

# A Unified Framework of Latent Feature Learning in Social Media

Zhaoquan Yuan, Jitao Sang, Changsheng Xu, *Fellow, IEEE*, and Yan Liu

**Abstract**—The current trend in social media analysis and application is to use the pre-defined features and devoted to the later model development modules to meet the end tasks. Representation learning has been a fundamental problem in machine learning, and widely recognized as critical to the performance of end tasks. In this paper, we provide evidence that specially learned features will addresses the diverse, heterogeneous, and collective characteristics of social media data. Therefore, we propose to transfer the focus from the model development to latent feature learning, and present a unified framework of latent feature learning on social media. To address the noisy, diverse, heterogeneous, and interconnected characteristics of social media data, the popular deep learning is employed due to its excellent abstract abilities. In particular, we instantiate the proposed framework by (1) designing a novel relational generative deep learning model to solve the social media link analysis task, and (2) developing a multimodal deep learning to lambda rank model towards the social image retrieval task. We show that the derived latent features lead to improvement in both of the social media tasks.

**Index Terms**—Deep learning, feature learning, india buffet process, social media.

## I. INTRODUCTION

**S**Ocial media is defined as the means of interactions among people in which they create, share, and exchange information and ideas in virtual communities and networks [1]. Recently, more and more users participate in content creation rather than just consumption in these social media networks. With the explosive growth of user generated data on the web, social media has become one of the most popular web applications, and plays an important role in related multimedia applications. Social media problems have been extensively investigated in multimedia research community, ranging from image/video annotation [2] and multimedia retrieval [3] to user recommendations [4] and target advertisement [5].

Manuscript received October 12, 2013; revised March 04, 2014; accepted April 28, 2014. Date of publication May 06, 2014; date of current version September 15, 2014. This work was supported in part by the National Basic Research Program of China under Grant 2012CB316304, the National Natural Science Foundation of China under Grants 61225009, 61373122, and 61303176, the Beijing Natural Science Foundation under Grant 4131004, and the Singapore National Research Foundation under its International Research Centre at the Singapore Funding Initiative and administered by the IDM Programme Office. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Cees Snoek.

Z. Yuan, J. Sang, and C. Xu are with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zqyuan@nlpr.ia.ac.cn; jtsang@nlpr.ia.ac.cn; csxu@nlpr.ia.ac.cn).

Y. Liu is with the Department of Computing, Hong Kong Polytechnic University, Kowloon 999077, Hong Kong (e-mail: clyliu@comp.polyu.edu.hk).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2014.2322338

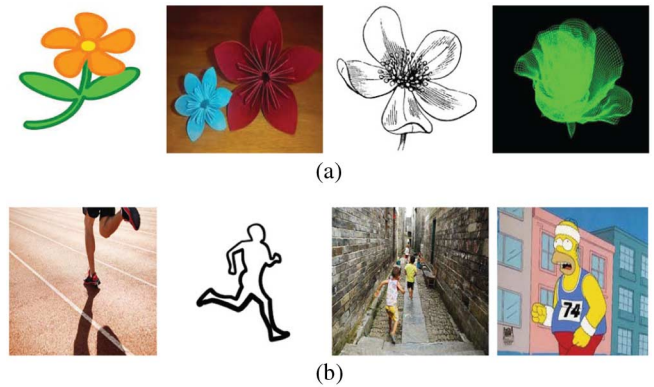


Fig. 1. Example images from Flickr to demonstrate the diversity characteristic of social media data.

Different from conventional multimedia data, social media network exhibits unique characteristics, which poses practical challenges to social media analysis and applications. Many solutions have been proposed to address these challenges. In the following, we elaborate the social media data characteristics and challenges in its generation, distribution and interaction, as well as briefly review the corresponding solutions.

Firstly, from the perspective of generation, social media data are noisy and diverse. The user-generated mechanism gives rise to their low quality and large quantity. Users with various backgrounds use social media to record their daily life, resulting in subjective social media data and featuring a diverse distribution of attributes like resources, appearance and degree of diffusion. The diversity characteristic poses challenges to the basic social media analysis tasks, such as social image classification. The social media data expressing the same concept, even from the same modality, may vary much from each other, making discriminative representation very difficult. Fig. 1 shows some example images from Flickr associated with user-generated tags of “flower” and “running”, respectively. While the images in the same row express the same concept, their appearances are very different. To address this challenge, the idea of social media data preprocessing is applied. By removing noise from user-generated content to obtain clean and refined data, application-specific algorithms are designed for solutions [6]. Typical research topics include video duplicate detection [7], image tag refinement [8] and social media organization [9]. While social media data preprocessing can address the noisy and diverse issues in some degree, it is hard to distinguish between the concept-related parts and the concept-free parts to get the concept semantic representation for diverse media.

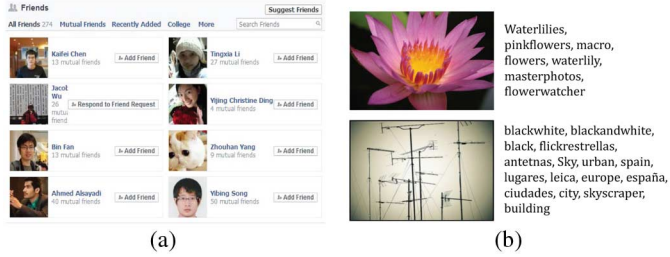


Fig. 2. (a) Linked friends information of a user in Facebook, and (b) two sample images uploaded by a user in Flickr associated with social tags.

Secondly, from the perspective of distribution, social media data are heterogeneous. It is very common that multiple forms of data, e.g., text and image, exist simultaneously on the same social media platforms. Fig. 2 shows part of the linked friends information of a user in Facebook (a), and two sample images uploaded by users in Flickr associated with social tags (b). Since social media data of different modalities follow much different statistical distributions, the latent feature structure is quite complicated. Data with different structures and distributions prevent from integrated social media understanding. Take user modeling as an example, to understand user preferences from the online activities, viable solutions need to model heterogeneous user data, e.g., registration profile, browsing history, shared images and videos, added comments and annotations, in a principled way. To overcome the difficulty of heterogeneity for social media data, social media semantic understanding methods are proposed by extracting semantics for each modality of data via multimedia content analysis, and gaining overall understanding of heterogeneous data in the derived semantic space. Typical research topics include cross-modal retrieval [10], topic and event identification [11], and social media knowledge mining [12][13]. However, the models in this kind of methods are usually shallow and have limited representation capabilities [14]. The task of learning the modality-free unified representation is still a challenge in social media analysis.

Last, but maybe the most important, from the perspective of interaction, most of the social media data are interconnected. We refer to it as the “collective” effect that social media data do not exist independently but interact with each other. The collective effect is either explicit or implicit, e.g., the interaction of observed user-user relation to their online behaviors is explicit, while the interaction of collaborative annotation to derive the final tag metadata is implicit. The collective effect among social media data violates the independently and identically distributed assumption in most statistical machine learning algorithms. Both content and collective information need to be considered for solutions. Social media network analysis methods are proposed to address this challenge, by embedding the interconnected social media data into tensors [15] or social graphs [5], where social network analysis methodology and metric is exploited for solution. Typical research topics include community detection [16], social behavior mining [17], and contextual social media analysis [18]. These methods mostly aim at the explicit “collective” effects, and ignore the implicit ones. Moreover, in these existing models, the latent representation is not in the unified semantic level, which cannot handle the complex

links in social media network, especially for the heterogeneous ones.

From the discussion above about the social media challenges and existing solutions, we can see that most of these current social media research efforts are devoted to the high-level model development, while representation learning, one of the fundamental and hot research topic in machine learning, is largely ignored. The importance of feature learning is far from being emphasized in current social media community, and we believe that it is a promising research line to take advantage of feature learning to address the challenges in social media networks.

In this paper, we present a novel framework on the latent feature learning to solve the high-level tasks in social media networks from an alternative feature representation perspective. We tackle the analysis and applications in social media network with latent features automatically learned from a specially designed deep architecture.

A preliminary version of this work was introduced in [19]. The extension in this paper mainly includes two aspects: 1) we explicitly present a unified framework of deep architecture-based latent feature learning in Section II-B; and 2) we extend the proposed framework towards unified multimodal feature learning by addressing the image retrieval application in Section IV.

The rest of the paper is organized as follows. In Section II, we firstly introduce the related work about latent feature learning, and then present a unified framework of latent feature learning in social media. Targeting at solving the social media link analysis tasks, following the proposed framework, we describe the novel RGDBN model in Section III. Likewise, targeting at solving social image retrieval tasks, in Section IV, a novel multi-modal deep learning to lambda rank model is developed. In Section V, we give a brief review of the related work about link analysis in social media network and image retrieval methods. We conclude the paper with an outlook on future work and challenges to be tackled in Section VI.

## II. FRAMEWORK OF LATENT FEATURE LEARNING

### A. Review of Feature Learning

In this section, we give a brief review of the related work about feature learning methods.

Feature learning is a classical problem in machine learning, with extensive efforts devoted to it. Besides the deep learning research line on which our proposed latent feature learning framework is based, we review several other feature learning algorithms.

One kind of latent feature learning methods aims to find a low-rank approximation for the raw features, including sparse coding [20], PCA [21], ICA [22], etc. They map the raw data to a lower-dimensional representation based on the assumption that the data lie (approximately) in an underlying low-dimensional linear subspace.

Another kind of feature learning methods is based on the single-layer network, e.g., RBM, auto-encoder, and  $K$ -means clustering. [23] shows that large numbers of hidden nodes and dense feature extraction are critical to achieving high performance for these methods.

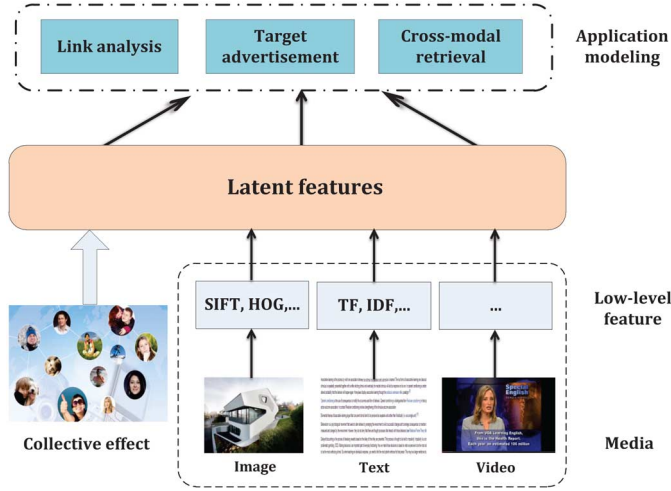


Fig. 3. Framework of latent feature learning-based social media analysis and application.

In addition, topic model [24] can also be regarded as a solution of latent feature learning, which aims to discover the abstract “topics” that occur in a collection of raw data and takes the “topics” distributions as the semantic representation of the raw data.

From the perspective of model architecture depth, above methods are shallow models. Shallow models are recognized as encountering the curse of dimensionality, and having limited capability in learning the distributed representation in complex situations [14]. Some related work [25][26] has shown that the shallow models failed to model the diverse data. Also, due to the gap between data in different modalities, shallow models are limited in handling the heterogeneous data [27], and it is hard to get the unified modality-free representation.

### B. Deep Learning-based Framework of Latent Feature Learning.

Regarding the characteristics of social media data and the limitations of shallow models, in this section, we describe a novel unified framework of latent feature learning in social media network. By integrating the collective prior into the deep learning architectures, we learn the latent representations of media in social media network to handle the challenges in social media analysis tasks described in Section I.

Our basic premise is that, if we model the collective effect and learn a unified feature representation for various social media data, later tasks of social media analysis and application can be solved with off-the-shelf machine learning algorithms. The idea is illustrated in Fig. 3. The latent features which can be considered as a kind of latent representation for media in social media network are learned by 1) composing and decomposing structure-specified low-level features; and 2) integrating the collective effect from interconnected social media data. Higher-level social media tasks, e.g., link analysis, cross-media retrieval, are performed based on the derived latent feature space. Positioned in the middle-level, the latent feature representation involves

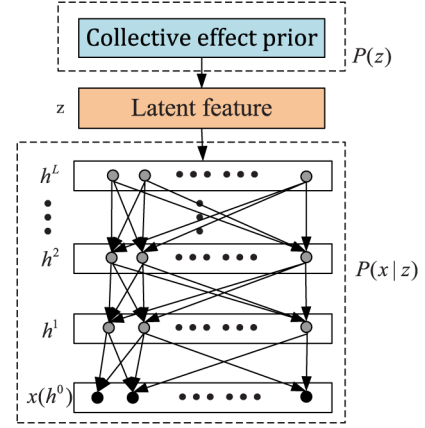


Fig. 4. Deep learning-based latent feature learning framework in social media.

with feature fusion from low-level features and captures the dependency between interconnected data. Moreover, since the latent feature is data-driven and learned automatically rather than pre-defined, the robustness is guaranteed.

To realize the framework mentioned above and handle the challenges described in Section I, we propose a deep learning-based framework of latent feature learning which is shown in Fig. 4.

Towards the challenges of diversity and heterogeneity, we build our latent feature learning framework on the deep architecture, where the low-level feature illustrated in Fig. 3 is set as the lowest layer  $x$ , and the latent feature layer  $z$  is positioned above the top level deep nets. Deep learning framework, which models the learning task using deep architecture composed of multi-layer nonlinear modules, provides a powerful tool for automatic feature learning. The ability of deep learning in high-level abstraction and distributed representation has been validated in both classical machine learning problems [28][29] and social media related problems [27][30]. The reason that we utilize the deep learning framework for feature learning is two-fold: 1) The mechanism of bottom-up greedy unsupervised pre-training and fine-tuning fits well to the characteristics of diversity and heterogeneity of social media data. The final feature representation is learned via a feature hierarchy, where higher-level features are formed by composition and decomposition of the lower-level features. Following this layer-wise learning structure, the derived features are expected to capture the dependency and interaction in the raw features and explain more abstract and robust semantics; 2) Deep learning can be explained from multiple related machine learning perspectives, such as neural network, probabilistic graphical model, etc. This theoretical flexibility makes it possible to solve the feature learning of social media data and final social media applications under a unified framework.

Towards the challenge that social media data are interconnected, where the collective effect captures the dependency among the media and plays an important role in data generation and relationship understanding, we place the collective effect and deep architecture within the Bayesian framework. With the collective effect illustrated in Fig. 3 modeled as prior for the latent feature, the observed data can be viewed as generated

from the latent feature through a deep architecture. The collective effect defines the prior distribution  $P(\mathbf{z})$  of latent feature  $\mathbf{z}$ , and the deep generative process determines  $P(\mathbf{x}|\mathbf{z})$ , the distribution over observed low-level feature conditioned on the latent feature. According to the Bayesian theory, the posterior distribution  $P(\mathbf{z}|\mathbf{x})$  of latent feature  $\mathbf{z}$  is

$$P(\mathbf{z}|\mathbf{x}) \propto P(\mathbf{z}) \cdot P(\mathbf{x}|\mathbf{z}). \quad (1)$$

In our proposed unified latent feature learning framework, we use a modified Deep Belief Nets [31] to model the deep generative process with multiple non-linear transactions to learn the high-level representation. In the conventional DBN, the top two layers constitute a Restricted Boltzmann Machine (RBM) [32], and the remaining hidden layers form a directed acyclic graph that convert the representations in the associative memory into observable variables. Different from the conventional DBN, our deep generative multi-layer network is composed of multiple layers  $\{\mathbf{h}^{(0)}, \mathbf{h}^{(1)}, \dots, \mathbf{h}^{(L)}\}$  and the data are generated in a top-down way directly. The generative process for  $\mathbf{x}_i$  is represented as follows:

$$\begin{aligned} P(\mathbf{x}_i|\mathbf{h}^{(1)}, \dots, \mathbf{h}^{(L)}, \mathbf{z}_i) \\ = P(\mathbf{z}_i) \left( \prod_{l=1}^{L-1} P(\mathbf{h}^{(l)}|\mathbf{h}^{(l+1)}) \right) P(\mathbf{x}_i|\mathbf{h}^{(1)}) P(\mathbf{h}^{(L)}|\mathbf{z}_i). \end{aligned} \quad (2)$$

The units in each layer are independent given the values of the units in the layer above and the parameterization of these conditional distributions is

$$\Pr(h_s^{(l)} = 1|\mathbf{h}^{(l+1)}) = \text{sigmoid}\left(b_s^{(l)} + \sum_t W_{s,t}^{(l+1)} h_t^{(l+1)}\right) \quad (3)$$

where  $h_s^{(l)}$  is the binary activation of hidden unit  $s$  in layer  $l$ ,  $\mathbf{h}^{(l)}$  is the vector  $(h_1^{(l)}, h_2^{(l)}, \dots)$ ,  $\text{sigmoid}()$  denotes the logistic function, and  $b_s^{(l)}$  and  $W_{s,t}^{(l+1)}$  denote the bias and weight between unit  $s$  in layer  $l$  and unit  $t$  in layer  $l+1$ , respectively. The transaction way from  $\mathbf{z}_i$  to  $\mathbf{h}_i^{(L)}$  is the same as Equation (3).

In most of the social media analysis tasks, the input feature  $x_i$  is continuous rather than binary. Therefore, the Gaussian RBM [33] is used to model the generative process in the lowest layer

$$P(x_s|\mathbf{h}^{(1)}) = \mathcal{N}\left(b_s + \sigma_s \sum_t W_{s,t}^{(1)} h_t^{(1)}, \sigma_s^2\right) \quad (4)$$

where  $\sigma_s^2$  denotes the variance of the unit  $s$ . We encourage the model to give high probability to generate the training data.

As the deep architecture re-uses and extracts the features in a layer-wise way during the learning process, when the learning process is completed, the “invariant” and “abstract” features are learned and lie in the higher layers of architecture which preserve the latent feature representation for the media content.

Combining the collective prior and deep generative learning process under the Bayesian framework, the collective effect and media content information are integrated into the unified latent features  $\mathbf{z}$ .

Similar to most deep architecture models, the whole learning processes include layer-wise pre-training and a following fine-tuning stages. The greedy layer-wise pre-training is the phase of constructing the deep architecture based on Restricted Boltzmann Machine (RBM) [32] which is an undirected graphical model with a hidden layer and another visible layer. The multi-layer deep network is built in a bottom-up fashion, where each pair of two adjacent layers can be regarded as a RBM by taking the lower layer as visible layer and the upper layer as hidden layer.

After having greedily pre-trained the deep multiple layers, the parameters are adjusted in the fine-tuning stage. Different from conventional deep models, both Bayesian sampling and discriminative supervised learning can be applied in our framework. The choice of inference methods for latent feature  $\mathbf{z}$  depends on the strategy of defining the collective effect prior, which is necessarily task-specific. For the social media data with strong interactivity, we explicitly define its prior distribution. In Section III, we combine the non-parametric Bayesian method (i.e., Indian Buffet Process, IBP) and Bayesian sampling method to infer the latent feature in link prediction tasks. However, for the tasks where the collective effects are weak, and the data are with weak interactivity, it is better to use the discriminative loss feedback to learn the latent feature. Our application of image retrieval in Section IV falls into this case.

The proposed deep learning-based latent feature learning framework is flexible, which can be instantiated into many concrete models to solve the corresponding social media applications. In this paper, we instantiate two examples to validate the effectiveness of the proposed framework to solve different social media problems. Firstly, we instantiate the framework into a relational generative deep learning model towards solving social media link analysis problems in Section III. In the proposed RGDBN model, we assume that the observed links between data are generated from the interactions of latent features. By integrating a flexible non-parametric Bayesian model of IBP into the modified Deep Belief Nets [31], we learn the optimal latent features that best embed both the media content and the explicit collective effects. The model is capable of analyzing the links between heterogeneous as well as homogeneous data, with effectiveness validated from social media applications in social image annotation and user recommendation. Secondly, we instantiate the proposed framework into a multimodal deep learning to lambda rank model towards solving the unified multimodal feature learning tasks with the application of image retrieval in Section IV. In the proposed model, visual image raw feature and textual query feature are fused through the deep learning architecture, and unified multimodal features are learned. During the learning process, discriminative pairwise lambda loss [34] is used to adjust the parameters in global fine-tuning stage.

The two instantiations have intrinsic relation and correspond to two kinds of typical applications in social multimedia. Their relations are featured in two-fold. On one hand, they illustrate how different applications are converted into the unified latent feature framework, and solved from the latent feature learning perspective. On the other hand, their models are different in the

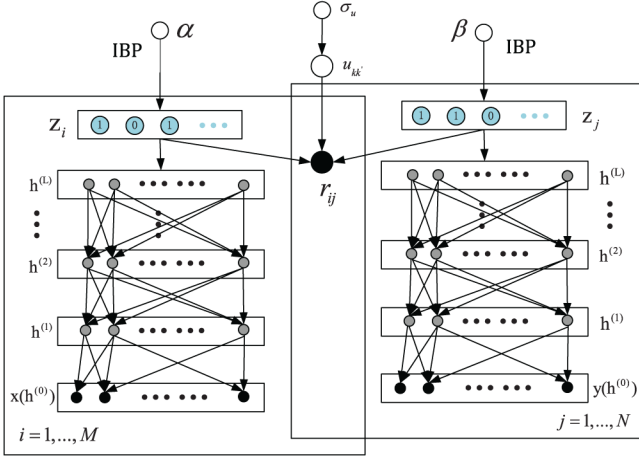


Fig. 5. The framework of relational generative deep belief nets.

training way, especially in the fine-tuning stage, which illustrate the two kinds of learning ways for latent features.

### III. LATENT FEATURE LEARNING FOR LINK ANALYSIS

In this section, we address the important link analysis problems in social media network so as to illustrate the latent feature learning methods for social media analysis. Due to the heterogeneity characteristic and collective effect discussed in the introduction section, link analysis is a challenging problem. Within the latent feature learning framework, we propose a novel Relational Generative Deep Belief Nets (RGDBN) model to handle this problem. In the RGDBN model, the link between items is generated from the interactions of their latent features. By integrating the Indian buffet process into the modified Deep Belief Nets, we learn the latent feature that best embeds both the media content and observed media relationships. The model is able to analyze the links between heterogeneous as well as homogeneous data.

#### A. Relational Generative Deep Belief Nets

Assume we observe the two kinds of set of items “A” and “B” with the raw feature matrices  $\mathbf{X}$ ,  $\mathbf{Y}$ , where each row  $\{\mathbf{x}_i\}_{i=1}^M$  and  $\{\mathbf{y}_j\}_{j=1}^N$  denote the raw feature vector of the  $i$ th and  $j$ th item respectively, and their link relationship matrix between them  $\mathbf{R} \in \mathbb{R}^{M \times N}$ , where,  $r_{ij} = 1$  if we observe a link between the item  $i$  and  $j$ , and  $r_{ij} = 0$  if we observe that there is no link. Unobserved links are unfilled, where  $M$  and  $N$  denote the number of items in category “A” and “B” respectively. Our goal is to learn a model to predict unfilled values based on the raw features of items and their observed links. Note that “A” and “B” may be heterogeneous or homogeneous in our model.

In our RGDBN model which is shown in Fig. 5, we assume that each of the two kinds of items is represented by a set of latent features. Let  $\mathbf{Z}^A$  and  $\mathbf{Z}^B$  be the binary matrices to represent the latent features of the two kinds of items, where each row corresponds to an item and each column corresponds to a feature such that  $z_{ik} = 1$  if the  $i$ th item has the feature  $k$  and  $z_{ik} = 0$  otherwise. Let  $\mathbf{z}_i^A$  and  $\mathbf{z}_j^B$  denote the feature vectors corresponding to the  $i$ th item in “A” item and  $j$ th item in the “B” item. For simplicity, we use  $\mathbf{z}_i$  and  $\mathbf{z}_j$  to represent them.

In the proposed generative model, the prior distribution over the latent features  $P(\mathbf{Z}^A)$  and  $P(\mathbf{Z}^B)$  specify the number of features and their probabilistic distributions, while the distributions over observed low-level features conditioned on those latent features  $P(\mathbf{X}|\mathbf{Z}^A)$ ,  $P(\mathbf{Y}|\mathbf{Z}^B)$  and relationships  $P(\mathbf{R}|\mathbf{Z}^A, \mathbf{Z}^B)$  determine how these latent features generate the observed raw features and their relationships.

On one hand, the latent features will generate the relation links in the proposed generative model, and the probability of having a link from item  $i$  to  $j$  is entirely determined by the combined effect of all pairwise latent feature interactions. Let  $\mathbf{U}$  be a real-valued weight matrix where  $u_{kk'}$  is the weight that affects the probability of there being a link from item  $i$  to item  $j$  if item  $i$  has feature  $k$  and item  $j$  has feature  $k'$ . We assume that links are independent conditioned on  $P(\mathbf{Z}^A)$ ,  $P(\mathbf{Z}^B)$  and  $\mathbf{U}$ , and only the features of item  $i$  and  $j$  influence the probability of a link between those items. This defines the likelihood

$$P(\mathbf{R}|\mathbf{Z}^A, \mathbf{Z}^B, \mathbf{U}) = \prod_{i,j} P(r_{ij}|\mathbf{Z}^A, \mathbf{Z}^B, \mathbf{U}) \quad (5)$$

where the product ranges over all pairs of items. Given the feature matrix  $\mathbf{Z}^A$ ,  $\mathbf{Z}^B$  and weight matrix  $\mathbf{U}$ , the probability that there is a link from item  $i$  to item  $j$  is

$$\Pr(r_{ij} = 1|\mathbf{Z}^A, \mathbf{Z}^B, \mathbf{U}) = \text{sigmoid}(\mathbf{z}_i \mathbf{U} \mathbf{z}_j^T) \quad (6)$$

where  $\text{sigmoid}()$  denotes the logistic function, and we give a prior on  $\mathbf{U}$  with  $\mathcal{N}(0, \sigma_u^2)$  for each entry  $u_{kk'}$  independently.

On the other hand, motivated by that there should be strong relationship for items with similar content features and the content information plays an important role in link analysis between items, the latent features also generate the observed raw low-level content features  $\mathbf{x}_i$  and  $\mathbf{y}_j$ . The generative process is as Equation (2) within the latent feature framework introduced in Section II.

In our model, the latent features are the key factor that determines the link relationship. However, on one hand, the existence of collective effect leads to that the social media data do not exist independently but influence with each other, which means that what latent features an item has are not only influenced by the content of the item itself, but also the features the other items possess. For example, an item should possess some “popular” features that most of items have with high probabilities though these features are not reflected from its content information. Therefore, the collective effect should be considered into the latent features for each items. On the other hand, from the model expressivity angle, if the number of the latent features is fixed in advance, the model is less flexible. Based on the above ideas, we want to seek a prior for the latent feature to allow us to simultaneously infer the number of features at the same time we learn which items have each feature. The Indian buffet process (IBP) is such a prior, and it offers hopes to integrate the collective effect into the unified latent features for our model.

IBP [35] is a stochastic process defining a probability distribution on binary matrices with a finite number of rows and an unbounded number of columns. A feature matrix drawn from it for a finite number of items will only have a finite number of

non-zero features. The generative process to sample matrices from the IBP can be described through a culinary metaphor. Each row of  $\mathbf{Z}$  corresponds to a diner at an Indian buffet and each column corresponds to a dish at the infinitely long buffet. If a customer takes a particular dish, then the entry that corresponds to the customer's row and the dish's column is 1 and the entry is 0 otherwise. The culinary metaphor describes how people choose the dishes. In the IBP, the first customer chooses a  $\text{Poisson}(\alpha)$  number of dishes to sample, where  $\alpha$  is a parameter of the IBP. The  $i$ th customer tries each previously sampled dish with probability proportional to the number of people that have already tried the dish and then samples a  $\text{Poisson}(\alpha/i)$  number of new dishes.

We use the IBP prior on  $\mathbf{Z}^A$  and  $\mathbf{Z}^B$  respectively by regarding the items corresponding to the customers, and latent features corresponding to the dishes. If some features are possessed by a number of items, then the current item also has the features with probability proportional to the number of items. We describe this prior as

$$\mathbf{Z}^A \sim \text{IBP}(\alpha) \quad (7)$$

$$\mathbf{Z}^B \sim \text{IBP}(\beta) \quad (8)$$

where  $\alpha$  and  $\beta$  control the growth rate of the new features.

Combining the IBP prior and deep generative learning process under the Bayesian framework, the collective effect and media content information are integrated into the unified latent features  $\mathbf{Z}^A$  and  $\mathbf{Z}^B$  which are just what we emphasize in this paper for social media network analysis.

Exact inference for our model is intractable. For simplicity's sake, we firstly learn the high-level representation (in layer  $\mathbf{h}^{(L)}$ ) from the content feature for each item, and then do the posterior inference on  $\mathbf{Z}^A$ ,  $\mathbf{Z}^B$  and  $\mathbf{U}$  using the approximate Markov Chain Monte Carlo [36] algorithm.

The learning of the high-level representation from content feature is mainly to learn the parameters (including weights between adjacent layers and biases). Similar to the DBN, the whole unsupervised learning processes include layer-wise pre-training described in Section II and a following fine-tuning stage. In the fine-tuning stage, we use the "wake-sleep" algorithm [37] to adjust the parameters of all the layers globally in the fine-tuning stage. In this training process, the "recognition" weights that are used for inference are untied from the "generative" weights that define the model.

For the posterior inference on  $\mathbf{Z}^A$ ,  $\mathbf{Z}^B$  and  $\mathbf{U}$ , we illustrate the process by taking the  $\mathbf{Z}^A$  for example, and the process for  $\mathbf{Z}^B$  is also in a similar way. Given  $\mathbf{U}$ , and the parameters  $\Theta$  (weights and biases) between layer  $\mathbf{Z}^A$  and  $\mathbf{h}^{(L)}$ , we sample the  $\mathbf{Z}^A$  by starting with an arbitrary binary matrix, and then iterate through the rows of the matrix. For non-zero columns  $k$  and row  $i$ ,

$$\begin{aligned} & \Pr(z_{ik} = 1 | \mathbf{Z}_{-ik}, \mathbf{z}_j, \mathbf{h}_i^{(L)}, \mathbf{h}_j^{(L)}, R) \\ & \propto \Pr(z_{ik} = 1 | \mathbf{Z}_{-ik},) P(r_{ij}, \mathbf{h}_i^{(L)} | z_{ik} = 1, \mathbf{Z}_{-ik}, \mathbf{z}_j, \mathbf{U}, \Theta) \\ & \propto \Pr(z_{ik} = 1 | \mathbf{Z}_{-ik},) P(r_{ij} | z_{ik} = 1, \mathbf{Z}_{-ik}, \mathbf{z}_j, \mathbf{U}) \\ & \quad \times P(\mathbf{h}_i^{(L)} | \mathbf{z}_i, \Theta) \\ & \propto m_k \cdot P(r_{ij} | z_{ik} = 1, \mathbf{Z}_{-ik}, \mathbf{z}_j, \mathbf{U}) P(\mathbf{h}_i^{(L)} | \mathbf{z}_i, \Theta) \end{aligned} \quad (9)$$

where  $m_k$  denotes the number of non-zero columns entries in column  $k$  excluding row  $i$ , and  $\mathbf{Z}_{-ik}$  is the set of assignments of other items, not including  $i$ , for feature  $k$ . Note that Equation (9) indicates the three parts we try to model:  $\Pr(z_{ik} = 1 | \mathbf{Z}_{-ik},)$  captures the collective effect, while  $P(r_{ij} | z_{ik} = 1, \mathbf{Z}_{-ik}, \mathbf{z}_j, \mathbf{U})$  makes the latent more likely to generate the relationship, and the part  $P(\mathbf{h}_i^{(L)} | \mathbf{z}_i)$  to generate the raw content feature along the deep architecture. Similarly the number of new features associated with item  $i$  should be drawn from a  $\text{Poisson}(\alpha/M)$  distribution. Given the  $\mathbf{Z}^A$  and  $\mathbf{Z}^B$  to sample  $\mathbf{U}$ , we simply use the method in [38] to use a Metropolis-Hastings step for each weight in which we propose a new weight from a normal distribution centered around the older one. Given  $\mathbf{Z}$ , we update values of  $\Theta$  in the same way as in contrastive divergence [39] algorithm.

Based on the results of posterior inference, it is easy to predict the missing data values given the observed link. We collect  $T$  samples  $\{\mathbf{Z}^{A,(t)}, \mathbf{Z}^{B,(t)}, \mathbf{U}^{(t)}\}_{t=1}^T$ , and estimate the predictive distribution of a unfilled link values as the average of the predictive distributions for each of the collected samples. Assuming that we want to predict the missing link  $r_{ij}$  between items  $i$  and  $j$ , the approximate predictive distribution will be as follows:

$$\begin{aligned} & \Pr(r_{ij} = 1 | \mathbf{R}_{train}, \mathbf{x}_i, \mathbf{y}_j) \\ & \approx \frac{1}{T} \sum_t \Pr(r_{ij} = 1 | \mathbf{z}_i^{(t)}, \mathbf{z}_j^{(t)}, \mathbf{U}^{(t)}) \end{aligned} \quad (10)$$

In conclusion, the proposed Relational Generative Deep Belief Nets model is summarized in Algorithm 1.

---

**Algorithm 1:** Relational Generative Deep Belief Nets

---

**Input:** Initial relation matrix  $\mathbf{R}$ ; low level features matrix of "A" items  $\mathbf{X}$ ; low level features matrix of "B" items  $\mathbf{Y}$ ; Number of deep network layers  $L$ ; Random initial bias parameters for units in layer  $\mathbf{h}^{(0)}, \mathbf{h}^{(1)}, \dots, \mathbf{h}^{(L)}$ ; Random initial weight parameters  $W^{(0)}, W^{(1)}, W^{(2)}, \dots, W^{(L-1)}$ ; Random initial latent feature matrix  $\mathbf{Z}^A$ ; Random initial latent feature matrix  $\mathbf{Z}^B$ .

**Output:** Unfilled link values  $r_{ij}$  of  $\mathbf{R}$

- 1 **for each layer from**  $\mathbf{h}^{(1)}$  **to**  $\mathbf{h}^{(L)}$  **do**
- 2     Greedy layer-wise pre-training by learning a RBM at a time;
- 3     wake-sleep algorithm for fine-tuning the parameters of deep nets;
- 4 **for a number of iterations do**
- 5     **for each row**  $i$  **in**  $\mathbf{Z}^A$  **do**
- 6         sample  $z_{ik}$ ;
- 7     **for each row**  $j$  **in**  $\mathbf{Z}^B$  **do**
- 8         sample  $z_{jk}$ ;
- 9     sample  $\mathbf{U}$ ;
- 10    sample  $\Theta$ ;

```

11  for each test entry  $r_{ij}$  do
12    for collection number  $T$  do
13      collect  $\mathbf{z}_i^{(t)}$  and  $\mathbf{z}_j^{(t)}$ ;
14      collect  $\mathbf{U}^{(t)}$ ;
15  Return test unfilled link values

```

$$r_{ij} = \frac{1}{T} \sum_t \Pr \left( r_{ij} = 1 | \mathbf{z}_i^{(t)}, \mathbf{z}_j^{(t)}, \mathbf{U}^{(t)} \right);$$

## B. Experiments

In order to evaluate the effectiveness of our proposed RGDBN model, we conduct a series of experiments on both self-developed dataset from Flickr for homogeneous user-user relationship and public MIRFlickr-25000 dataset [40] for heterogeneous image-tag relationship.

For both experiments of link prediction, we treat it as positive sample if we observe  $r_{ij} = 1$  in the dataset, and negative sample if  $r_{ij} = 0$ . The prediction results are ranked in increasing sequences according to the prediction probabilities. AUC, the area under the ROC (Receiver Operating Characteristic) curve, serves as the performance metric, which is calculated by the following equation:

$$\text{AUC} = \frac{S_0 - n_0(n_0 + 1)/2}{n_0 n_1} \quad (11)$$

where  $n_0$  and  $n_1$  are the number of positive and negative samples respectively, and  $S_0 = \sum \text{Pos}(q)$ , where  $\text{Pos}(q)$  is the rank position of the  $q$ th positive example in the ranked list. The more positive examples are ranked higher (with higher probability of being a positive example), the higher the term  $S_0$ . Therefore, AUC measures the quality of ranking, which is a more elegant metric in our problem than accuracy.

1) *Link Analysis Between Homogeneous Items*: We first conduct the experiment for link prediction between homogeneous items: user recommendation task. In our experiment, we crawl 3,500 users' data from Flickr, and the data information includes user's name, contact list, tag list in his/her profile, uploaded and "like" photos and related tags for each photo. For the data, we initially select about ten users with more than 30 friends, and crawl all the friends of them and the associated information, and then the friends of the friends, and goes on. Finally, we select the top 3,500 users with noticeable number of friends. The contact number distribution for users is shown in Fig. 6, from which we can see that most users' contact numbers are less than 100. In our experiment, the 1,000 most frequent tags construct the vocabulary, and each user is represented by the tags in user's profile and that associated his/her uploaded and "like" photos. We use TF-IDF to construct the low level feature for each user. In the contact list of user  $i$ , if user  $j$  is in it, we set  $r_{ij} = 1$ ; if the user  $j$  is not in it, and meanwhile, there is no common tag for them, we set  $r_{ij} = 0$ . The remaining entries of matrix  $\mathbf{R}$  are left as unfilled. Note that, the matrix  $\mathbf{R}$  is not symmetric, which indicates that the fact user  $i$  is in the contact list of user  $j$ , does not guarantee user  $j$  is in the contact list of user  $i$ . We hold out

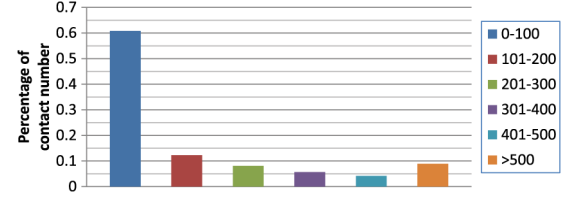


Fig. 6. Histogram of contact number for users.

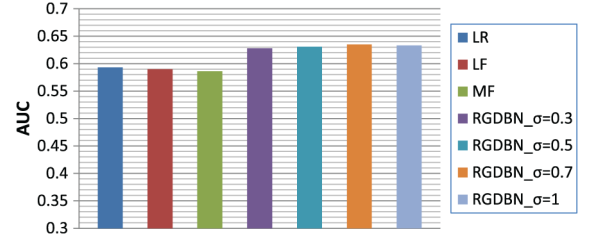


Fig. 7. AUC results for link prediction between user-user relationship.

20% of the data during training and report the AUC results for the held-out data.

In the proposed RGDBN model, we set the hyper-parameter both  $\alpha$  and  $\beta$  as 1. There are 4 layers in the deep architecture, and the number of units for layer  $\mathbf{h}^{(0)}$ ,  $\mathbf{h}^{(1)}$ ,  $\mathbf{h}^{(2)}$ ,  $\mathbf{h}^{(3)}$  is 1,000, 800, 500, 300 respectively.

We then compare our model against logistic regression (LR) and matrix factorization (MF). In logistic regression, we regard the link problem as supervised binary classification problem, and each data point corresponding to a pair of users. Similar to the related work [41], the features include the number of common tags, sum of tags, Jaccard's coefficient, shortest distance, and Katz, etc. In matrix factorization, we take the relationship matrix as input, and non-negative matrix factorization is used. For RGDBN, we conduct the experiments with different  $\sigma_u$  values 0.3, 0.5, 0.7, 1 for the prior on  $\mathbf{U}$ , and the AUC results are illustrated in Fig. 7.

We observe that the proposed RGDBN model outperforms both the matrix factorization method and classification-based method using the topographical features and proximity features. From the view of features used in these models, we think that the features in classification-based method are raw features extracted from the graph topology and attribute, and those in matrix factorization method could be considered as a shallow learned feature based on the low-rank approximation, while the latent features learned in RGDBN model are deep learned and more representative.

2) *Link Analysis Between Heterogeneous Items*: We also conduct the experiment for link analysis applied in social image annotation between heterogeneous items for image-tag relationship on public collection dataset [40]. The data set consist of 25,000 annotated images which are collected from Flickr along with their tags. The average number of tags per image is 8.94. It includes 24 labeled categories and each image may belong to multiple categories. Fig. 8 shows some sample images from the dataset.

In our experiment, we consider analyzing the image-tag relationship. For image  $i$ , if it is associated with the ground truth

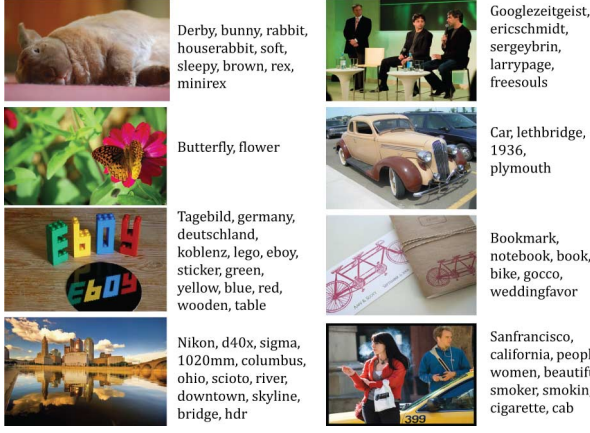


Fig. 8. Sample images from MIRFLICKR-25000 dataset.

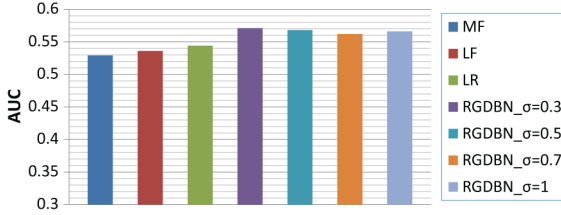


Fig. 9. AUC results for link prediction between image-tag relationship.

label or tag  $j$ , we set  $r_{ij} = 1$ , otherwise  $r_{ij} = 0$ . Only the 800 most frequent tags are considered. For the hold-out data, we select 20% of training data associated with the labels.

In our proposed RGDBN model, we set the hyper-parameter  $\alpha = 1$  and  $\beta = 0.5$  respectively. There are 4 layers for the image deep architecture, and the number of units is 1,024, 500, 200, 50 respectively, while for the tag deep architecture, we only use 3 layers with the number of units 800, 200, 50 respectively. The reasons are two-fold. First, we think that the tag and the low-level visual feature are not in the same semantic level. Tag is closer to the learned latent feature level with respect to visual feature. Second, in the experimental dataset, the number of tag training samples is less than that of images. Reducing the number of layers leads to less number of parameters to train. We use HOG as raw feature for images, while for the tags, we consider the co-occurrence patterns among the tags and take the normalized co-occurrence counts of tags as features.

For comparison, other than classic matrix factorization (MF) method, we regard the image-tag link prediction as an image classification problem where we take each image as a data point. If image  $i$  is classified into tag  $j$ , we regard there is a link between image  $i$  and tag  $j$ . The logistic regression model is used. As in the experiment in the link analysis between homogeneous items, we also compare our model with the only latent feature-based method and conduct the experiments with different  $\sigma_u$  values 0.3, 0.5, 0.7, 1 for the prior on  $\mathbf{U}$ , and the AUC results are shown in Fig. 9.

We can see that though our RGDBN outperforms the other methods, the whole performance is worse than that in the experiment with homogeneous items. In our opinion, other than

the reason of heterogeneity, the link sparsity is also an important factor. The fact that the average number of tags per image is less than 10 indicates that most of the entries in  $\mathbf{R}$  are zero. Since we expect the model to produce high probability in generating the observed relationship, the derived low probability in Equation (6) results in a small number of latent features to be learned in the posterior sample process. Under such circumstance, the advantage of the latent feature is limited.

#### IV. LATENT FEATURE LEARNING FOR IMAGE RETRIEVAL

In this section, from the perspective of unified multimodal feature learning, we design a Multimodal Deep Learning to Lambda Rank (MDLLR) model by instantiating the latent feature learning framework to address the image retrieval task. In the MDLLR model, latent features are learned from the multimodal media content and discriminative lambda loss is used for supervised parameters learning.

##### A. Multimodal Deep Learning to Lambda Rank

Text-based image retrieval has found growing importance due to its popularity through Web image search engines. In the task, the input is a text query and the retrieval system outputs a ranking set of images in which the images relevant to the query should appear above the others. Assume that we observe a set of queries  $\mathbf{Q}$  and images  $\mathbf{X}$ , and their low-level raw features are  $\{q_i\}_{i=1}^M$  and  $\{x_j\}_{j=1}^N$  respectively. When the  $q_i$  is as query for image retrieval, the image  $x_j$  is clicked by  $s(x_j, q_i)$  times. The task is that for another query  $q$ , we need to predict the image ranking list for the corresponding query  $q$ .

In the information retrieval system, ranking is the core factor, for the IR evaluation is directly based on the ranking. Learning to rank is a kind of classical machine learning method for ranking, and they integrate the query features into the document features, which is available in the text retrieval task. However, in the image retrieval task, the different statistical properties of different modal data make it a difficult task. The proposed latent feature learning framework provides a way to learn the multimodal unified representation for query and image. Combining the deep multimodal architecture and lambda ranking, we design a Multimodal Deep Learning to Lambda Rank (MDLLR) model to solve the image retrieval task.

In our designed MDLLR model which is shown in Fig. 10, separate networks are firstly utilized to learn the latent representations of image and query respectively, and then their joint representation is learned through a combined deep network. By the multimodal deep network, the latent features are learned from multimodal heterogeneous raw features through the multimodal deep network. Due to its ability of feature fusion, the unified latent representation integrates the information of both the image and query.

In the feature fusion layers, there is a multimodal RBM, where high level image feature  $\mathbf{h}_x^L$  and query feature  $\mathbf{h}_q^L$  are fused by treating  $\mathbf{h}_x^L$  and  $\mathbf{h}_q^L$  as visible layer together and treating  $\mathbf{h}^0$  as hidden layer. Assume that the dimensions of layers  $\mathbf{h}_x^L$ ,  $\mathbf{h}_q^L$ ,  $\mathbf{h}^0$  are  $G$ ,  $D$ ,  $F$  respectively, then the joint

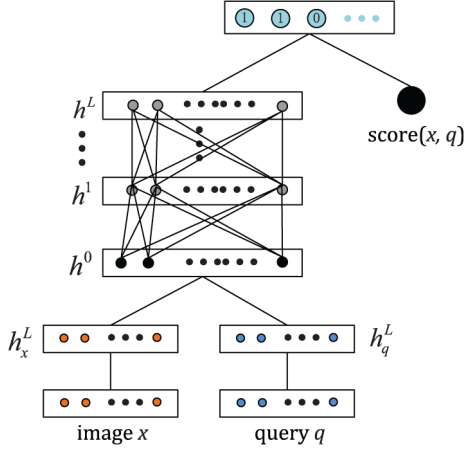


Fig. 10. Multimodal deep learning to lambda rank model.

energy configuration of the multimodal RBM is defined as follows:

$$\begin{aligned}
 E(\mathbf{h}_x^L, \mathbf{h}_q^L, \mathbf{h}^0) = & - \sum_{i=1}^D \sum_{j=1}^F h_{q,i}^L W_{x,ij} h_j^0 - \sum_{g=1}^G \sum_{j=1}^F h_{x,g}^L \\
 & \cdot W_{q,ij} h_j^0 - \sum_{i=1}^D a_{q,i} h_{q,i}^L - \sum_{j=1}^F a_j^0 h_j^0 \\
 & - \sum_{g=1}^G a_{x,g} h_{x,g}^L. \quad (12)
 \end{aligned}$$

Exact inference likelihood learning in this model is intractable. An analogous Contrastive Divergence approximation algorithm [39] is applied.

Since the latent features are linear with the click scores  $s(x, q)$ , the linear regression method is used in our model.

The training includes pre-training and fine-tuning two stages. Pre-training follows the way described in Section II. In the fine-tuning stage, the loss between the model ranking list where the images are sorted by the model scores and groundtruth ranking list is used to tune the parameters of the whole network globally. Here, we use the Lambda rank method [34] to compute the pairwise loss. For a query  $q$ , a lambda function which is the pairwise loss between the returned image  $u$  and  $v$  is defined as

$$\begin{aligned}
 \lambda_{u,v} = & Z \cdot \frac{2^{l_u} - 2^{l_v}}{1 + \exp\{s(u, q) - s(v, q)\}} \\
 & \cdot \left( \log \left( \frac{1}{1 + r(u)} \right) - \log \left( \frac{1}{1 + r(v)} \right) \right) \quad (13)
 \end{aligned}$$

where  $Z$  is the normalization factor for the query,  $s(u, q)$  and  $s(v, q)$  denote the model scores between query  $q$  and image  $u$ ,  $v$  respectively, which are the values of linear regression to the latent features and the function of parameters in the deep network. The update of these scores will lead to the update of the parameters, in which way, the parameters are learned.  $r(u)$  and  $r(v)$  are the position in the ranking list corresponding to image  $u$  and  $v$  for query  $q$ , and  $l_u, l_v$  denote their relevance labels. For each pair of image  $u$  and  $v$ , in each round of optimization, their scores are updated by  $+\lambda_{u,v}$  and  $-\lambda_{u,v}$  respectively. In

our model, we use batch learning per query-image where we accumulate  $\lambda$  for each image summing its contributions from all the image pairs, and then do the update, which speedup the training time.

In conclusion, the proposed Multimodal Deep Learning to Lambda Rank model is summarized in Algorithm 2.

---

**Algorithm 2:** Multimodal Deep Learning to Lambda Rank

---

**Input:** Number of separate deep network layers  $L_{q,x}$ ; Number of combine deep network layers  $L$ ; Random initial bias parameters for units in each layer; Random initial weight parameters for each RBM; Image raw features  $\{x_j\}_{j=1}^N$ ; Query raw features  $\{q_i\}_{i=1}^M$ ; Click score  $\{s(x_j, q_i)\}_{i,j}$ ;

**Output:** Image ranking lists for test queries;

- 1 **for each layer from  $\mathbf{h}_{q,x}^{(1)}$  to  $\mathbf{h}_{q,x}^{(L)}$  do**
  - 2 Greedy layer-wise pre-training by learning a separate RBM at a time;
  - 3 Train the multimodal RBM;
  - 4 **for each layer from  $\mathbf{h}^{(1)}$  to  $\mathbf{h}^{(L)}$  do**
  - 5 Greedy layer-wise pre-training by learning a combine RBM at a time;
  - 6 Compute the Lambda loss based on the equation (13)
  - 7 **for each query  $q$  in training data do**
  - 8 **for each image  $u$  in image list of the query  $q$  do**
  - 9 Compute the  $s(u, q)$ ;
  - 10 **for each image  $v (v \neq u)$  in image list of  $q$  do**
  - 11 Compute the  $s(v, q)$ ;
  - 12 Compute the  $\lambda_{u,v}$ ;
  - 13 Compute the  $\lambda_u = \sum_v \lambda_{u,v}$ ;
  - 14 Backpropagation with loss  $\lambda_u$ ;
  - 15 **for each query  $q$  in test data do**
  - 16 **for each image  $x$ ; do**
  - 17 Compute the  $s(x, q)$ ;
  - 18 Sort the images according to  $s(x, q)$ ;
  - 19 Return the image ranking list for query  $q$ ;
- 

## B. Experiment

In order to evaluate the effectiveness of the novel instantiated MDLLR model, we conduct several experiments on the MSR-Bing Image Retrieval Challenge dataset. The data provided by Microsoft includes the training set which is a sample of Bing user click log and the development dataset which is a manually labeled set. Each triads in the training set is a clicked image-query component which contains image, query, and click count. Each record in the development set includes query, image and judgment (relevance level). There are about 23 million triads

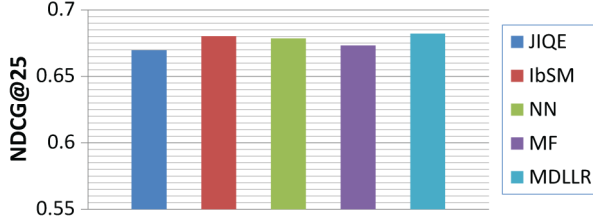


Fig. 11. NDCG@25 results for image retrieval.

and 1 million images in the training data. In the develop dataset, the number of queries is 1,000, and the number of triads is 79,926. Note that the online test dataset of the challenge is not publicly available. Hence we conduct the experiments on the training set and development set.

In the development data, there are three relevance levels: *excellent*, *good*, and *bad*. NDCG(Normalized Discounted Cumulative Gain)@25 is used for evaluation metric, which is computed by

$$NDCG@25 = Z \cdot \sum_{j=1}^{25} (2^{l_j} - 1) / \log(1 + j) \quad (14)$$

where  $Z$  is the normalization factor,  $j$  is position in the ranking list and  $l_j$  denotes the relevance label of image in the position  $j$ . From the evaluation metric, we can see that NDCG is position sensitive, and ranking is a very important factor for image retrieval.

In our experiments, we train the individual models on the training set and half of the development set is used as the test set. The click counts between images and queries in the training set are transformed into the ground truth for training. The HOG and TF-IDF are used as low-level raw features of image and query respectively. There are 2 hidden separate layers for image and query respectively and 2 hidden combined layers in the deep architecture.

We compare our MDLLR model against classic Matrix Factorization (MF) [42] model, Nearest Neighbor-based (NN) method, Image-based Score model (IbSM) and Joint Image-Query Embedding (JIQE) method. In the Nearest Neighbor-based (NN) method, we use the test image-query pair  $(x; q)$  to search the similar neighbors  $(\mathbf{X}; \mathbf{Q})$  using the image annotation set and calculate the similarity between the test pair and neighbor pairs. We combine the text matching method and LSH to complete the nearest neighbor search. In the Image-based Score model (IbSM), we first use the test image to retrieve the  $t$  nearest neighbor images using the image annotation set and then measure the text similarity between the query and the associated tags with the  $t$  nearest neighbor set  $H$ . BM25 score is used for computing text similarity. The Joint Image-Query Embedding method learns a mapping onto a feature space where images and text queries are both embedded. The mapping functions for images and queries are learned jointly to optimize the supervised loss for the image-query ranking. This could be regarded as a latent feature learning method based on the shallow architecture. The evaluation results are shown in Fig. 11.

From the experimental results, we can see that our latent feature learning-based MDLLR model outperforms the other classic methods. We think that the effectiveness mainly owes to the learned latent feature based on the multimodal deep architecture which has strong ability of feature fusion. In the process, each layer learns successively higher-level representations and removes modality-specific correlations. Eventually, the latent feature layer in the network can be seen as a modality-free latent representation which is more representative than those learned by shallow architectures.

## V. RELATED WORK

In this section, we give a brief review of the related work about link analysis in social media network and image retrieval methods.

### A. Link Analysis in Social Media Network

The existing link analysis techniques in social media network can be roughly divided into graphical topology-based, low-rank approximation-based, and Bayesian relational methods.

In the graphical topology-based methods, the proximity features including common keywords [41] and topological features [43]–[4] are usually extracted to represent linked pairwise items. These features are then used in conventional models for solution. For example, [41] utilizes these features for supervised pairwise classification.

Low-rank approximation-based methods represent the observed links into graph adjacency matrices, and utilize matrix factorization techniques to reconstruct the original matrices by low-rank approximation [45]. This kinds of methods can be regarded as the special case of latent feature learning, i.e., with only one layer.

Bayesian relational models assume that there are some latent factors to generate the observed links. The basic idea is to set prior on these factors, with typical work including Infinite Relational Model [46], Mixed Membership Stochastic Block-model [47], Nonparametric Latent Feature Model [38], etc. These models are of some flexibility for modeling the observed links. However, in these models, the latent factors are not in the unified semantic level, which cannot handle the complex links in social media network, especially for the heterogeneous ones.

### B. Image Retrieval

The related work for image retrieval can be divided into two categories: text-based image retrieval and content-based image retrieval. There are some surveys about content-based image retrieval [48][49]. As this paper focuses on the former case, in this section, we briefly introduce the related work about text-based image retrieval.

In the task of text-based image retrieval, the system is given a set of images and some query concepts. For a text query, it outputs a list of ranked images according to the relevance to the query. For the image retrieval system, the similarity between the queries and images plays an important role, for it affects the final ranking performance directly. Roughly, two types of methods to model the query-image score have been introduced in the literature: annotation-based methods and direct methods.

In the annotation-based methods, there is an intermediate image annotation step, and the relevance computing is then based on the image annotation. Various models can be used for annotation, such as SVM [50], k-NN [51] and so on. For the application, the images in the test dataset are ranked according to the relevance score outputted by the annotation corresponding to the query concept [52].

Different from the annotation-based methods, direct methods do not rely on the intermediate annotation task. Instead, they model the query-image score directly. [53] uses matrix factorization to model the relationship between visual content and the text keywords, and the similarity can be measured in the latent space. [54] joins the image and word queries discrimination to determine the relevance for image retrieval. [55] adopts a learning criterion related to the final retrieval performance to rank the images from text queries.

## VI. CONCLUSION

The novel idea of this work is to tackle the analysis and applications in social media with latent features automatically learned from a specially designed deep architecture. This goes against the current trend which advocates the use of pre-defined features and focuses on higher-level model development. Based on the discussions on the characteristics in social media data and challenges in social media analysis, we justify why it is necessary to bring attention to this new research line and what benefits of this new research line may bring to the social media community and the broader social media applications.

We have introduced a unified deep architecture-based latent feature learning framework in social media. To test the idea and the proposed framework, we design a novel RGDBN model and apply it in the link analysis tasks. Our analysis confirmed that the feature learning process preserves sufficient information to capture the interactions between heterogeneous as well as homogeneous data. We demonstrate that the learned latent feature is more representative to embed the linked media content, and more effective to generate the observed links, which outperforms non-feature learning methods in user recommendation and image annotation tasks. Meanwhile, we design a multimodal deep learning to lambda rank model for image retrieval within the latent feature learning framework, and the experimental results validate the effectiveness of our proposed framework.

Future work will focus, first of all, on evaluating the idea and framework in more social media tasks and applications, investigating in particular whether the derived latent feature is helpful to represent the complexly interconnected social media data. Besides, reasoning and inference are critical to a practical probabilistic graphical model. We are working towards developing more efficient posterior inference techniques for model learning.

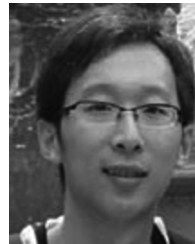
## REFERENCES

- [1] T. Ahlqvist, "Social media roadmaps: Exploring the futures triggered by social media," *VTT*, 2008.
- [2] L. Cao, J. Yu, J. Luo, and T. S. Huang, "Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 125–134.
- [3] J. Sang and C. Xu, "Right buddy makes the difference: An early exploration of social relation analysis in multimedia applications," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 19–28.
- [4] Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao, and H. Cao, "Link prediction and recommendation across heterogeneous social networks," in *Proc. IEEE 12th Int. Conf. Data Mining*, 2012, pp. 181–190.
- [5] D. Liu, G. Ye, C.-T. Chen, S. Yan, and S.-F. Chang, "Hybrid social media network," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 659–668.
- [6] H. Gao, X. Wang, J. Tang, and H. Liu, "Network denoising in social media," in *Proc. ASONAM*, 2013.
- [7] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Practical elimination of near-duplicates from web video search," in *Proc. 15th Int. Conf. Multimedia*, 2007, pp. 218–227.
- [8] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. Int. Conf. Multimedia*, 2010, pp. 461–470.
- [9] G. Friedland, O. Vinyals, and T. Darrell, "Multimodal location estimation," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1245–1252.
- [10] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. Int. Conf. Multimedia*, 2010, pp. 251–260.
- [11] L. Liu, L. Sun, Y. Rui, Y. Shi, and S. Yang, "Web video topic discovery and tracking via bipartite graph reinforcement model," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 1009–1018.
- [12] H. Sundaram, L. Xie, M. De Choudhury, Y.-R. Lin, and A. Natsev, "Multimedia semantics: Interactions between content and community," *Proc. IEEE*, vol. 100, no. 9, pp. 2737–2758, Sep. 2012.
- [13] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 363–376, Jan. 2013.
- [14] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
- [15] J. Sang, C. Xu, and J. Liu, "User-aware image tag refinement via ternary semantic analysis," *IEEE Trans. Multimedia*, vol. 14, no. 3, pt. 2, pp. 883–895, Jun. 2012.
- [16] Y.-R. Lin, J. Sun, H. Sundaram, A. Kelliher, P. Castro, and R. Konuru, "Community discovery via metagraph factorization," *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 3, p. 17, 2011.
- [17] J.-I. Biel and D. Gatica-Perez, "Vlogsense: Conversational behavior and social attention in YouTube," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 7, no. 1, p. 33, 2011.
- [18] G.-J. Qi, C. Aggarwal, Q. Tian, H. Ji, and T. S. Huang, "Exploring context and content links in social media: A latent space method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 850–862, May 2012.
- [19] Z. Yuan, J. Sang, Y. Liu, and C. Xu, "Latent feature learning in social media network," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 253–262.
- [20] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," *Advances Neural Inf. Process. Syst.*, vol. 19, p. 801, 2007.
- [21] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, pp. 498–520, 1933.
- [22] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.
- [23] A. Coates, H. Lee, and A. Y. Ng, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. Int. Conf. AI Statist.*, 2011, pp. 215–223.
- [24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
- [25] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.
- [26] G. Mesnil, Y. Dauphin, X. Glorot, S. Rifai, Y. Bengio, I. Goodfellow, E. Lavoie, X. Muller, G. Desjardins, and D. Warde-Farley *et al.*, "Unsupervised and transfer learning challenge: A deep learning approach," in *Proc. Unsupervised Transfer Learn. Workshop*, 2011.
- [27] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *Advances Neural Inf. Process. Syst.* 25, pp. 2231–2239, 2012.
- [28] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.
- [29] R. Salakhutdinov and G. Hinton, "Semantic hashing," *Int. J. Approximate Reason.*, vol. 50, no. 7, pp. 969–978, 2009.
- [30] Y. Kang and S. Choi, "Restricted deep belief networks for multi-view learning," in *Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer, 2011, pp. 130–145.
- [31] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

- [32] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in *Parallel Distributed Processing*, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA, USA: MIT Press, 1986, vol. 1, ch. 6, pp. 194–281.
- [33] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances Neural Inf. Process. Syst.*, vol. 19, p. 153, 2007.
- [34] C. J. Burges, R. Ragno, and Q. V. Le, "Learning to rank with non-smooth cost functions," *NIPS*, vol. 6, pp. 193–200, 2006.
- [35] T. Griffiths and Z. Ghahramani, "Infinite latent feature models and the indian buffet process," *Sciences New York*, vol. 18, no. GCNU TR 2005-001, p. 475, 2005.
- [36] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 1st ed. New York, NY, USA: Springer, 2004.
- [37] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, "The "wake-sleep" algorithm for unsupervised neural networks," *Science*, pp. 1158–1161, 1995.
- [38] K. T. Miller, T. L. Griffiths, and M. I. Jordan, "Nonparametric latent feature models for link prediction," *Advances Neural Inf. Process. Syst.*, vol. 22, 2009.
- [39] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, pp. 1771–1800, 2002.
- [40] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 39–43.
- [41] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *Proc. SDM 06: Workshop Link Anal., Counter-Terrorism, Security*, 2006.
- [42] S. Rendle, "Factorization machines with libFM," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, p. 57, 2012.
- [43] H. Kashima and N. Abe, "A parameterized probabilistic model of network evolution for supervised link prediction," in *Proc. 6th Int. Conf. Data Mining*, 2006, pp. 340–349.
- [44] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.
- [45] A. K. Menon and C. Elkan, "Link prediction via matrix factorization," in *Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer, 2011, pp. 437–452.
- [46] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *Proc. 21st Nat. Conf. AI*, vol. 1, pp. 381–388.
- [47] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *J. Mach. Learn. Res.*, vol. 9, pp. 1981–2014, 2008.
- [48] Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *J. Visual Commun. Image Representation*, vol. 10, no. 1, pp. 39–62, 1999.
- [49] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surveys*, vol. 40, no. 2, p. 5, 2008.
- [50] M. R. Naphade, "On supervision and statistical learning for semantic multimedia analysis," *J. Visual Commun. Image Representation*, vol. 15, no. 3, pp. 348–369, 2004.
- [51] X. Li, C. G. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1310–1322, Nov. 2009.
- [52] J. Vogel and B. Schiele, "Natural scene retrieval based on a semantic modeling step," in *Proc. Image Video Retrieval*, 2004, pp. 207–215.
- [53] J. C. Caicedo and F. A. González, "Multimodal fusion for image retrieval using matrix factorization," in *Proc. 2nd ACM Int. Conf. Multimedia Retrieval*, 2012, p. 56.
- [54] A. Lucchi and J. Weston, "Joint image and word sense discrimination for image retrieval," in *Computer Vision – ECCV 2012*. Berlin, Germany: Springer, 2012, pp. 130–143.
- [55] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1371–1384, Aug. 2008.



**Zhaoquan Yuan** received the B.E. degree from the Department of Computer Science and Technology at the University of Science and Technology of China (USTC), Hefei, China. He is currently pursuing the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include social media mining, machine learning, and pattern recognition.



**Jitao Sang** received the B.E. degree from Southeast University, Nanjing, China, in 2007, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2012. He is currently an Assistant Professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. He was awarded the Special Prize of President Scholarship by Chinese Academy of Sciences. He has published several refereed research papers and coauthored the Best Student Paper in Internet Multimedia Modeling 2013 and the Best Paper Finalist papers in ACM Multimedia 2012 and 2013. He has served as Guest Editor and as an organizing committee member in several journals and conferences. His research interests include multimedia content analysis, social media mining, and social network analysis.



**Changsheng Xu** (M'97–SM'99–F'14) is a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, and an Executive Director of the China-Singapore Institute of Digital Media, Singapore. His current research interests include multimedia content analysis/indexing/retrieval, pattern recognition, and computer vision. He is an Associate Editor of the *IEEE TRANSACTIONS ON MULTIMEDIA* and *ACM Transactions on Multimedia Computing, Communications, and Applications*. He served as a Program Chair of ACM Multimedia in 2009. He has served as an Associate Editor, Guest Editor, General Chair, Program Chair, Area/Track Chair, Special Session Organizer, Session Chair, and TPC Member for over 20 prestigious IEEE and ACM multimedia journals, conferences, and workshops. He holds 30 granted/pending patents and has published over 200 refereed research papers. He is an ACM Distinguished Scientist.



**Yan Liu** received the B.Eng. degree from the Department of Electrical Engineering, Southeast University, Nanjing, China, the M.Sc. degree from the School of Business, Nanjing University, Nanjing, China, and the Ph.D. degree from the Department of Computer Science, Columbia University, New York, NY, USA. She is an Associate Professor with the Department of Computing, Hong Kong Polytechnic University, Hong Kong. Her current research interests include multimedia content analysis, machine learning, and cognitive science.