular food to the fourth user.

While a lot of works have been devoted to mining users' interest from texts, links, click data and social information (Qiu and Cho 2006; Li et al. 2008; White, Bailey, and Chen 2009; Wen and Lin 2010; Kim et al. 2012; Hong, Doumith, and Davison 2013; Wang et al. 2013), image based user interests mining is largely unexplored. Feng and Qian (Feng and Qian 2013; 2014), Wang *et al* (Wang et al. 2009) leveraged both personal photos and their associated textual information such as tags and comments to discover users' interest. Their methods are not applicable if the texts data are absent.

In this paper, we build a model which can mine users' interests directly from their personal photos and does not reply on any text information. We propose a User Image La-

tent Space Model to jointly model user interests and image contents. User interests are modeled as latent factors. Each user has a distribution over these latent interests. For example, in Figure 1(b), there exist three interests: car, tiger and tree. The first user likes tiger and tree a lot, thereby, has high probabilities over tiger and tree. The second user loves car, hence, has high probabilities over car. If these latent interests and users' distributions over them can be identified, we can discover what the users are interested in. We learn these latent interests from users' personal photos by defining a probabilistic generative model where image contents are generated from users' interests. For example, in Figure 1(c), the first image is likely to be generated from two latent factors: tiger and tree; the second image is likely to be generated from car and tree. In the image model, image content is organized into a four-level hierarchical structure: themes, semantic regions, visual words and pixels. To incorporate the spatial coherence of neighboring pixels, we define Markov Random Field on latent layers to encourage nearby pixels to share the same label. We use variational inference to approximate the posteriors of latent variables and learn model parameters.

The major contribution of this paper is summarized as follows:

- We propose a latent space model which can directly mine users' interest from personal photos, without any requirement of text information.
- We evaluate our model on 180K Flickr photos. Qualitative and quantitative analysis both demonstrate the effectiveness of our methods.

The rest of the paper is organized as follows. Section 2 reviews related work. In Section 3, we introduce the model and inference technique. Section 4 presents experimental results and Section 5 concludes the paper.

Related Works

Many works have been proposed to mine user interests. Qiu and Cho (Qiu and Cho 2006) detected user interests from their past search histories. Li *et al* (Li et al. 2008) tried to mine user interests from the locations they visited. White *et al* (White, Bailey, and Chen 2009) propose to model user interests based on contextual information including social, historic, task, collection, and user interaction. Wang *et al* (Wang et al. 2013), Wen and Lin (Wen and Lin 2010) proposed to infer user interests from users' social connections and interactions. Kim *et al* (Kim et al. 2012) characterized user interests by reading level and topic distributions. Hong *et al* (Hong, Doumith, and Davison 2013) modeled users' interests by analyzing information gathered from Twitter, such as tweets, hash tags, followers. All of these works focus on texts, links, clicks, meta data and social clues. Personal photos, as a crucial medium to convey user interests, have been largely ignored nevertheless. Wang *et al* (Wang et al. 2009) investigated photo based interest mining. However, they use standalone image annotation tools to transform images into textual tags and subsequently extract interests from tags. Thereby, this work is essentially still text based. Feng and Qian (Feng and Qian 2013; 2014) leveraged personal photos and their associated texts such as photo tags and comments to mine users' interests. These methods rely heavily on textual information and are not applicable when the side texts are absent. Our method directly mines users' interests from images and does not require the presence of any textual information.

Our work is also closely related with image modeling (Fei-Fei and Perona 2005; Sivic et al. 2005; Russell et al. 2006; Cao and Fei-Fei 2007; Verbeek and Triggs 2007; Niebles, Wang, and Fei-Fei 2008; Zhao, Fei-Fei, and Xing 2010). Fei-Fei and Perona (Fei-Fei and Perona 2005) proposed a Bayesian hierarchical model for learning natural scene categories. Niebles *et al* (Niebles, Wang, and Fei-Fei 2008), Russell *et al* (Russell et al. 2006), Sivic *et al* (Sivic et al. 2005) used probabilistic latent semantic analysis (pLSA) (Hofmann 1999) and Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) to learn action categories and detect object categories. Cao and Fei-Fei (Cao and Fei-Fei 2007) proposed a spatially coherent latent topic model which enforces spatial coherency by assigning only one single latent topic to all patches in each image region. Verbeek and Triggs (Verbeek and Triggs 2007), Zhao *et al* (Zhao, Fei-Fei, and Xing 2010) combined topic models (Hofmann 1999; Blei, Ng, and Jordan 2003) with Markov Random Field to do region classification and image segmentation. A common property of all the above mentioned models is that image contents are assumed to be generated from some latent factors and high-level semantic information is to be discovered by inferring the latent factors. Our model is devised in the same spirit, but with a deep structure, which covers high level semantic information and low level visual information. Deep learning models (Le et al. 2012) is able to learn deep hierarchical representations of images, however, they lack the flexibility to incorporate side information such as users or to incorporate the spatial coherence of pixels. Our model is the first one simultaneously modeling users and images, with the goal to discover user interests.

Model

We propose a User Image Latent Space Model (UILSM) to jointly model user interests and image contents. By inferring the latent variables of this model, we can discover users' interests. We use variational inference to approximate the posteriors of latent variables and learn model parameters. In this section, we first describe how to model images, then based on that, we propose the User Image Latent Space Model.

Image Modeling

Image content can be organized into a four-level hierarchical structure: themes, semantic regions, visual words and pixels, from top to bottom. Themes and semantic regions are high-level information units while visual words and pixels are low-level information units. Themes (or topics) reflect the central semantics of an image, such as football game, party, campus and so on. For example, the photo in Figure 2 contains two themes: picnic and nature. A semantic region is a set of connected pixels which depict a meaningful concept. It can be a scene region (e.g., sky, ocean) or an object (e.g.,

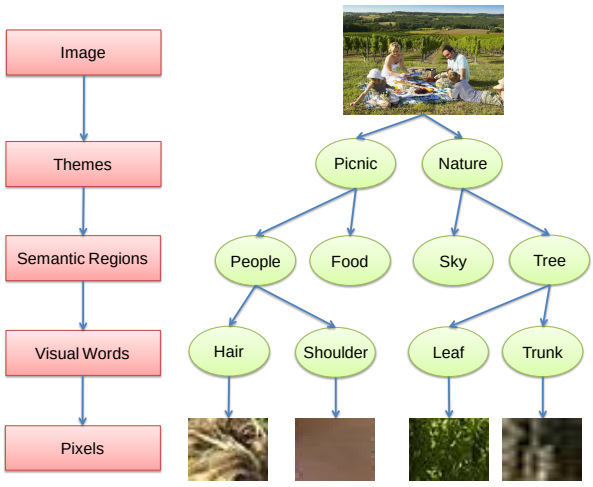

Figure 2: Hierarchical structure of image content. We model image content using a hierarchical structure which contains four levels of elements: themes, semantic regions, visual words and pixels. Higher level element is composed of lower level elements. The exemplar image contains two themes: people and nature. The picnic theme is composed of two semantic regions: people and food. The people region consists of two kinds of visual words: hair and shoulder.

bird, computer). For instance, in Figure 2, the photo is comprised of scenes like sky, trees, grass and objects like people, food, canvas, etc. A theme may be characterized by multiple semantic regions. In Figure 2, the picnic theme is revealed by objects of people, food, bottle, cup and canvas while the nature theme is reflected by scenes like sky, forest and grass. A semantic region is semantically coherent, but not necessarily visually consistent. For instance, while a human face can be deemed as a semantic region, within human face there exist multiple visual elements like eye, nose, cheek, mouth, which are visually quite heterogeneous from each other. In other words, each semantic region is composed of a set of homogeneous visual "words" (elements). A visual word is a set of connected pixels which are highly similar in appearance. The women object of Figure 2 consists of four visual words: blond hair, shoulder, black eye and white dress. The tree region is composed of visual words like leaf and trunk. On the bottom of the hierarchy are the observed pixels. From the bottom-up view, lower level information units construct higher level information units layer by layer: pixels form visual words, visual words form semantic region and semantic regions form theme.

One salient characteristic of image content is spatial coherence. Neighboring pixels are likely to be from the same visual word, the same semantic region and the same theme. In image model design, it is indispensable to ensure the spatial coherence.

User Image Latent Space Model

The User Image Latent Space Model (UILSM) is depicted in Figure 3. In image collection, we assume themes are of two categories: background themes and user interested themes. The assumption is based on two observations. First, not all photos are taken because they appeal to users' interest. Second, even in a photo containing interesting stuff, background scenes inevitably occur a lot. For user-interested themes, we assume each user has a unique profile distribution over them, to emphasize that different users have different interests.

Suppose we are given U users and each user j possesses N_j personal photos. There exist T_b background themes and T_u user-interested themes. Each theme has a multinomial distribution $\boldsymbol{\beta}$ over semantic regions. Suppose the image collection contains O semantic regions and W visual words. Each semantic region has a multinomial distribution $\boldsymbol{\gamma}$ over visual words and each visual word has a multivariate Gaussian distribution $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ over the feature descriptors of pixels. Each user has a Dirichlet distribution $\boldsymbol{\alpha}^{(u)}$ over user interested themes and all images share a Dirichlet distribution $\boldsymbol{\alpha}^{(b)}$ over the background themes. Each image has a multinomial distribution $\boldsymbol{\theta}^{(u)}$ over user-interested themes and a multinomial distribution $\boldsymbol{\theta}^{(b)}$ over background themes. $\boldsymbol{\theta}^{(u)}$ is sampled from $\boldsymbol{\alpha}^{(u)}$ corresponding to the user to whom the image belongs to and $\boldsymbol{\theta}^{(b)}$ is sampled from $\boldsymbol{\alpha}^{(b)}$. Each pixel in the image can be either generated from a background theme or a user-interested theme. To make this binary decision, we introduce a Bernoulli variable δ for each pixel. $\delta = 1$ denotes that the pixel is generated from a user-interested theme and $\delta = 0$ denotes that the pixel is generated from a background theme. δ is generated from Bernoulli distribution ω and ω is sampled from Beta distribution $\boldsymbol{\zeta}$. Each image has a unique ω to emphasize the fact that some images contain more user-interested stuff while some contain less. All images share a single $\boldsymbol{\zeta}$.

First we describe how pixels in an image can be generated independently without considering the spatial relationship among them. Given an image, we sample a Bernoulli distribution ω from Beta prior $\boldsymbol{\zeta}$, sample a multinomial distribution $\boldsymbol{\theta}^{(b)}$ over background themes from Dirichlet prior $\boldsymbol{\alpha}^{(b)}$, sample a multinomial distribution $\boldsymbol{\theta}^{(u)}$ over user-interested themes from the Dirichlet prior $\boldsymbol{\alpha}^{(u)}$ corresponding to the user to whom this image belongs. Then, for each pixel, we sample binary variable δ from ω. If $\delta = 1$, we sample a user-interested theme t from $\boldsymbol{\theta}^{(u)}$; otherwise, sample a background theme t from $\boldsymbol{\theta}^{(b)}$. Given t, we sample a semantic region label o from the multinomial $\boldsymbol{\beta}$ corresponding to t. Given o, sample a visual word w from the multinomial $\boldsymbol{\gamma}$ corresponding to o. Finally, we sample the feature descriptor $\mathbf{p}$ of this pixel from the multivariate Gaussian distribution $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ corresponding to w.

As stated before, image content exhibits strong coherence. Neighboring pixels are likely to share the same label on binary decision layer, theme layer, semantic region layer and visual word layer. To encode the spatial coherence, similar to (Verbeek and Triggs 2007; Zhao, Fei-Fei, and Xing 2010), we define Markov Random Field (MRF) over each layer to encourage neighboring pixels to share the same label.

In UILSM, the generative process of an image belonging to user j can be summarized as follows.

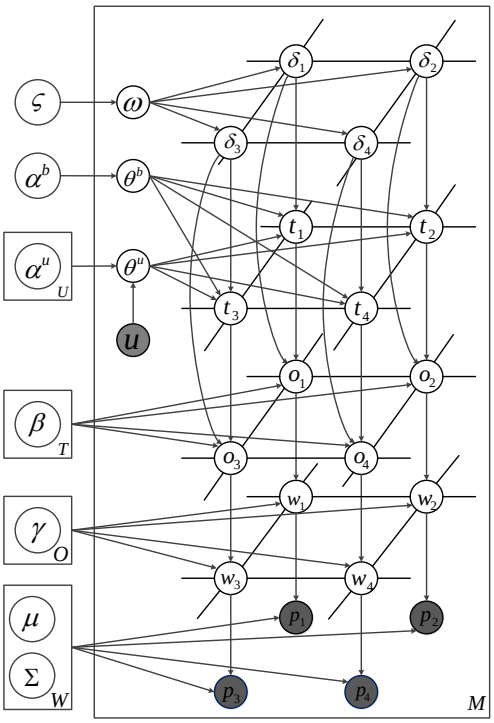

Figure 3: User Image Latent Space Model.

- Sample user interest proportion $\boldsymbol{\theta}^{(u)} \sim Dir(\alpha_j^{(u)})$
- Sample background proportion $\boldsymbol{\theta}^{(b)} \sim Dir(\alpha^{(b)})$
- Sample Bernoulli parameter $\omega \sim Beta(\boldsymbol{\zeta})$
- Sample binary decisions $\boldsymbol{\delta}$

$$p(\boldsymbol{\delta}|\omega, \lambda^{(\delta)}) = \frac{1}{A(\omega,\lambda^{(\delta)})} \prod_{n=1}^{N} p(\delta_n|\omega) \exp\{\lambda^{(\delta)} \sum_{(r,s)\in\mathcal{P}} \mathbb{I}(\delta_r = \delta_s)\} \tag{1}$$

- Sample theme labels $\mathbf{t}$

$$p(\mathbf{t}|\boldsymbol{\delta}, \boldsymbol{\theta}^{(u)}, \boldsymbol{\theta}^{(b)}, \lambda^{(t)}) = \frac{1}{B(\boldsymbol{\delta},\boldsymbol{\theta}^{(u)},\boldsymbol{\theta}^{(b)},\lambda^{(t)})} \prod_{n=1}^{N} p(t_n|\delta_n, \boldsymbol{\theta}^{(u)}, \boldsymbol{\theta}^{(b)}) \exp\{\lambda^{(t)} \sum_{(r,s)\in\mathcal{P}} \mathbb{I}(t_r = t_s)\} \tag{2}$$

- Sample object labels $\mathbf{o}$

$$p(\mathbf{o}|\boldsymbol{\delta}, \mathbf{t}, \boldsymbol{\beta}^{(u)}, \boldsymbol{\beta}^{(b)}, \lambda^{(o)})) = \frac{1}{C(\boldsymbol{\delta},\mathbf{t},\boldsymbol{\beta}^{(u)},\boldsymbol{\beta}^{(b)},\lambda^{(o)})} \prod_{n=1}^{N} p(o_n|\delta_n, t_n, \boldsymbol{\beta}^{(u)}, \boldsymbol{\beta}^{(b)}) \exp\{\lambda^{(o)} \sum_{(r,s)\in\mathcal{P}} \mathbb{I}(o_r = o_s)\} \tag{3}$$

- Sample visual word labels $\mathbf{w}$

$$p(\mathbf{w}|\mathbf{o}, \boldsymbol{\gamma}, \lambda^{(w)}) = \frac{1}{D(\mathbf{o},\boldsymbol{\gamma},\lambda^{(w)})} \prod_{n=1}^{N} p(w_n|o_n, \boldsymbol{\gamma}) \exp\{\lambda^{(w)} \sum_{(r,s)\in\mathcal{P}} \mathbb{I}(w_r = w_s)\} \tag{4}$$

- For each super pixel n, sample its descriptor $\mathbf{p}_n$

$$p(\mathbf{p}_n|w_n, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{W}) = \prod_{k=1}^{W} \mathcal{N}(\mathbf{p}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{w_{nk}} \tag{5}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\mathcal{P}$ denotes neighboring pixel pairs. A, B, C, D are partition functions. $\lambda^{(\delta)}$, $\lambda^{(t)}$, $\lambda^{(o)}$, $\lambda^{(w)}$ are tradeoff parameters indicating how much we emphasize the spatial coherence on each layer.

Inference and Learning

We employ variational inference (Wainwright and Jordan 2008) technique to approximate the posterior of latent variables and learn model parameters.

The variational distribution q is defined as

$$\begin{aligned} &q(\omega, \boldsymbol{\theta}^{(b)}, \boldsymbol{\theta}^{(u)}, \boldsymbol{\delta}, \mathbf{t}, \mathbf{o}, \mathbf{w}) = q(\omega|\boldsymbol{\xi}) q(\boldsymbol{\theta}^{(b)}|\boldsymbol{\eta}^{(b)}) \\ &q(\boldsymbol{\theta}^{(u)}|\boldsymbol{\eta}^{(u)}) \prod_{n=1}^{N} q(\delta_n|\tau_n) q(t_n^{(u)}|\boldsymbol{\phi}_n^{(u)}) \\ &q(t_n^{(b)}|\boldsymbol{\phi}_n^{(b)}) q(o_n|\boldsymbol{\varphi}_n) q(w_n|\boldsymbol{\psi}_n) \end{aligned} \tag{6}$$

where $\boldsymbol{\xi}$ is Beta parameter. $\boldsymbol{\eta}^{(b)}$ and $\boldsymbol{\eta}^{(u)}$ are Dirichlet parameters. $\{\tau_n\}_{n=1}^{N}$ are Bernoulli parameters. $\{\boldsymbol{\phi}_n^{(u)}\}_{n=1}^{N}$,$\{\boldsymbol{\phi}_n^{(b)}\}_{n=1}^{N}$ $\{\boldsymbol{\varphi}_n\}_{n=1}^{N}$ and $\{\boldsymbol{\psi}_n\}_{n=1}^{N}$ are multinomial parameters. Given the variational distribution, we derive a variational lower bound and use EM algorithm to optimize it. In the E step, we fix model parameters and infer the variational variables.

$$\begin{aligned} \psi_{nj} \propto \exp\{ &\sum_{k=1}^{O} \varphi_{nk} \log \gamma_{kj} + \lambda^{(w)} \sum_{r\in\mathcal{N}(n)} \psi_{rj} \\ &+ \log \mathcal{N}(\mathbf{p}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\} \end{aligned} \tag{7}$$

$$\begin{aligned} \varphi_{nj} \propto \exp\{ &\tau_n \sum_{k=1}^{T_u} \phi_{nk}^{(u)} \log \beta_{kj}^{(u)} \\ &+(1-\tau_n) \sum_{k=1}^{T_b} \phi_{nk}^{(b)} \log \beta_{kj}^{(b)} \\ &+\lambda^{(o)} \sum_{r\in\mathcal{N}(n)} \varphi_{rj} + \sum_{k=1}^{W} \psi_{nk} \log \gamma_{jk}\} \end{aligned} \tag{8}$$

$$\begin{aligned} \phi_{nk}^{(u)} \propto \exp\{ &\tau_n(\Psi(\eta_k^{(u)}) - \Psi(\sum_{j=1}^{K} \eta_j^{(u)})) \\ &+\lambda^t \sum_{r\in\mathcal{N}(n)} \tau_n \tau_r \phi_{rk}^{(u)} + \tau_n \sum_{j=1}^{O} \varphi_{nj} \log \beta_{kj}^{(u)}\} \end{aligned} \tag{9}$$

$$\begin{aligned} \phi_{nk}^{(b)} \propto \exp\{ &(1-\tau_n)(\Psi(\eta_k^{(b)}) - \Psi(\sum_{j=1}^{K} \eta_j^{(b)})) \\ &+\lambda^t \sum_{r\in\mathcal{N}(n)} (1-\tau_n)(1-\tau_r) \phi_{rk}^{(b)} \\ &+(1-\tau_n) \sum_{j=1}^{O} \varphi_{nj} \log \beta_{kj}^{(b)}\} \end{aligned} \tag{10}$$

$$\tau_n = 1/(1+\exp(-h)) \tag{11}$$

$$h = \Psi(\xi_1) - \Psi(\xi_2) + \lambda^{(\delta)} \sum_{r \in \mathcal{N}(n)} (2\tau_r - 1)$$
$$+ \sum_{k=1}^{T_u} \phi_{nk}^{(u)} (\Psi(\eta_k^{(u)}) - \Psi(\sum_{j=1}^{K} \eta_j^{(u)}))$$
$$- \sum_{k=1}^{T_b} \phi_{nk}^{(b)} (\Psi(\eta_k^{(b)}) - \Psi(\sum_{j=1}^{K} \eta_j^{(b)}))$$
$$+ \lambda^{(t)} \sum_{r \in \mathcal{N}(n)} (\tau_r \sum_{k=1}^{T_u} \phi_{nk}^{(u)} \phi_{rk}^{(u)} + (\tau_r - 1) \sum_{k=1}^{T_b} \phi_{nk}^{(b)} \phi_{rk}^{(b)})$$
$$+ \sum_{k=1}^{T_u} \sum_{j=1}^{O} \phi_{nk}^{(u)} \varphi_{nj} \log \beta_{kj}^{(u)} - \sum_{k=1}^{T_b} \sum_{j=1}^{O} \phi_{nk}^{(b)} \varphi_{nj} \log \beta_{kj}^{(b)} \tag{12}$$

$$\eta_k^{(u)} = \alpha_k^{(u)} + \sum_{n=1}^{N} \tau_n \phi_{nk}^{(u)} \tag{13}$$

$$\eta_k^{(b)} = \alpha_k^{(b)} + \sum_{n=1}^{N} (1 - \tau_n) \phi_{nk}^{(b)} \tag{14}$$

$$\xi_1 = \zeta_1 + \sum_{n=1}^{N} \tau_n, \quad \xi_2 = \zeta_2 + \sum_{n=1}^{N} (1 - \tau_n) \tag{15}$$

In M step, we fix the variational variables and learn model parameters.

$$\beta_{kj}^{(u)} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \tau_n \phi_{nk}^{(u)} \varphi_{nj} \tag{16}$$

$$\beta_{kj}^{(u)} \propto \beta_{kj}^{(b)} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} (1 - \tau_n) \phi_{nk}^{(b)} \varphi_{nj} \tag{17}$$

$$\gamma_{kj} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \varphi_{nk} \psi_{nj} \tag{18}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \psi_{nk} \mathbf{p}_n}{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \psi_{nk}}, \quad \boldsymbol{\Sigma}_k = \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \psi_{nk} \mathbf{p}_n \mathbf{p}_n^{\mathsf{T}}}{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \psi_{nk}} \tag{19}$$

We optimize $\boldsymbol{\zeta}$, $\boldsymbol{\alpha}^b$ and $\{\boldsymbol{\alpha}_j^{(u)}\}_{j=1}^{U}$ using Newton's method. Instead of learning $\lambda^{(\delta)}$, $\lambda^{(t)}$, $\lambda^{(o)}$, $\lambda^{(w)}$ from data, we choose to empirically tune them.

Experimental Result

In this section, we evaluate our model qualitatively and quantitatively on Flickr dataset.

Experimental Settings

We crawl 183723 personal photos from 227 Flickr users. Each user has about 800 images. These images are quite diverse, covering topics like flower, car, bird, animal, people, etc. We oversegment each image into superpixels using the Simple Linear Iterative Clustering (SLIC) (Achanta et al. 2010) algorithm. SLIC clusters pixels in the combined five-dimensional color and image plane space to efficiently generate compact, nearly uniform superpixels. We depict each superpixel with color, texture (Li, Socher, and Fei-Fei 2009) and SIFT (Lowe 2004) based bag-of-words (BOW) features. The total dimension of descriptor is 107 (3-dim color, 4-dim texture, 100-dim BOW). We set the number of background themes to 100, the number of user-interested themes to 1000, the number of semantic regions to 1500 and the number of visual words to 500. The tradeoff parameters $\lambda^{(\delta)}$, $\lambda^{(t)}$, $\lambda^{(o)}$, $\lambda^{(w)}$ are all set to 1. The model is initialized randomly.

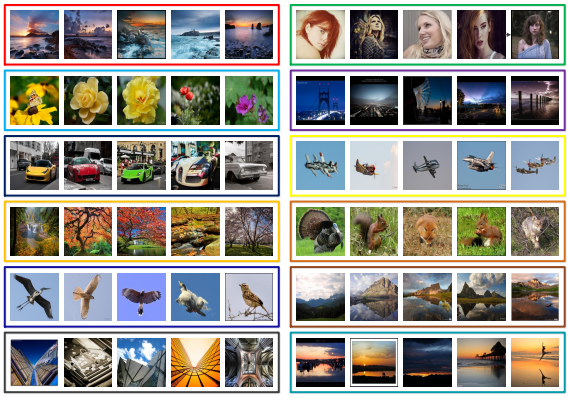

Figure 4: User-interested themes. These themes are about ocean, beautiful women, flower, night view, car, plane, tree, animal, bird, mountain, building, sunset respectively. They are likely to be interested by users. For example, people loving plants may show interests in the flower theme and tree theme.

Figure 5: Background themes. These themes contain street views and random stuff. They are unlikely to be interested by people. They show up in users' albums mostly due to random shots.

Qualitative Evaluation

In this section, we present qualitative results. For each learned theme, we visualize it with the most representative images. Figure 4 shows 12 user-interested themes. They correspond to ocean, beautiful women, flower, night view, car, plane, tree, animal, bird, mountain, building, sunset respectively. These themes are well-aligned with people's interests. For example, people who love traveling are fond of taking photos of ocean and mountains. Those who love plants prefer to take photos of flowers and trees. Figure 5 shows 6 background themes, which are unlikely to be interested by users. Basically, these themes correspond to street views and random stuff. The reason why user-interested themes and background themes can be distinguished is: images from background themes are likely to appear in the photostreams of many users while those from user-interested themes tend to show up in a few users' albums. Our model is capable to capture these frequency patterns to tell these two kinds of themes apart.

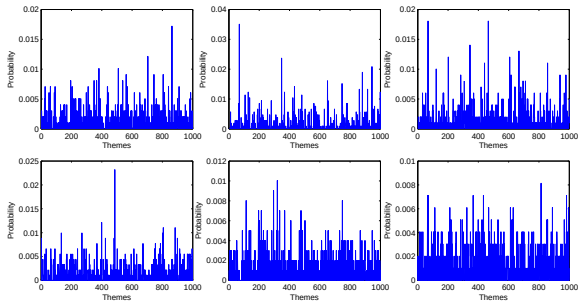

Figure 6: Users' distributions over user-interested themes. In each subfigure, the horizontal axis corresponds to user-interested themes. The vertical axis indicates the probability that the user likes a theme. Through analyzing these distributions, we can figure out what users are interested in.

Figure 6 shows the distributions over 1000 user-interested themes of 6 users. By analyzing each user's distribution over these themes, we can identify user's interest. For example, user 2 (row 1, column 2) has high probabilities over theme 69 and 348. Theme 69 corresponds to beautiful girls and theme 348 is about people in tribe. Observing this, we can conjecture that user 2 has great interests in people photos. And the conjecture is confirmed by inspecting his photo set, where most images are about people. In the distribution of user 4 (row 2, column 1), there is a sharp peak over theme 485, which is about painting. This strongly indicates that this user is a painting fan. Browsing his photo album, we observe a number of painting images. Thus, our model successfully identifies his interest over paintings.

Quantitative Evaluation

In this section, we evaluate our model from a quantitative perspective. For each of the 227 Flickr users in our experiment, we download 100 images which are marked as "favorites" by the user and we assume the user is interested in these images. We assemble these 22700 images into a test set and predict whether a user likes an image. The groundtruth is assumed as follows: if an image is in a user's favorite set (which contains 100 images), this image is supposed to be interested in by the user; otherwise, the image is not liked by the user. Given a user and an image, we predict whether this user likes this image or not in the following way: we infer both the user's distribution and the image's distribution over the 1000 user-interested themes and compare how similar these two distributions are; if they are similar enough, we predict that the user shows interest in this image. Image's distribution is given by the approximated posterior expectation $\eta^{(u)}$ of $\theta^{(u)}$ in Eq.(13), and user distribution is computed as the average of $\eta^{(u)}$ of all the images belonging to this user. Specifically, we use Euclidean distance to measure the similarity and set a threshold to make the decision. By varying the threshold, we obtain the recall-precision curve. We compare with two baseline methods: k-means and LDA. We use k-means to cluster all images into 1000 clusters and assume each cluster is a theme. We compute the distributions of users and images over the 1000 clusters and make predictions. For LDA, we represent images into bag-of-words and learn 1000 topics from them. Again, we make predictions based on the inferred distributions over the LDA topics.

Table 1: Average precision (AP) of user-interested image prediction on Flickr dataset

Method	K-means	LDA	Our method
AP	0.5563	0.5940	0.6907

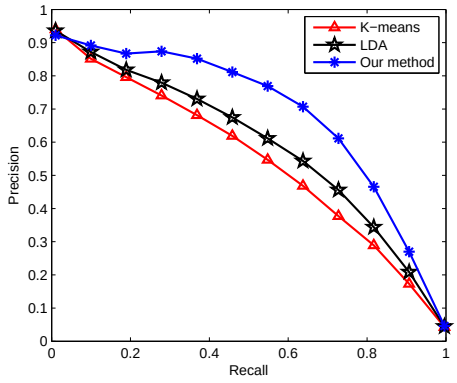

Figure 7: The precision-recall curves of predicting user-interested images on Flickr dataset.

Figure 7 compares the recall-precision curves of our method and two baseline methods. Table 1 summarizes the average precision (AP) of three methods. From the experimental results, we can see that our method is significantly better than the baselines. The average precision of our method is 10 percent higher than LDA and 13 percent higher than k-means. The superiority of our model is due to three reasons. First, our model seamlessly unifies user modeling and image modeling while k-means and LDA lack the mechanism to simultaneously model users and images. Jointly modeling users and images makes our model suitable for discovering user interests from image contents. Second, our model uses a four-level hierarchical structure to model image, which can capture the high-level semantics of images while k-means and LDA use shallow machineries to do the modeling, which are insufficient to discover complex patterns. Our model is reduced to LDA if we remove the theme layer and is reduced to a Gaussian Mixture Model (a probabilistic counterpart of K-means) if we remove both the theme and semantic region layers. This indicates the necessity and effectiveness of using a four-layer model. Third, our model is able to distinguish user-interested themes and background themes while the baselines lack this mechanism. This merit makes our model more robust to noise resulted from users' random shots.

Conclusion

In this paper, we study the problem of user interests mining from personal photos. We propose a User Image Latent Space Model to jointly model user interests and image con-

tents. By inferring the latent interests and users' distributions over them, we can discover what users are interested in. To model images, we organize image content into a four-level hierarchical structure: themes, semantic regions, visual words and pixels. MRFs are defined over latent layers to incorporate spatial coherence of neighboring pixels. Experiments on 180K photos belonging to 227 Flickr users qualitatively and quantitatively demonstrate the effectiveness of our model. Our model successfully identifies background themes and user-interested themes. By analyzing users' distributions over the latent themes, we can figure out what they show interests in. In the user-interested image prediction task, our method beats the baselines with a large margin.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions and thank Bin Zhao and Gunhee Kim for reviewing the preliminary draft. This work was supported by NSF IIS1111142, NSF IIS1447676 and AFOSR FA95501010247.

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2010. Slic superpixels. *École Polytechnique Fédéral de Lausssanne (EPFL), Tech. Rep* 149300.

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

Cao, L., and Fei-Fei, L. 2007. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*, 1–8. IEEE.

Fei-Fei, L., and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, volume 2, 524–531. IEEE.

Feng, H., and Qian, X. 2013. Recommend social network users favorite brands. In *Advances in Multimedia Information Processing*. Springer. 730–739.

Feng, H., and Qian, X. 2014. Mining user-contributed photos for personalized product recommendation. *Neurocomputing* 129:409–420.

Hofmann, T. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 289–296. Morgan Kaufmann Publishers Inc.

Hong, L.; Doumith, A. S.; and Davison, B. D. 2013. Co-factorization machines: modeling user interests and predicting individual decisions in twitter. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 557–566. ACM.

Kim, J. Y.; Collins-Thompson, K.; Bennett, P. N.; and Dumais, S. T. 2012. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 213–222. ACM.

Le, Q. V.; Monga, R.; Devin, M.; Corrado, G.; Chen, K.; MarcAurelio Ranzato, J. D.; and Ng, A. Y. 2012. Building high-level features using large scale unsupervised learning. In *ICML*.

Li, Q.; Zheng, Y.; Xie, X.; Chen, Y.; Liu, W.; and Ma, W.-Y. 2008. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, 34. ACM.

Li, L.-J.; Socher, R.; and Fei-Fei, L. 2009. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2036–2043. IEEE.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60(2):91–110.

Niebles, J. C.; Wang, H.; and Fei-Fei, L. 2008. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision* 79(3):299–318.

Qiu, F., and Cho, J. 2006. Automatic identification of user interest for personalized search. In *Proceedings of the 15th international conference on World Wide Web*, 727–736. ACM.

Russell, B. C.; Freeman, W. T.; Efros, A. A.; Sivic, J.; and Zisserman, A. 2006. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, volume 2, 1605–1614. IEEE.

Sivic, J.; Russell, B. C.; Efros, A. A.; Zisserman, A.; and Freeman, W. T. 2005. Discovering objects and their location in images. In *ICCV*, volume 1, 370–377. IEEE.

Verbeek, J., and Triggs, B. 2007. Region classification with markov field aspect models. In *CVPR*, 1–8. IEEE.

Wainwright, M. J., and Jordan, M. I. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1(1-2):1–305.

Wang, X.-J.; Yu, M.; Zhang, L.; Cai, R.; and Ma, W.-Y. 2009. Argo: intelligent advertising by mining a user's interest from his photo collections. In *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*, 18–26. ACM.

Wang, T.; Liu, H.; He, J.; and Du, X. 2013. Mining user interests from information sharing behaviors in social media. In *Advances in Knowledge Discovery and Data Mining*. Springer. 85–98.

Wen, Z., and Lin, C.-Y. 2010. On the quality of inferring interests from social neighbors. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 373–382. ACM.

White, R. W.; Bailey, P.; and Chen, L. 2009. Predicting user interests from contextual information. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 363–370. ACM.

Zhao, B.; Fei-Fei, L.; and Xing, E. P. 2010. Image segmentation with topic random field. In *ECCV*. Springer. 785–798.