Learning Geographical Hierarchy Features for Social Image Location Prediction

Xiaoming Zhang[†], **Xia Hu**[‡*], **Zhoujun Li**[†]
[†]State Key Laboratory of Software Development Environment, Beihang University, China
[‡]Department of Computer Science and Engineering, Texas A&M University, USA
[†]{yolixs, lizj}@buaa.edu.cn, [‡]huxia001@gmail.com

Abstract

Image location prediction is to estimate the geolocation where an image is taken. Social image contains heterogeneous contents, which makes image location prediction nontrivial. Moreover, it is observed that image content patterns and location preferences correlate hierarchically. Traditional image location prediction methods mainly adopt a single-level architecture, which is not directly adaptable to the hierarchical correlation. In this paper, we propose a geographically hierarchical bi-modal deep belief network model (GH-BDBN), which is a compositional learning architecture that integrates multi-modal deep learning model with non-parametric hierarchical prior model. GH-BDBN learns a joint representation capturing the correlations among different types of image content using a bi-modal DBN, with a geographically hierarchical prior over the joint representation to model the hierarchical correlation between image content and location. Experimental results demonstrate the superiority of our model for image location prediction.

1 Introduction

With the rising popularity of GPS-equipped mobile terminals, the amount of geo-tagged social images is increasing rapidly. Image location is an important type of information that could support many applications. First, it allows users in searching and organizing images more effectively. Second, by analyzing the correlation between geolocation and image content, we can discover the geographical knowledge about the popularity of image content patterns. For example, images about NYC might cover entirely different events compared to those about Beijing, and the choices of tags and visual features are different for the two cities. In addition, there are also large-scale and ever-growing images that are not tagged with GPS records. Hence, it is becoming increasingly important for an effective image location prediction algorithm, i.e., predicting where an image is taken [Serdyukov *et al*., 2009].

Some approaches have been proposed for image location prediction [Hays and Efros, 2008; Hauff and Houben, 2012]. These approaches can be roughly categorized into three classes. In the first category, user-generated text tags are used as a basis for location predication [Kling *et al*., 2014; Laere *et al*., 2011; Serdyukov *et al*., 2009; Yin *et al*., 2011]. Those approaches use a pure language model to identify the relation between text content and location. Since there are many noisy tags and different locations have their own characteristics of vision pattern, combining visual content could contribute to the performance of the proposed methods. Moreover, there are a large percentage of images that are not associated with any tags, and they cannot be geolocated with these text-based approaches. The second class is vision-based methods which estimate the location of query image based on the locations of visually similar images [Hays and Efros, 2008; Li *et al*., 2013; Li et al., 2009]. Due to the variety of visual content and the "semantic gap" [Smeulders *et al*.,2000; Zhang *et al*., 2012], exploiting visual content is challenging. The third class of approaches combines different types of image contents, i.e., text tags and visual content, for landmark prediction [Crandall *et al*., 2009]. However, the linear combination strategy of these approaches cannot be directly applied to social image since the text space and visual space have inherently different structures.

Meanwhile, the distribution of image content across geographical locations presents unique characteristics. First, some patterns of image content are popular in a large-scale region, while the sub-regions also have their location specific characteristics. It has been reported that these preferences correlate hierarchically [Wu and David, 2002]. For example, it is likely that images taken in Philadelphia and Camden are more similar than those in Chicago. Second, social images are far from uniformly distributed over the globe. The widely used single-level assumption by existing approaches is affected by the problem of unbalanced sampling on geographical regions. For example, there might be many geo-tagged images about the famous location Los Angeles but less geo-tagged images for the adjacent location Riverside. Given a query image from Riverside, it may be located to other densely photographed locations by the single-level method due to the little training images for Riverside. However, when using the hierarchal structure, it may first locate the query image to the super region that covers both Los Angeles and

[*]The majority of this work was done while the author was affiliated with Computer Science and Engineering at Arizona State University

Riverside, which increases the chance to estimate a more accurate location. This is because that images within the same super-region may share some features and can help to represent the location Riverside better other than the little training samples of region Riverside itself. Therefore, these characteristics motivate us to arrange image content and location preferences in a hierarchical structure and learn geographical hierarchy features which capture the hierarchical correlation for location prediction.

To learn the hierarchical features, it is quite challenging due to the following reasons. First, social images contain different modalities of content, i.e., text tags and visual content. These contents are presented in different feature space. It needs a joint representation that correlates both modalities and is thus ideal surrogate for learning the hierarchical features. Second, to capture the hierarchical structure, the high-level features should be able to express the distinctive perceptual structure of a specific region. That is, a hierarchy of super regions for sharing abstract knowledge among adjacent regions should be learned. Third, the hierarchical structure is implicit and not pre-defined.

To tackle the challenges, we propose to take advantage of the hierarchical correlation between image content patterns and locations for image location prediction. In particular, we investigate: (1) how to learn a joint representation which correlates visual content and text content; (2) how to automatically learn a hierarchical structure as well as hierarchical features that capture the hierarchical correlation between image content patterns and locations. Our solutions to these questions result in a new architecture for image location prediction. In particular, we propose a geographically hierarchical bi-modal deep belief network model (GH-BDBN) that integrates multi-modal deep models with a non-parametric hierarchical prior model. The bi-modal deep model (BDBN) is used to correlate visual content and text content through the top hidden level, and a geographically hierarchical prior over the top-level features of BDBN is proposed to model the hierarchical correlation between image content patterns and locations. The main contributions are outlined as:

- We propose a unified model to effectively integrate visual content and text content through their correlation for image location prediction.
- We propose a compound model GH-BDBN that integrates a deep learning model with a non-parametric hierarchical model, which learns the hierarchical features as well as the hierarchical structure.
- We empirically evaluate the proposed model on a real-world Flickr dataset and elaborate the effects of each type of content information for location prediction.

2 Problem Statement

In this section, we first introduce the notations used in the paper and then formally define the problem we study.

Notation: In this paper, the hierarchical structure is represented by a tree. We denote by C_i^l a node assignment of i-th input in the l^{th} level of the tree, where the superscript denotes the level of the variable indicated by the subscript, and other notations are defined similarly. $l(i)$ denotes the level of node i and $\pi(i)$ denotes its parent. Each geo-tagged image $x = \{x^p, x^t, l_x\}_{x \in \mathbf{I}}$ consists of three atoms: $x^p \in R^D$ is the visual feature vector for the visual content; $x^t \in R^M$ is the text feature vector for the text content; l_x is a real-valued 2-D vector containing the latitude and longitude where the image is taken; and $\mathbf{I}$ denotes the training dataset.

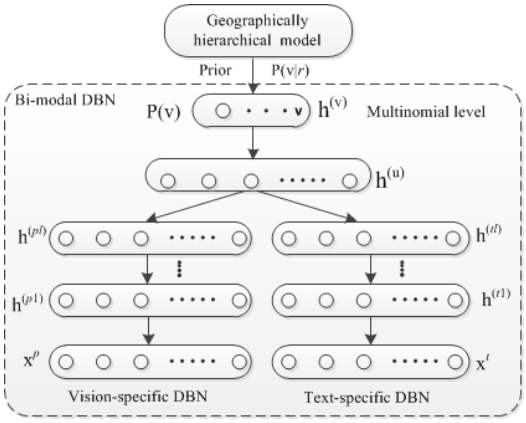

Figure 1: A graphical representation of GH-BDBN

With the given notations, we formally define the problem of location prediction for social image as follows:

Given a set of geo-tagged social images $\mathbf{I}$ *in which each image contains text features, visual features, and geo-coordinate, we aim to automatically estimate the geo-coordinate* l_e *for query image* e *based on* GH-BDBN.

3 Geographically Hierarchical bi-modal DBN

Usually, social images are uploaded with text descriptions such as tags and comments. Both of the visual content and text content relate to the location where the image was taken. Moreover, these types of content are correlated with location in a hierarchical structure. That is, adjacent locations always share some common features, such as language habit and natural scenery, while each location also has its own specific features. To model these information for image location prediction, we propose a geographically hierarchical bi-modal deep belief network model (GH-BDBN) to learn the geographically hierarchical features by incorporating the visual content, text content and geographical information of social images. A graphical representation of GH-BDBN is shown in Figure 1. The bi-modal deep model is used to correlated visual content x^p and text content x^t through the top hidden level $\mathbf{h}^{(v)}$ which is in turn used as the low-level features of the geographically hierarchical model. Then, we place a geographically hierarchical prior over the $\mathbf{h}^{(v)}$, which obtains both a layered hierarchy of increasingly abstract features and a tree-structured hierarchy of geographical regions.

3.1 Joint Representation of Social Image

To combine different types of image content for location prediction, some methods have been proposed using a linear combination strategy [Crandall *et al.*, 2009]. This method

cannot be directly applied to the problem that these contents are represented in heterogeneous feature. Therefore, we learn a joint representation that is expected to be a higher-level abstraction representation for both modalities and is thus a ideal surrogate for location prediction.

In particular, we use the bi-modal DBN (BDBN) [Srivastava and Salakhutdinov, 2012] to learn latent features from the raw features of visual content and text content. The deep learning model has been validated as an effective tool for feature learning and can deal with noisy data [Erhan *et al*., 2014; Chen *et al*., 2013]. As shown in the bottom of Figure 1, each data modality is modeled with a separate DBN which uses a Gaussian RBM to model the distribution over the input values. The two models are combined by learning a joint level $\mathbf{h}^{(u)}$ on top of them. The conditional distribution of $\mathbf{h}^{(u)}$ based on the two low levels is:

$$P(\mathbf{h}^{(u)}|\mathbf{h}^{(pl)},\mathbf{h}^{(tl)}) = \delta(\mathbf{W}_p^{(u)}\mathbf{h}^{(pl)} + \mathbf{W}_t^{(u)}\mathbf{h}^{(tl)} + \mathbf{b}^{(u)}) \quad (1)$$

where $\delta(x) = 1/(1+\exp(-x))$. The parameters of BDBN can be learned approximately by greedy lay-wise training and using contrastive divergence algorithm [Hinton *et al*., 2006].

To enable BDBN to express more information and introduce more structured hierarchical priors, we add another level $\mathbf{h}^{(v)}$ on the top. The activities of this level units are modeled by a conditional multinomial distribution. Specifically, $\mathbf{h}^{(v)}$ consists of M softmax units each of which is 1-of-K encoding and contains a set of K weights. The k^{th} discrete value of the hidden unit is represented by a vector containing 1 at the k^{th} location and zeros elsewhere. The conditional probability of a softmax unit of $\mathbf{h}^{(v)}$ is estimated by:

$$P(h_k^{(v)}|\mathbf{h}^{(u)}) = \frac{\exp(-b_k^v + \sum_j w_{kj}^v h_j^{(u)})}{\sum_{k'} \exp(-b_{k'}^v + \sum_j w_{k'j}^v h_j^{(u)})} \quad (2)$$

In Eq. (2), all M separate softmax units share the same set of weights, connecting them to binary hidden units at the lower level. This means that M separate copies of softmax units can be viewed as a single multinomial units that are sampled M times from the conditional distribution of Eq. (2) [Srivastava *et al*., 2013]. It is the same as the bag-of-words representation. The conditional distribution of $h_k^{(u)}$ based on softmax units is estimated as following:

$$P(h_k^{(u)} = 1|\mathbf{h}^{(v)}) = \delta(-\sum_j W_{kj}^{(v)} \hat{h}_j^{(v)} + a_k^{(u)}) \quad (3)$$

where $\hat{h}_j^{(v)} = \sum_{m=1}^{M} h_j^{(v,m)}$ denotes the count for the j^{th} discrete value of a hidden unit.

3.2 Geographically Hierarchical Prior

Usually, image content patterns and locations are hierarchically correlated, and locations are not uniformly photographed. Therefore, the single-level models [Kling *et al*., 2014; Li *et al*., 2013; Crandall *et al*., 2009] are not directly adaptable to this problem. It is reasonable to arrange image content patterns and location preferences in a tree structure, and thus the sparsely photographed regions can share

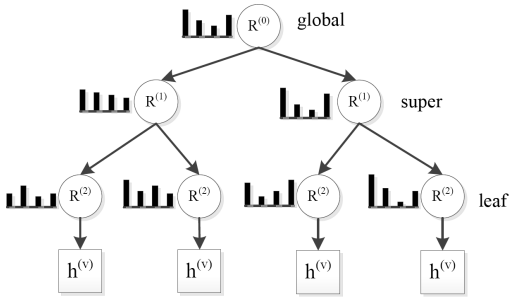

Figure 2: Geographically hierarchical prior over $\mathbf{h}^{(v)}$

the prior knowledge learned from the super region. Also, the joint representation is variety for different images and thus is unsuitable to model the more abstract knowledge of the larger scale regions. Therefore, we employ a geographically hierarchical prior over $\mathbf{h}^{(v)}$ which is used as the low-level features that provide a useful initial representation for all image content patterns. Then, the high-level features are learned to express the distinctive perceptual structure of a specific region, which is in terms of region-specific correlations over low-level features.

Specifically, we propose a geographically hierarchical topic model of prior over the activities of $\mathbf{h}^{(v)}$. The multinomial unit $\mathbf{h}^{(v)}$ is referred as document, and the M samples of $\mathbf{h}^{(v)}$ represent M "words" of the document. Locations are automatically clustered into hierarchical regions, and each region has its own distribution over topics that are modeled as mulinomial distribution over words [Blei *et al*., 2003] and are referred as our learned higher level features. A joint generative process is used to model the generation of image location and "words" as follows: for each image, a region is first chose, and then, its "words" and geographical location are generated from their corresponding distributions in this region.

Generation of Tree Structure

For simplicity, we present our hierarchical model with a three-level tree as shown in Figure 2. Each node denotes a region, and the parent region is a sum of the children regions.

Hierarchical regions: We use a hierarchical multivariate Gaussian model to capture the hierarchical structure of geographical locations. Each region r has a region-dependent multivariate normal distribution of location. To accomplish conjugacy of the joint prior distribution of the mean and the covariance to the likelihood, the distribution of the mean has to depend on the covariance. The prior distribution of the mean μ_r is Gaussian conditioned on Σ_r: $\mu_r \sim N(\xi_r, \rho_0^{-1}\Sigma_r)$. The prior distribution of Σ_r is inverse-Wishart: $\Sigma_r \sim IW(v_0, H_r)$. The joint distribution of μ_r and Σ_r is the Normal/inverse-Wishart distribution:

$$(\mu_r, \Sigma_r) \sim N\text{-}IW(\xi_r, \rho_0, v_0, H_r) \quad (4)$$

where ρ_0, and v_0 are hyperparameters, ξ_r and H_r are level-specific parameters: $\xi_r = \mu_{\pi(r)}$, $H_r = \alpha\Sigma_{\pi(r)} (0 < \alpha < 1)$.

Latent topic ϕ: The number of topics ϕ is assumed to be finite (it can be alleviated that the number is infinite).

Each topic is a multinomial distribution on word vocabulary, and is drawn from a Dirichlet distribution over words: $\phi_i \sim Dir(\eta)$.

Region-specific distribution of topics θ: The distribution over latent topics of each node is also modeled hierarchically. It is reasonable that geographical proximity is a good prior for similarity in region-specific distribution of topics. The Hierarchical Dirichlet Process (HDP) [Teh *et al*. 2006] is used to model this hierarchical distribution. First, we draw the global-level $\boldsymbol{\theta}^{(0)}$ for the global region $\boldsymbol{R}^{(0)}$: $\boldsymbol{\theta}^{(0)} \sim Dir(\beta)$. At the lower level node r, $\boldsymbol{\theta}_r$ is drawn using the parent $\boldsymbol{\theta}_{\pi(r)}$ as a prior: $\boldsymbol{\theta}_r \sim Dir(\varepsilon\boldsymbol{\theta}_{\pi(r)})$. In this way, the sparsely sampled regions can share the prior knowledge of the larger-scale region. Similarly, the topic distribution $\boldsymbol{\theta}_i^d$ of each document i is drawn using $\boldsymbol{\theta}_{\pi(i)}$ of the node which generates the document as a prior.

Prior of the tree structure

Since the tree structure is not pre-defined, we need to automatically infer the distribution over the possible hierarchical structures. We place a nested Chinese restaurant prior (nCRP) [Blei *et al*., 2010] over the path of node selection in the tree. The main building block of the nCRP is the Chinese restaurant process (CRP). To select a node, we define a CRP process over the children of a parent node i:

$$P(C_n^{l(i)+1} = j | C_1^{l(i)+1} ... C_{n-1}^{l(i)+1}) = \begin{cases} \frac{n_j}{n-1+\lambda} & n_j > 0 \\ \frac{\lambda}{n-1+\lambda} & j \text{ is new} \end{cases} \tag{5}$$

where n_j is the number of previous observations selecting child node j, λ is the concentration parameter, and $C_n^{l(i)+1} = j$ denotes that the n^{th} observation selects child node j. nCRP extends CRP to a nested sequence of partition, one for each level of the tree. This model allows for a nonparametric prior learns an arbitrary tree taxonomy.

3.3 Parameter Inference

We use Markov Chain Monte Carlo to infer parameters at all levels. When the tree structure $\boldsymbol{C}$ is unknown, the inference process alternates between fixing $\boldsymbol{C}$ while sampling the space of model parameters, and vice verse.

Sampling HDP parameters: Given the node assignments $\boldsymbol{C}$ and the states of $\mathbf{h}^{(v)}$, we use the posterior representation sampler to sample [Teh and Jordan, 2010]. That is, the HDP sampler maintains $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of each level, topic distribution of each level and document, i.e., $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, $\boldsymbol{\theta}^{d}$, and topic indicator variables $\boldsymbol{t}$ of words. The sampler alternates between: 1) sampling topic indicator t_{ni} for each word using Gibbs updating in the Chinese restaurant franchise (CRF) representation of HDP; 2) updating $\boldsymbol{\theta}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ at all levels conditioned on $\boldsymbol{t}$ and $\boldsymbol{C}$ with the usual posterior of a DP (Dirichlet Process). We use the MAP estimating method to update $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for each node r:

$$\mu_r = \frac{\rho_0 \mu_{\pi(r)} + n_r \bar{l}_r}{\rho_0 + n_r}, \Sigma_r = \frac{\rho_r + 1}{\rho_r (v_r - d + 1)} H_r \tag{6}$$

$$H_r = \alpha \Sigma_{\pi(r)} + S_r + \frac{\rho_0 n_r (\bar{l}_r - \mu_{\pi(r)})(\bar{l}_r - \mu_{\pi(r)})^T}{\rho_0 + n_r} \tag{7}$$

$$\rho_r = \rho_0 + n_r, v_r = v_0, S_r = \sum_{i=1}^{n_r} (l_i - \bar{l}_r)(l_i - \bar{l}_r)^T \tag{8}$$

where $\bar{l}_r$ denotes the average value of geographical coordinates in node *r*, and *d* is the dimensionality of geographical record. The posteriors over leaf node are independent, and the inference can be speeded up by sampling in parallel.

Sampling node assignments $\boldsymbol{C}$: Given the current instantiation of $\boldsymbol{\theta}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, for each input n we have:

$$(\theta_{n,i}^d, ..., \theta_{n,T}^d) \sim Dir(\varepsilon\theta_{C_n,1}^{(2)}, ..., \varepsilon\theta_{C_n,T}^{(2)}) \tag{9}$$

$$l_n \sim N(\mu_{C_n}^{(2)}, \Sigma_{C_n}^{(2)}) \tag{10}$$

where T is the number of topics. Combining the above likelihood terms with the CRP prior Eq. (5), the posterior over the node assignment is estimated as follows:

$$\begin{aligned} &P(\boldsymbol{C}_n | \boldsymbol{\theta}_n^d, \boldsymbol{C}_{-n}, \boldsymbol{\theta}^{(2)}, l_n, \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}^{(2)}) \\ &\propto P(\boldsymbol{\theta}_n^d | \boldsymbol{\theta}^{(2)}, \boldsymbol{C}_n) P(l_n | \boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma}^{(2)}, \boldsymbol{C}_n) P(\boldsymbol{C}_n | \boldsymbol{C}_{-n}) \end{aligned} \tag{11}$$

where $\boldsymbol{C}_{-n}$ denotes variables $\boldsymbol{C}$ for all observations other than n. When computing the probability of placing $\boldsymbol{\theta}_n^d$ and l_n under a new node, its parameters are sampled from the prior.

Sampling hidden unit $\mathbf{h}^{(v)}$: Given the states of $\mathbf{h}_n^{(v)}$, conditional samples from $P(\mathbf{h}_n^{(pl)}, \mathbf{h}_n^{(tl)}, ..., \mathbf{h}_n^{(u)} | \mathbf{h}_n^{(v)}, \mathbf{v}_n)$ can be obtained by running a Gibbs sampler that alternates between sampling the states of units in one level independently given the unit in the adjacent level. Conditioned on topic assignments t_{ni} and $\mathbf{h}_n^{(u)}$, the states of $\mathbf{h}_n^{(v)}$ for each input n are sampled using Gibbs condition:

$$P(h_{ni}^{(v)} | \mathbf{h}_n^{(u)}, \mathbf{h}_{n,-i}^{(v)}, \mathbf{t}_n) \propto P(\mathbf{h}_n^{(u)} | \mathbf{h}_n^{(v)}) P(h_{ni}^{(v)} | t_{ni}) \tag{12}$$

where the first term is estimated by the product of logistic functions of Eq. (3) as following:

$$P(\mathbf{h}_n^{(u)} | \mathbf{h}_n^{(v)}) = \prod_j P(h_{nj}^{(u)} | \mathbf{h}_n^v) \tag{13}$$

The second term of Eq.(12) is estimated by the multinomial, i.e., $Multi(\boldsymbol{\phi}_{t_{ni}})$. Since it has a conjugate prior of Dirichlet, the parameter $\boldsymbol{\phi}$ can be integrated out.

Fine-tuning bi-modal DBN: Given the states of the top level $\mathbf{h}^{(v)}$, the bi-modal DBN parameters in the low levels are fine-tuned in the same way as in contrastive divergence algorithm. Fine-tuning low-level BDBN features can improve the model by encoding the geographical information in the learning of joint representation, and thus the learned representation is more effective to reflect the geographical information.

4 Location Prediction

Based on the results of posterior inference, it is straightforward to predict the node assignment of a query image e. That is, we first infer the posterior over $\mathbf{h}_e^{(v)}$ on the bi-modal DBN, and then we run the Gibbs sampler to get approximate samples from posterior over the node assignments. For calculation efficiency, we approximate the document-specific topic

distribution by the leaf node-specific distribution $\boldsymbol{\theta}^{(2)}_{C_e}$. Then the posterior probability of a node assignment is:

$$P(\mathbf{h}_e^{(v)}|\boldsymbol{C}_e, \boldsymbol{\phi}, \boldsymbol{\theta}) \propto P(\mathbf{h}_e^{(v)}|\phi, \boldsymbol{\theta}^{(2)}_{C_e}) = \prod_j (\sum_k P(h^{(v)}_{ej}|\boldsymbol{\phi}_k, \boldsymbol{\theta}^{(2)}_{C_e,k})) \quad (14)$$

where the topic assignments of words can be integrated out. By combining the nCRP prior, this likelihood can be used to approximately infer the posterior over node assignment. Then, we select top-*m* images that are most similar to the query image from the images located in the assigned node. The similarity between two images is estimated based on the JSD (Jensen-Shannon-Divergence) between their probability distributions over latent topics. Finally, the location of the query image *e* is estimated as following:

$$l_e = \frac{\sum_{g \in N_m(e)} Sim(g, e) \cdot l_g}{\sum_{g \in N_m(e)} Sim(g, e)} \quad (15)$$

where $N_m(e)$ denotes the *m* most similar images.

5 Experiments

In this section, we conduct experiments to assess the effectiveness of GH-BDBN. Through the experiments, we aim to answer:

1. What are the effects of content type and hierarchy prior on image location prediction?
2. How effective is the proposed model compared with other methods of image location prediction?

We begin by introducing the experiment settings and then analyze the experimental results.

5.1 Experiment Settings

We use the dataset MediaEval2012 [Rae and Kelm, 2012] which is a community-driven benchmark and is run by the MediaEval organizing committee to evaluate our model. Since MediaEval2012 doesn't includes the raw images and some images were removed after the dataset was collected, we download 1,020,130 raw images with their tag lists using the links in the metadata from Flickr. We concatenated 1024-D HOG features, 144-D color correlogram in HSV color space, and 128-D wavelet texture to set a 1296-D representation for image's visual content. Each dimension was mean-centered. As the raw tags are sparse and noisy, we include a new presentation for text content. First, we remove the stop words and words that appear in more than 10% of the images. Word2Vec [Mikolov *et al*., 2013] is used to represent each tag with a 500-D vector, which is trained on Wikipedia articles of about 100 million words. Then, all the image tag lists are clustered into 2000 groups based on the geographical information using GMM algorithm, and a 2000-D geographical dictionary is obtained. Finally, each tag is represented by a 2000-D sparse codes learned from the geographical dictionary, and the tag list of each image is represented by a 2000-D feature by max pooling all the words in the list. For BDBN, there are three levels for the vision-specific DBN, and the numbers of units are 1296, 1000, and 800 respectively. Since text is closer to the learned latent feature level, we use two levels to model the text-specific DBN with the numbers of units being 2000 and 1000 respectively. For $\mathbf{h}^u$ and $\mathbf{h}^v$, we set the numbers of units to be 1500 and 1000 respectively. We split the dataset and use 80% for training and 20% for testing.

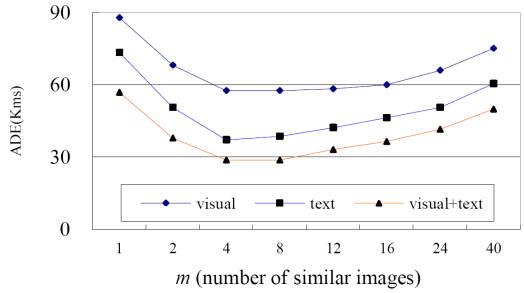

Figure 3: ADE of using different types of content

There is a parameter involved in the experiments, i.e., *m* which denotes the number of most similar images in Eq.(15). The effect of the parameter on location prediction will be further discussed in Section 5.2. For quantitative evaluation, we use average distance error (ADE) as the evaluation metric: $ADE = \sum_{i=1}^{N}(\hat{l}_i, l_i)/N$, where $dis(\hat{l}_i, l_i)$ is the Euclidean distance between the estimated location and the true location.

We compare our approach with three groups of four approaches that are effective for location prediction based on different types of image content. The first group of approach is the text-based LMSS (language model and similarity search) [Laere *et al*., 2011]. It uses the text model to first locate the region and then precisely determine the location in this region based on text similarity measuring. The second group of approaches are vision-based IM2GPS [Hays and Efros, 2008] and GVR [Li *et al*., 2013]. They estimate the location of query image based on the locations of visual neighbors. The third group of approach is Mapping [Crandall *et al*., 2009] which combines both visual feature and text content to classify the query image into a region. We revise it to estimate a more precise location based on the locations of neighbor images similar to Eq.(15).

5.2 Effects of Content Type and Hierarchy Prior

In this subsection, we study the importance of image content type and hierarchy prior, which accordingly answers the first question asked in the beginning of Section 5.

First, we evaluate the performance of our model on predicting location by using different types of image content, i.e., only visual content, only tags, and both of visual content and tags, respectively. Figure 3 shows the performance of predicting location using different types of content, with various values of *m* shown in the *X* axis and ADE of using different types of image content shown in the *Y* axis. From the results in Figure 3, we draw several observations. First, the performance is not always proportionate to the number of similar images used in location estimation, and the optimal choice of *m* is 4. When *m* is too small, the influence of the falsely selected images increases. When *m* is too great, the chance

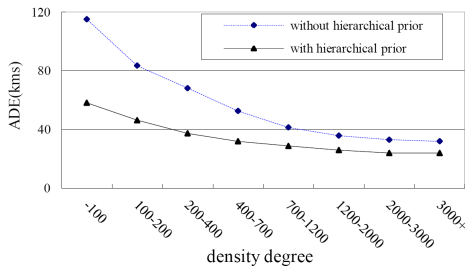

Figure 4: Evaluation of hierarchical prior

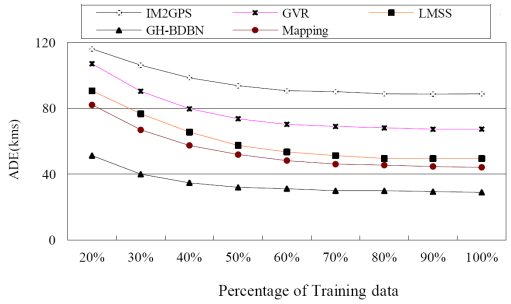

Figure 5: Performance comparison of different approaches

of selecting more noisy images increases. Second, location prediction using both of visual content and tags has the best performance, and the text content is more effective in location prediction than visual content. This is because that, usually, text description is more effective in conveying the semantic content of social image than visual features. Since there are many noisy tags and visual content is of large variety in each geographical location, combining both text content and visual content has a better performance than using text content only.

Then, we test whether the hierarchical prior is effective for location prediction. The prediction without hierarchical prior has following steps. First, we cluster the training images based on their geo-coordinates using GMM, and each cluster is labeled as a region. Then, we use the semi-supervised classification model [Liu *et al.*, 2011] to train a classifier on the bi-modal DBN. To predict the location, the query image is classified to a region, and then a precise location is estimated similar to Eq.(15). To evaluate the performance on unbalanced sampling regions, we split the geo-coordinates into grid cells whose sizes are 50km*50km, and density degree is used to define the number of images in a cell. Figure 4 illustrates the performances of GH-BDBN and the deep model without hierarchical prior respectively, with different degrees of image density shown in the X axis. The dashed line denotes the performance of the model without hierarchical prior, and the solid line denotes the performance of GH-BDBN. It can be seen that GH-BDBN performs better at relatively low densities. This is because the low density region can share prior knowledge of the geographical neighbor regions in our model, which decreases the effect of unbalanced sampling. Without the hierarchical prior, the query image located in the low density region will has a great chance to be wrongly classified to other density regions that are far away from the target region. It can also be seen that GH-BDBN outperforms the deep model without hierarchical prior across the board, which demonstrates that exploiting the hierarchical correlation help improve the performance of image location prediction.

5.3 Performance Comparison

To answer the second question asked in the beginning of Section 5. We compare GH-BDBN with the baseline approaches. Figure 5 shows the experimental result with the percentage of training dataset varied from 20% to 100%, by fixing m=4. The results show that all the performances are affected by the volume of training data. This is because when the training dataset is too small, the distributions of images in many locations are very sparse, especially in the locations that are less frequently photographed. The performance of our model is less sensitive to the volume of training data because we can enrich the knowledge of the sparsely photographed region by exploiting the knowledge of the large-scale region with geographically hierarchical prior model. Moreover, our deep learning model is more effective to learn latent feature from few examples. It also shows that our approach outperforms other ones consistently. IM2GPS and GVR predict location mainly based on the similarity between visual feature directly. However, the visually similar images might be semantically far away from each other. The text-based approach LMSS performs better than the vision-based approaches. It uses a pure text model to discover the link between image tags and locations, which ignores the vision patterns of each location. Therefore the performance is also affected. Mapping is better than other three baselines since it exploits both visual feature and textual feature. However, the linear combination strategy is not directly adaptable to the features that belong to different representation spaces. Our approach learns the joint representation to correlate different types of image content, and the geographical hierarchy features are also learned to more effectively capture the hierarchical correlation between image content patterns and location preferences.

6 Conclusion and Future Work

In this paper, we proposed a compositional learning architecture GH-BDBN that integrates multi-modal deep learning models with non-parametric hierarchical prior models. Experimental results on real-world dataset demonstrate the effectiveness of GH-BDBN. The novelty of this work is to tackle the analysis and application in geo-tagged multi-modal data with latent feature automatically learned from a multi-modal deep architecture with nonparametric geography hierarchy prior. This complements the current research which focuses on single-level model for raw data directly and neglects the inherent and hierarchical relation between multi-modal data and geolocations.

There are many potential future extensions of this work. It would be interesting to investigate image owner's other social information, like the micro-blogs published in the travel and the interesting points that the owner often visit, for image location prediction.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 61202239, No. 60973105, and No. 61170189), the Fundamental Research Funds for the Central Universities (No. YWF-14-JSJXY-16), and the Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE-2015ZX-11).

References

[Blei *et al*., 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation, *Journal of Machine Learning Research*, 3: 993-1022, 2003.

[Blei *et al*., 2010] David.M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. *Journal of the ACM*, 57(2), 2010.

[Chen *et al*., 2013] Fei Chen, Huimin Yu, Roland Hu, Xunxun Zeng. Deep Learning Shape Priors for Object Segmentation. In *Proceedings of the First International Con-ference on Learning Representations*, pages 1870-1877, 2013.

[Crandall *et al*., 2009] David Crandall, Lars Backstrom, Dan Huttenlocher, and Jon Kleinberg. Mapping the worlds photos. In *Proceedings of the 18th International World Wide Web Conference*, pages 761-770, 2009.

[Erhan *et al*., 2014] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable Object Detection using Deep Neural Networks. In *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition*, pages 2155-2162, 2014.

[Hauff and Houben, 2012] Claudia Hauff, Geert-Jan Houben. Placing images on the world map: a microblog-based enrichment approach. In *Proceedings of the.35th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval*, 2012.

[Hays and Efros, 2008] James Hays, and Alexei Efros. im2gps: estimating geographic information from a single image. In *Proceedings of the 21th IEEE Conference on Computer Vision and Pattern Recognition*, pages 1-8, 2008.

[Hinton *et al*., 2006] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527-1554, 2006.

[Kennedy and Naaman, 2008] Lyndon Kennedy, and Mor Naaman. Generating diverse and representative image search results for landmarks. In *Proceedings of the 17th International World Wide Web Conference*, pages 297-306, 2008.

[Kling *et al*., 2014] Christoph C. Kling, Jerome Kunegis, Sergej Sizov, and Steffen Staab. Detecting Non-Gaussian Geographical Topics in Tagged Photo Collections. In *Proceedings of the 7th ACM international conference on Web Search and Data Mining*, pages 603-612, 2014.

[Laere *et al*., 2011] Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. Finding locations of Flickr resources using language models and similarity search. In *Proceedings of the First Annual ACM International Conference on Multimedia Retrieval*, pages 48-55, 2011.

[Li *et al*., 2013] Xinchao Li, Martha Larson, and Alan Hanjalic. Geo-visual ranking for location prediction of social images. In Proceedings of the 21st ACM International Conference on Multimedia Retrieval, pages 81-88, 2013.

[Li *et al*., 2009] Yunpeng Li, David J. Crandall, and Daniel P. Huttenlocher. Landmark classification in large-scale image collections. In *Proceedings of the 13rd International Conference on Computer Vision*, pages 1957-1964, 2009.

[Liu *et al*., 2011] Yan Liu, Shusen Zhou, Qingcai Chen. Discriminative deep belief networks for visual data classification. P*attern Recognition*, 44:2287-2296, 2011.

[Mikolov *et al*., 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the First International Conference on Learning Representations*, pages 4089-4114, 2013.

[Rae and Kelm, 2012] Adam Rae and Pascal Kelm. Working notes for the Placing Task at MediaEval 2012. In *Proceedings of MediaEval 2012 Workshop*, 2012.

[Serdyukov *et al*., 2009] Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. Placing Flickr Photos on a Map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and Development on Information Retrieval*, pages 484-491, 2009.

[Smeulders *et al*.,2000] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, Ramesh Jain, Content-Based Image Retrieval at the End of the Early Years, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1349-1380, 2000.

[Srivastava and Salakhutdinov, 2012] Nitish Srivastava, and Ruslan Salakhutdinov. Learning Representations for Multimodal Data with Deep Belief Nets. In *Proceedings of ICML Representation Learning Workshop*, 2012.

[Srivastava *et al*., 2013] Nitish Srivastava, Ruslan Salakhutdinov, and Geoffrey Hinton. Modeling Documents with a Deep Boltzmann Machine. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 2013.

[Teh *et al*. 2006] Yee Whye Teh, Michael I. Jordan , Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476): 1566-1581, 2006.

[Teh and Jordan, 2010] Yee Whye Teh, and Michael I. Jordan. Hierarchical Bayesian Nonparametric Models with Applications. *Bayesian Nonparametrics in Practice*, UK: Cambridge Univ. Press, 2010.

[Wu and David, 2002] Jianguo Wu, John L. David. A spatially explicit hierarchical approach to modeling complex ecological systems: theory and applications. *Ecological Modelling*, 153: 7-26, 2002.

[Yin *et al*., 2011] Zhijun Yin, Liangliang Cao, Jiawei Han, and Chengxiang Zhai. Geographical topic discovery and comparison. In *Proceedings of the 20th International World Wide Web Conference*, pages 247-256, 2011.

[Zhang *et al*., 2012] Xiaoming Zhang, Heng Tao Shen, Zi Huang, Zhoujun Li, Yang Yang. Automatic Tagging by Exploring Tag Information Capability and Correlation. *Journal of World Wide Web*, 15(3): 233-256, 2012.