Unconstrained Multimodal Multi-Label Learning

Yan Huang, Wei Wang, and Liang Wang, *Senior Member, IEEE*

***Abstract*—Multimodal learning has been mostly studied by assuming that multiple label assignments are independent of each other and all the modalities are available. In this paper, we consider a more general problem where the labels contain dependency relationships and some modalities are likely to be missing. To this end, we propose a multi-label conditional restricted Boltzmann machine (ML-CRBM), which handles modality completion, fusion, and multi-label prediction in a unified framework. The proposed model is able to generate missing modalities based on observed ones, by explicitly modelling and sampling their conditional distributions. After that, it can discriminatively fuse multiple modalities to obtain shared representations under the supervision of class labels. To consider the co-occurrence of the labels, the proposed model formulates the multi-label prediction as a max-margin-based multi-task learning problem. Model parameters can be jointly learned by seeking a balance between being generative for modality generation and being discriminative for label prediction. We perform a series of experiments in terms of classification, visualization, and retrieval, and the experimental results clearly demonstrate the effectiveness of our method.**

***Index Terms*—Multi-label learning, multi-task learning, multimodal learning, restricted Boltzmann machine.**

I. Introduction

IN REAL life, along with various ways of data acquisition, a concept can be represented by multiple data modalities. For example, image contents can be represented by either images themselves or their associated tags. In social network, identities are characterized by various attributes (modalities) such as age, gender and personal photo. Compared to single data modality, multiple modalities provide complementary representations of the same concept, which can greatly facilitate pattern recognition tasks such as classification and retrieval. To take advantage of this, various multimodal learning methods have recently been proposed, e.g., multiple feature concatenation [15], multiple kernel learning [6], multi-view Markov network [4], multimodal metric learning [47], multimodal deep autoencoder [24], multimodal deep Boltzmann machine [37] and transfer learning with tree-based priors [38].

To simplify the multimodal learning problem, existing methods usually assume that all the modalities to be analyzed are available. But such an assumption is not always practical since there often exist missing modalities in real world. For example, it is very easy to collect images, while obtaining their associated tags is difficult which requires a great deal of human labor. Several recent methods (e.g., [24]) have made attempts to generate the missing modalities. However, their models are especially designed to explain the complete multimodal data, rather than generate missing modalities. The generation scheme is not explicitly incorporated into their learning objectives. Accordingly, the classification performance drops dramatically when comparing with the common case that all the modalities are available. Some other methods (e.g., [6]) formulate the generation goal into their learning objectives for explicit optimization, but they only focus on solving a limited tag-generation case in a classification manner. In fact, such a classification strategy cannot be directly extended to handle more general cases since the missing modalities cannot be always regarded as class labels.

In addition to the constrained assumption about modalities, most multimodal learning methods assume that multiple class labels of multimodal data are independently of each other, which completely ignores the dependency relationships of labels. In fact, class labels usually have co-occurrence relationships. Taking bimodal data (image and tag) classification as an example, some highly correlated label pairs such as *Sky* and *Cloud*, are more likely to be assigned to one same image. Without considering such label co-occurrence, a model would assign the label *Sky* together with some irrelevant labels such as *Fish*. Unfortunately, existing methods (e.g., [24] and [37]) usually treat the multimodal multi-label learning problem as multiple independent single-label assignments by using a one-vs-all logistic classifier for each assignment, which fails to model the label co-occurrence. Moreover, under the label independence assumption, some labels with fewer training samples can not be adequately modelled, since they cannot potentially leverage knowledge from their relevant labels, e.g., learning shared features [38].

In this paper, we aim to deal with the problem of unconstrained multimodal multi-label learning when some modalities are missing or incomplete, and the labels are not necessarily independent of each other. To this end, we propose a

Manuscript received March 14, 2015; revised June 26, 2015; accepted August 15, 2015. Date of publication September 03, 2015; date of current version October 20, 2015. This work was supported by the National Natural Science Foundation of China under Grant 61175003, Grant 61202328, Grant 61572504, and Grant 61420106015, and by the National Basic Research Program of China under Grant 2012CB316300. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Guo-Jun Qi.

Y. Huang and W. Wang are with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China (e-mail: yhuang@nlpr.ia.ac.cn; wangwei@nlpr.ia.ac.cn).

L. Wang is with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, CASIA, Beijing 100190, China (e-mail: wangliang@nlpr.ia.ac.cn).

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.2015.2476658

Multi-Label Conditional Restricted Boltzmann Machine (ML-CRBM) which can jointly deal with modality completion, fusion and multi-label prediction. To generate the missing modalities, ML-CRBM models the conditional distribution over the missing modalities given observed ones, and then samples the missing modalities with a Gibbs sampler. Then, ML-CRBM discriminatively fuses all the modalities to obtain shared representations under the supervision of class labels. To further model the label co-occurrence, it handles the multi-label assignment as a multi-task learning problem where multiple labels are utilized for co-supervision.

The proposed model is learned by optimizing the objectives of both modality generation and multi-label prediction, where intractable inferences can be efficiently approximated by variational methods. We perform a series of experiments including unconstrained multimodal multi-label classification and retrieval, and label co-occurrence visualization on two publicly available datasets. The experimental results show that our method outperforms the state-of-the-art methods.

Our contributions can be summarized as follows.

1) We study a rarely investigated but practically important problem, namely unconstrained multimodal multi-label learning, and propose a new RBM-style model which can jointly handle incomplete modalities, data fusion and label relationships.
2) To the best of our knowledge, it is the first work that applies the idea of conditional Restricted Boltzmann Machine (RBM) in the context of multimodal learning, and demonstrates the validity for modality generation.
3) We find that the proposed multi-task encoding is effective for modelling label co-occurrence, and can significantly improve the classification performance.
4) Different from the existing unsupervised learning of conditional RBMs, we explore two efficient algorithms for supervised learning.

The rest of the paper is organized as follows. In Section II, we briefly review related work. In Section III, we detail the proposed model. In Section IV, we apply the proposed model to unconstrained multimodal multi-label classification and retrieval. Finally, we conclude the paper in Section V.

II. Related Work

Multimodal data analysis has been widely studied in the recent literature. By concatenating image and their associated tag features as inputs to Support Vector Machines (SVM), Huiskes *et al.* [15] significantly improve image classification. Using such feature concatenation strategy ignores the incompatibility of heterogeneous multimodal data, some other methods project multiple modalities into a shared latent space and perform multimodal tasks in this space. Guillaumin *et al.* [6] train a multiple kernel learning based classifier to score unlabeled images, and then predict labels with both images and tags. Xing *et al.* [48] present a variant of undirected graphical model, known as the multi-wing harmonium, for fusing image and text modalities. Xie and Xing [47] project multiple modalities into a shared latent space by preserving the relationships of similar and dissimilar pairs between modalities. Nguyen *et al.* [25] propose multimodal multi-instance multi-label latent Dirichlet allocation for image annotation. In addition to these shared latent space learning methods, Qi *et al.* [28] propose a novel multimodal transfer learning framework for image classification, which can effectively transfer semantics from texts to images via cross-domain label propagation.

Since the resurgence of deep neural network in 2006, several deep models have been proposed for multimodal learning. By connecting all the multimodal inputs to a shared hidden layer, Ngiam *et al.* [24] extend a multimodal version of deep autoencoder [10] to fuse audio and video modalities for further classification and retrieval tasks. Wang *et al.* [45] exploit stacked autoencoders as nonlinear mapping functions to project heterogeneous features into a common latent space, which can capture both intra-modal and inter-modal semantic relationships for multimodal retrieval. Srivastava and Salakhutdinov [36], [37] propose multimodal deep belief nets and multimodal deep Boltzmann machines which serve as generative models to well explain multimodal inputs.

The above methods make a rigid assumption that all the data modalities are available. However, in practical applications, some modalities are often missing or incomplete. Srivastava and Salakhutdinov [37] make attempts to perform experiments with missing modalities, but the aim of their model is to encode the complete multimodal inputs but not generate the missing modalities. To effectively infer the missing modalities, Sohn *et al.* [34] train a multimodal deep recurrent neural network by minimizing the variation of information of all the modalities. Different from these models, our method takes advantage of the homogeneity of multimodal data, and generates the missing modalities from the observed ones by explicitly optimizing their conditional distributions.

Multimodal data usually belong to multiple classes, which makes multimodal learning become the problem of multimodal multi-label learning. Since the topics covered in the multi-label learning literature are numerous, readers may refer to [52] for a comprehensive introduction. We will next focus on reviewing some works that handle multi-label learning based on deep learning models. To directly apply deep neural network to multi-label learning, Zhang and Zhou [51] propose a novel neural network architecture which aims to minimize pairwise ranking error of multiple assigned labels. From the prospective of multi-task learning, Huang *et al.* [13] develop a multi-task deep neural network which decomposes the multi-label assignment problem into multiple tasks, each of which is a binary classification. Kiros and Szepesvari [16] take advantage of convolutional neural network [17] to learn feature representations for images, and then exploit existing TagProp [44] for image annotation.

In the context of multimodal learning, most methods usually handle the multimodal multi-label learning problem in a multi-class learning way, i.e., assigning multiple labels to multimodal data independently, which completely ignores potential relationships among labels. To deal with this issue, Srivastava and Salakhutdinov [38] formulate the label hierarchical relationship as a tree hierarchy, and then transfer knowledge between class labels for improving classification. Xu *et al.* [49] model the label correlation as a semi-supervised label diffusion process on a unified bi-relational graph, and utilize multiple

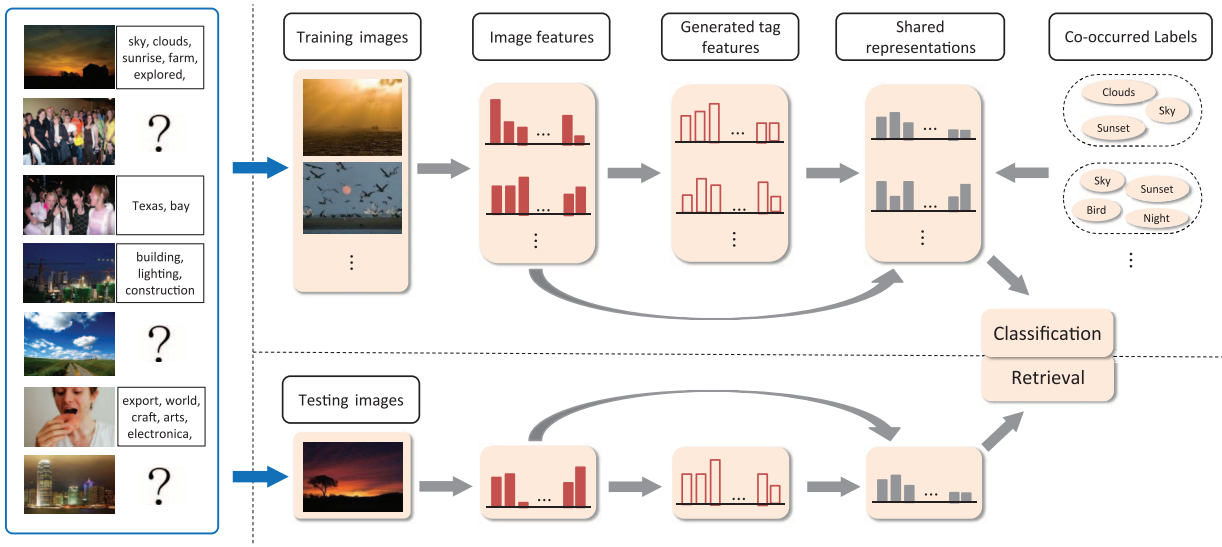


Fig. 1. Pipeline of unconstrained multimodal multi-label learning. Given incomplete multimodal data (e.g., bimodal image and tag), during training, we first extract features for images and then use them to generate missing tag features. All the features are discriminatively fused into shared representations by using co-occurred labels as supervision. During testing, given only images, the model can generate the corresponding missing tag features, and then obtain the shared representations to perform the classification and retrieval tasks.

data modalities for image classification. Qi *et al.* [29] propose a correlative multi-label framework to simultaneously classify video concepts and model their correlations. By considering the multi-level semantic relationship among category labels, Hua *et al.* [11] perform cross-modal correlation learning using adaptive hierarchical semantic aggregation. Qi *et al.* [30] exploit a two-dimensional multi-label active learning algorithm for image annotation, where only partial selected samples need to be labeled and the remaining samples can be inferred using learned label correlations. Our method proposes an alternative to automatically learn the label co-occurrence with a multi-task learning framework, in which multiple labels are jointly utilized to co-supervise the discriminative fusion of multimodal data.

Our model is also related to RBM [33] based methods, especially Conditional RBM [40] and Classification RBM [18]. Sutskever and Hinton propose a Conditional RBM for sequence modelling, where the historical visible variables are treated as dynamically changing biases for current variables by directed connections. Such a conditional scheme is further applied to human motion analysis [42], motion style modelling [43] and collaborative filtering [31]. To the best of our knowledge, it has not yet been investigated in the context of multimodal learning, and our method is the first work that demonstrates its usefulness for missing modality generation. Classification RBM is proposed by Larochelle and Bengio, which aims to extend common RBM as a non-linear classifier for supervised learning, especially multi-class learning. The model can also be regarded a special case of Conditional RBM by replacing a set of visible variables with class label variables. But different from it, our method proposes a multi-task encoding for class labels and focuses on multi-label learning on multimodal data.

III. Unconstrained Multimodal Multi-Label Learning

In contrast to the existing multimodal learning which mainly focuses on modality fusion, unconstrained multimodal multi-label learning further considers other two challenges: missing modalities and label dependency. As shown in Fig 1, we take the bimodal data (e.g., image and tag) for illustration, and assume that the tag modality is incomplete. During training, we first extract features for observed modalities (e.g., image), and then generate features for missing modalities (e.g., tag) from the observed ones. After modality generation, all the features are fused to obtain shared representations. It should be noted that the fusion procedure is under the supervision of class labels where the label co-occurrence is particularly considered. During testing, given only the observed modalities (e.g., image), the learned model will generate the missing modalities, and then obtain shared representations which can be used for the subsequent classification and retrieval tasks. In the next, we will give a detailed introduction on using Multi-Label Conditional Restricted Boltzmann Machine (ML-CRBM) for unconstrained multimodal multi-label learning. Before that, we will briefly review RBM, which is the foundation of ML-CRBM.

A. Multimodal Restricted Boltzmann Machine

Due to the power of representation learning [2], [1], RBM has been successfully applied to various tasks [10], [42], [9], [20], [8], especially multimodal learning [24], [36], [37]. Fig. 2(a) illustrates a multimodal RBM, where visible variables $\mathbf{m} \in \{0,1\}^M$ and $\mathbf{t} \in \{0,1\}^N$ denote images and associated tags, respectively, and hidden variables $\mathbf{h} \in \{0,1\}^H$ denote fused representations. Each visible variable is connected to each hidden variable, and no internal connection exists within layers. RBM has been widely used to model the distribution over binary-valued data, while Gaussian RBM [46] and Replicated Softmax Model [9] can be used to handle integer-valued and real-valued inputs, respectively. The energy function of RBM is defined as follows:

$$E(\mathbf{m},\mathbf{t},\mathbf{h}) = -\mathbf{m}^T\mathbf{W}^{\mathbf{mh}}\mathbf{h} - \mathbf{t}^T\mathbf{W}^{\mathbf{th}}\mathbf{h} - \mathbf{a}^T\mathbf{m} - \mathbf{b}^T\mathbf{t} - \mathbf{d}^T\mathbf{h} \quad (1)$$

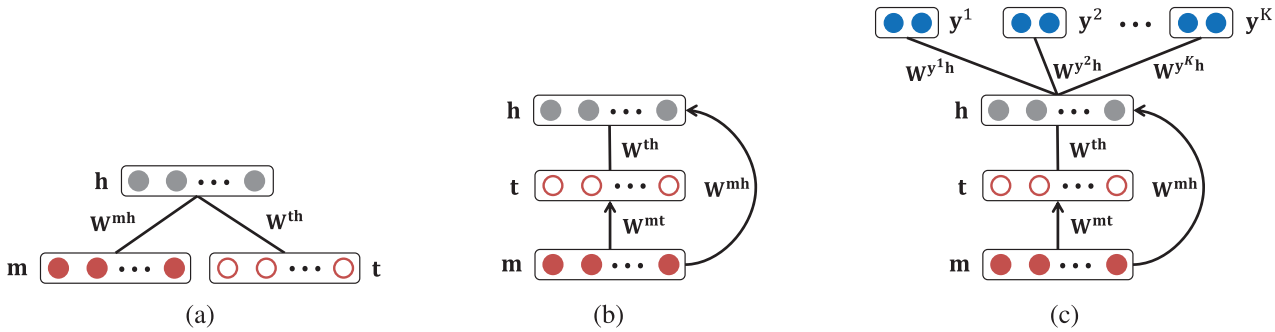


Fig. 2. Multimodal RBM, multimodal conditional RBM, and multi-label conditional restricted Boltzmann machine (ML-CRBM). Note that the ML-CRBM is actually a hybrid graph containing both undirected and directed connections. For modality generation, we regard the visible variables $\mathbf{m}$ as an additional fixed input and model dependency relationships across modalities by the directed connections. (a) Multimodal RBM. (b) Multimodal conditional RBM. (c) ML-CRBM.

where $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{d}$ are biases corresponding to $\mathbf{m}$, $\mathbf{t}$ and $\mathbf{h}$, respectively, and $\mathbf{W^{mh}}$ and $\mathbf{W^{th}}$ are the network weights. Based on the energy function, the distribution $p(\mathbf{m},\mathbf{t})$ can be obtained by marginalizing the joint distribution over all the variables

$$p(\mathbf{m},\mathbf{t}) = \sum\nolimits_{\mathbf{h}} p(\mathbf{m},\mathbf{t},\mathbf{h}) = \frac{1}{Z}\sum\nolimits_{\mathbf{h}} e^{-E(\mathbf{m},\mathbf{t},\mathbf{h})} \tag{2}$$

where $Z = \sum_{\mathbf{m},\mathbf{t},\mathbf{h}} e^{-E(\mathbf{m},\mathbf{t},\mathbf{h})}$ is the partition function.

To learn the network weights, for each pair of image and tag $\{\mathbf{m}^d,\mathbf{t}^d\}_{d\in D}$, we can minimize the negative log-likelihood via gradient descent

$$\mathcal{L} = -\log p(\mathbf{m}^d,\mathbf{t}^d). \tag{3}$$

The gradient of the objective function with respect to $\theta \in \{\mathbf{W^{mh}},\mathbf{W^{th}}\}$ can be computed as follows:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \mathbb{E}_{\mathbf{h}|\mathbf{m}^d,\mathbf{t}^d}\left[\frac{\partial}{\partial\theta}E(\mathbf{m}^d,\mathbf{t}^d,\mathbf{h})\right] - \mathbb{E}_{\mathbf{m},\mathbf{t},\mathbf{h}}\left[\frac{\partial}{\partial\theta}E(\mathbf{m},\mathbf{t},\mathbf{h})\right] \tag{4}$$

where $\mathbb{E}_{\mathbf{h}|\mathbf{m}^d,\mathbf{t}^d}[\frac{\partial}{\partial\theta}E(\mathbf{m}^d,\mathbf{t}^d,\mathbf{h})]$ and $\mathbb{E}_{\mathbf{m},\mathbf{t},\mathbf{h}}[\frac{\partial}{\partial\theta}E(\mathbf{m},\mathbf{t},\mathbf{h})]$ denote two expectations with respect to the data distribution and the model distribution, respectively. Since the model expectation is intractable, we use Contrastive Divergence [7] for an efficient approximation.

B. Multimodal Conditional Restricted Boltzmann Machine

Note that the underlying assumption of multimodal RBM is that all the modalities need to be available, i.e., without missing or incomplete modalities. Such limited assumption is not very practical in real applications since the multimodal data is usually incomplete. Even though we can alternatively use zero-valued vectors to replace the missing modalities for fusion, this will significantly degenerate the shared representations and thus drop the classification performance [37].

Although different modalities exhibit various modality-specific properties, they intrinsically represent the same concept. An intuitive idea is that we can generate the missing modalities from observed ones by taking advantage of this homogeneous property. As shown in Fig 2(b), we apply such idea to a Conditional RBM, where we assume that the tag modality ($\mathbf{t}$) is missing while the image modality ($\mathbf{m}$) is observed. Pairwise layers of variables, $\mathbf{m}$-$\mathbf{t}$, $\mathbf{m}$-$\mathbf{h}$ and $\mathbf{t}$-$\mathbf{h}$ are fully-connected by weights $\mathbf{W^{mt}}$, $\mathbf{W^{mh}}$ and $\mathbf{W^{th}}$, respectively. Note that the directed connections aim to model the dependency relationships across modalities. In particular, we regard the modality $\mathbf{m}$ as an additional fixed input or a dynamically changing bias for the tag-specific RBM (consisting of layers $\mathbf{t}$ and $\mathbf{h}$). Learning proceeds by optimizing the negative conditional log-likelihood via gradient descent

$$\mathcal{L} = -\log p(\mathbf{t}^d|\mathbf{m}^d) \tag{5}$$

for each training data $\{\mathbf{m}^d,\mathbf{t}^d\}_{d\in D}$. When the number of modalities is more than two, we can use a more general form

$$\mathcal{L} = -\log p(\mathbf{T}^d|\mathbf{M}^d) \tag{6}$$

where $\mathbf{T}^d = \{(\mathbf{t}^d)_1,\cdots\}$ and $\mathbf{M}^d = \{(\mathbf{m}^d)_1,\cdots\}$ denote two sets of missing and observed modalities, respectively.

After modality generation, we are able to fuse the complete multimodal data in a similar way as RBM. In fact, the data fusion can be carried out simultaneously with modality generation by optimizing the following objective function:

$$\mathcal{L}_{cond} = -\log p(\mathbf{m}^d,\mathbf{t}^d) - \lambda \log p(\mathbf{t}^d|\mathbf{m}^d) \tag{7}$$

where λ is a tuning parameter for balancing two goals. Note that the two goals are not mutually exclusive, since they both seek modality-free representations in the hidden variables $\mathbf{h}$. In detail, for fusion, with the aim to well explain the multimodal inputs, $\mathbf{h}$ is forced to ignore modality-specific properties and encode modality-free ones. While for generation, $\mathbf{h}$ serves as a transitional state between observed and missing modalities, which mainly captures modality-free characteristics.

C. Multi-Label Conditional Restricted Boltzmann Machine

The Conditional RBM described in Section III-B is an unsupervised learning model, i.e., without using any class label. When dealing with discriminative tasks such as classification, it has to exploit a two-phase learning procedure [24], [36], [37]: 1) fusing multiple modalities to obtain shared representations, and 2) feeding the representations to a prediction classifier. However, such two-phase learning scheme could make the fused representations sub-optimal for label prediction, since the model mainly seeks to improve the generative power rather than discriminative capability in the first phase.

To directly uncover the discriminative properties in the fused representations, we propose a Multi-Label Conditional Restricted Boltzmann Machine (ML-CRBM) which incorporates the supervised information into fusion, and performs modality fusion and label prediction in a one-phase learning procedure. In detail, it fuses multiple modalities under the supervision of class labels, and models the label co-occurrence by a multi-task encoding.

As shown in Fig 2(c), the proposed model consists of $K+3$ layers: an image layer $\mathbf{m}$, a tag layer $\mathbf{t}$, a shared hidden layer $\mathbf{h}$ and K output layers $\mathbf{y}=(\mathbf{y}^1,\mathbf{y}^2,\cdots,\mathbf{y}^K)$, where K is the number of class labels $\{c_k\}_{k=1}^K$. To consider the label dependency, instead of performing each single-label assignment separately, the model regards the multi-label assignment as multiple binary classification tasks, and jointly handles them in a multi-task learning framework. As a result, it can simultaneously use multiple co-occurred labels for supervision. In particular, each binary classification task is associated with an output layer. For the kth task, the two variables (y_1^k, y_2^k) in the corresponding output layer $\mathbf{y}^k$ indicate whether the input multimodal data $\{\mathbf{m},\mathbf{t}\}$ belong to the label c_k or not

$$\begin{cases} \text{if } y_1^k=1,\ y_2^k=0, & \mathbf{m},\mathbf{t}\in c_k \\ \text{if } y_1^k=0,\ y_2^k=1, & \mathbf{m},\mathbf{t}\notin c_k \end{cases}$$

where the layer $\mathbf{y}^k$ uses one-hot encoding [19]. Pairwise adjacent layers $\mathbf{m}$-$\mathbf{h}$, $\mathbf{t}$-$\mathbf{h}$, $\mathbf{y}$-$\mathbf{h}$ and $\mathbf{m}$-$\mathbf{t}$ are fully-connected with network parameters $\mathbf{W^{mh}}$, $\mathbf{W^{th}}$, $\mathbf{W^{yh}}$ and $\mathbf{W^{mt}}$, respectively.

Given the above notations, we define the energy function of ML-CRBM as

$$\begin{aligned} E(\mathbf{m},\mathbf{t},\mathbf{h},\mathbf{y}) = &-\mathbf{m}^T\mathbf{W^{mh}}\mathbf{h}-\mathbf{t}^T\mathbf{W^{th}}\mathbf{h}-\mathbf{y}^T\mathbf{W^{yh}}\mathbf{h} \\ &-\mathbf{m}^T\mathbf{W^{mt}}\mathbf{t} \end{aligned} \tag{8}$$

where all the bias terms are omitted for simplicity. Furthermore, we can obtain the joint distribution

$$p(\mathbf{m},\mathbf{t},\mathbf{y})=\sum\nolimits_{\mathbf{h}} p(\mathbf{m},\mathbf{t},\mathbf{h},\mathbf{y})=\sum\nolimits_{\mathbf{h}}\frac{1}{Z}e^{-E(\mathbf{m},\mathbf{t},\mathbf{h},\mathbf{y})} \tag{9}$$

where $Z=\sum_{\mathbf{m},\mathbf{t},\mathbf{h},\mathbf{y}} e^{-E(\mathbf{m},\mathbf{t},\mathbf{h},\mathbf{y})}$ is the partition function.

D. Inference and Learning

Inheriting the conditionally independent property from RBM, $p(\mathbf{t}|\mathbf{h},\mathbf{m})$ and $p(\mathbf{h}|\mathbf{m},\mathbf{t},\mathbf{y})$ factor over the variables, and their inferences are

$$p(\mathbf{t}|\mathbf{h},\mathbf{m})=\prod\nolimits_j \sigma\left(\sum\nolimits_i x_i W_{ij}^{\mathbf{mt}}+\sum\nolimits_r h_r W_{jr}^{\mathbf{th}}\right) \tag{10}$$

$$\begin{aligned} p(\mathbf{h}|\mathbf{m},\mathbf{t},\mathbf{y}) = \prod\nolimits_r \sigma\Big(&\sum\nolimits_i x_i W_{ir}^{\mathbf{mh}} \\ &+\sum\nolimits_j t_j W_{jr}^{\mathbf{th}}+\sum\nolimits_{k,z} y_z^k W_{z,r}^{\mathbf{yh}_k}\Big). \end{aligned} \tag{11}$$

Note that each pairwise variables (y_1^k, y_2^k) in a class label layer $\mathbf{y}^k$ denote the c_k-label assignment, by regarding each pairwise variables as a group, $p(\mathbf{y}|\mathbf{h})$ factors over K groups. For each group, we can perform inference as

$$p(y_1^k|\mathbf{h})=\frac{1}{Z^{\mathbf{y_k}}}e^{\sum_r h_r W_{1,r}^{\mathbf{yh}_k}} \tag{12}$$

$$p(y_2^k|\mathbf{h})=\frac{1}{Z^{\mathbf{y_k}}}e^{\sum_r h_r W_{2,r}^{\mathbf{yh}_k}} \tag{13}$$

where $Z^{\mathbf{y_k}}=\sum_z e^{\sum_r h_r W_{z,r}^{\mathbf{yh}_k}}$.

To learn the model parameters $\Theta=\{\mathbf{W^{mh}},\mathbf{W^{th}},\mathbf{W^{yh}},\mathbf{W^{mt}}\}$, for each training instance $\{\mathbf{m}^d,\mathbf{t}^d,\mathbf{y}^d\}_{d\in D}$, we jointly optimize the two goals in terms of label prediction and modality generation. For the label prediction, a general way is to optimize the negative log-likelihood (NLL) term $-\log p(\mathbf{m}^d,\mathbf{t}^d,\mathbf{y}^d)$. It can be decomposed as follows:

$$-\log p(\mathbf{m}^d,\mathbf{t}^d,\mathbf{y}^d)=-\log p(\mathbf{y}^d|\mathbf{m}^d,\mathbf{t}^d)-\log p(\mathbf{m}^d,\mathbf{t}^d)$$

which indicates that the NLL term pays partial attention to model the marginal distribution $p(\mathbf{m}^d,\mathbf{t}^d)$ to explain the multimodal inputs. But in supervised learning settings, we only care about learning discriminative shared representations and achieving accurate multi-label predictions. The conditional term $-\log p(\mathbf{y}^d|\mathbf{m}^d,\mathbf{t}^d)$ is more discriminative than the NLL term, so we replace the NLL term with $-\log p(\mathbf{y}^d|\mathbf{m}^d,\mathbf{t}^d)$. For the modality generation, we use the corresponding objective function $-\log p(\mathbf{t}^d|\mathbf{m}^d)$ in (5).

Combining the above two goals, we formulate our learning objective as follows:

$$\mathcal{L}_{dis}=-\log p(\mathbf{y}^d|\mathbf{m}^d,\mathbf{t}^d)-\lambda\log p(\mathbf{t}^d|\mathbf{m}^d) \tag{14}$$

where the two terms are associated to label predication and modality completion, respectively, and λ is a tuning parameter for balance. In our experiments, we achieve the best performance when $\lambda=0.12$.

Stochastic gradient descent is adopted to optimize the objective function $\mathcal{L}_{dis}$, and the gradient of $\mathcal{L}_{dis}$ with respect to $\theta\in\Theta$ is

$$\begin{aligned} \frac{\partial\mathcal{L}_{dis}}{\partial\theta} = &\mathbb{E}_{\mathbf{h}|\mathbf{m}^d,\mathbf{t}^d,\mathbf{y}^d}\left[\frac{\partial}{\partial\theta}E(\mathbf{m}^d,\mathbf{t}^d,\mathbf{y}^d,\mathbf{h})\right] \\ &-\mathbb{E}_{\mathbf{y},\mathbf{h}|\mathbf{m},\mathbf{t}}\left[\frac{\partial}{\partial\theta}E(\mathbf{m},\mathbf{t},\mathbf{y},\mathbf{h})\right] \\ &+\lambda\mathbb{E}_{\mathbf{h}|\mathbf{m}^d,\mathbf{t}^d}\left[\frac{\partial}{\partial\theta}E(\mathbf{m}^d,\mathbf{t}^d,\mathbf{h})\right] \\ &-\lambda\mathbb{E}_{\mathbf{t},\mathbf{h}|\mathbf{m}}\left[\frac{\partial}{\partial\theta}E(\mathbf{m},\mathbf{t},\mathbf{h})\right] \end{aligned} \tag{15}$$

where the first and third terms are data-dependent expectations, the second and fourth terms are model-dependent expectations. Considering that exact minimum objective learning is intractable, we perform efficient approximated learning where a MCMC based procedure [7] is utilized to estimate the model-dependent expectations. As summarized in Algorithm 1, the model can be learned by alternatively performing 1-step Gibbs sampling from the tractable inferences, e.g., $p(\mathbf{t}|\mathbf{h},\mathbf{m})$

and $p(\mathbf{h}|\mathbf{m},\mathbf{t},\mathbf{y})$, and updating the model parameters to minimize the objective.

Algorithm 1 Learning for the ML-CRBM

Input: training data $\{\mathbf{m}^d,\mathbf{t}^d,\mathbf{y}^d\}$, learning rate α
Parameters: $\Theta=\{\mathbf{W^{mh}},\mathbf{W^{th}},\mathbf{W^{yh}},\mathbf{W^{mt}}\}$
Notation: $a\leftarrow b$: setting a as value b
$a\sim p$: sampling a from p
// M update iterations
for m = 1 to M
// Positive phase
$\mathbf{m}^+\leftarrow\mathbf{m}^d,\mathbf{t}^+\leftarrow\mathbf{t}^d,\mathbf{y}^+\leftarrow\mathbf{y}^d$,
$\mathbf{h}^+\leftarrow p(\mathbf{h}|\mathbf{m}^+,\mathbf{t}^+,\mathbf{y}^+)$
$\mathbf{h}_*^+\leftarrow p(\mathbf{h}|\mathbf{m}^+,\mathbf{t}^+)$
// Negative phase
$\hat{\mathbf{h}}^+\sim p(\mathbf{h}|\mathbf{m}^+,\mathbf{t}^+,\mathbf{y}^+),\mathbf{y}^-\sim p(\mathbf{y}|\hat{\mathbf{h}}^+)$,
$\mathbf{h}^-\leftarrow p(\mathbf{h}|\mathbf{m}^+,\mathbf{t}^+,\mathbf{y}^-),\hat{\mathbf{h}}_*^+\sim p(\mathbf{h}|\mathbf{m}^+,\mathbf{t}^+)$,
$\mathbf{t}_*^-\sim p(\mathbf{t}|\hat{\mathbf{h}}_*^+,\mathbf{m}^+),\mathbf{h}_*^-\leftarrow p(\mathbf{h}|\mathbf{m}^+,\mathbf{t}_*^-)$
// Update
for $\theta\in\Theta$ do
$\Delta\theta\leftarrow[\frac{\partial}{\partial\theta}E(\mathbf{m}^+,\mathbf{t}^+,\mathbf{h}^+,\mathbf{y}^+)$
$-\frac{\partial}{\partial\theta}E(\mathbf{m}^+,\mathbf{t}^+,\mathbf{h}^-,\mathbf{y}^-)]$
$+\lambda[\frac{\partial}{\partial\theta}E(\mathbf{m}^+,\mathbf{t}^+,\mathbf{h}_*^+)-\frac{\partial}{\partial\theta}E(\mathbf{m}^+,\mathbf{t}_*^-,\mathbf{h}_*^-)]$
$\theta\leftarrow\theta-\alpha\Delta\theta$
end for
end for

Max-Margin Learning: The objective in (14) utilizes the negative log-likelihood estimation for supervised learning, where the normalization factor of the conditional distribution could make the inference very difficult. Furthermore, for the c_k-label assignment, from a discriminative aspect, the optimal predictions are

$$\begin{cases}\log p(y_1^k|\mathbf{m},\mathbf{t})\gg\log p(y_2^k|\mathbf{m},\mathbf{t}), & \mathbf{m},\mathbf{t}\in c_k\\ \log p(y_1^k|\mathbf{m},\mathbf{t})\ll\log p(y_2^k|\mathbf{m},\mathbf{t}), & \mathbf{m},\mathbf{t}\notin c_k\end{cases}$$

where $p(y_1^k|\mathbf{m},\mathbf{t})$ and $p(y_2^k|\mathbf{m},\mathbf{t})$ represent the probabilities of whether the inputs $\mathbf{m}$ and $\mathbf{t}$ belong to the class c_k or not.

To eliminate the normalization factor and guarantee the separability of the positive and negative cases, it is better to enlarge the distance S_k between $\log p(y_1^k|\mathbf{m},\mathbf{t})$ and $\log p(y_2^k|\mathbf{m},\mathbf{t})$. By taking advantage of the max-margin principle, we obtain a more discriminative objective as follows:

$$\begin{aligned}\mathcal{L}_{mar}&=\sum\nolimits_k\max(\Delta-S_k,0)-\lambda\log p(\mathbf{t}^d|\mathbf{m}^d)\\ S_k&=\mathrm{sgn}[(y^d)_1^k]\left[\log p((y^d)_1^k|\mathbf{m}^d,\mathbf{t}^d)\right.\\ &\quad\left.-\log p((y^d)_2^k|\mathbf{m}^d,\mathbf{t}^d)\right]\end{aligned}\tag{16}$$

where $\mathrm{sgn}(x)$ is 1 for $x=1$ and -1 for $x=0$, and Δ is the margin. Note that in the context of deep learning, hinge loss is also exploited by convolutional neural networks for multi-class classification [41], and by RBM for object segmentation [50].

E. Making Predictions

During testing, to generate missing modalities and predict labels, we are particularly interested in estimating $p(\mathbf{t}|\mathbf{m})$ and $p(\mathbf{y}|\mathbf{m},\mathbf{t})$. However, since both $\mathbf{t}$ and $\mathbf{y}$ have exponential numbers of possible configurations, exact inferences are intractable.

We exploit two factorial distributions $q_f(\mathbf{t},\mathbf{h})$ and $q_s(\mathbf{y},\mathbf{h})$ to approximate the true distributions $p(\mathbf{t}|\mathbf{m})$ and $p(\mathbf{y}|\mathbf{m},\mathbf{t})$, respectively

$$\begin{aligned}q_f(\mathbf{t},\mathbf{h})&=\prod\nolimits_j{\tau_j}^{t_j}(1-\tau_j)^{1-t_j}\prod\nolimits_r{\eta_r}^{h_r}(1-\eta_r)^{1-h_r}\\ q_s(\mathbf{y},\mathbf{h})&=\prod\nolimits_k(\omega_1^k)^{y_1^k}(\omega_2^k)^{y_2^k}\prod\nolimits_r{\pi_r}^{h_r}(1-\pi_r)^{1-h_r}\end{aligned}\tag{17}$$

where τ_j, η_r, ω_1^k, ω_2^k and π_r are variational parameters which estimate $p(t_j=1|\mathbf{m})$, $p(h_r=1|\mathbf{m})$, $p(y_1^k=1|\mathbf{m},\mathbf{t})$, $p(y_2^k=1|\mathbf{m},\mathbf{t})$ and $p(h_r=1|\mathbf{m},\mathbf{t})$, respectively. Note that $q_s(\mathbf{y},\mathbf{h})$ is a structured factorial distribution. In particular, since the two variables $\{y_1^k,y_2^k\}$ in the $\mathbf{y}^k$ denote the k class label, we treat their estimations $\{\omega_1^k,\omega_2^k\}$ as a single group, and factorize $q_s(\mathbf{y},\mathbf{h})$ into the product of a series of group distributions.

Minimizing the two Kullback-Leibler (KL) divergences $KL(q_f(\mathbf{t},\mathbf{h})||p(\mathbf{t},\mathbf{h}|\mathbf{m}))$ and $KL(q_s(\mathbf{y},\mathbf{h})||p(\mathbf{y},\mathbf{h}|\mathbf{m},\mathbf{t}))$, we obtain the following mean-field fixed-point equations:

$$\begin{aligned}\tau_j&\leftarrow\sigma\left(\sum\nolimits_i m_iW_{ij}^{mt}+\sum\nolimits_r\eta_rW_{jr}^{th}\right)\\ \eta_r&\leftarrow\sigma\left(\sum\nolimits_i m_iW_{ir}^{mh}+\sum\nolimits_j\tau_jW_{jr}^{th}\right)\\ \omega_1^k&\leftarrow\sigma\left(\sum\nolimits_r\eta_rW_{1,r}^{y^kh}\right),\omega_2^k\leftarrow\sigma\left(\sum\nolimits_r\eta_rW_{2,r}^{y^kh}\right)\end{aligned}\tag{18}$$

$$\begin{aligned}\pi_r\leftarrow\sigma\Big(&\sum\nolimits_i m_iW_{ir}^{mh}+\sum\nolimits_j t_jW_{jr}^{th}\\ &+\sum\nolimits_k\sum\nolimits_z\omega_z^kW_{z,r}^{y^kh}\Big)\end{aligned}\tag{19}$$

where $\sigma(\cdot)$ is the sigmoid function. Note that we use two different message-passing procedures corresponding to (18) and (19), respectively. Both approximated inferences are performed by iterating the fixed-point equations until convergence. In our experiments, we observe that 20 iterations are sufficient. In fact, using a joint message-passing procedure for these two approximations is feasible but very difficult, since we need to first generate the missing modality, and then use it to perform label prediction in two successive phases.

IV. Experiments

To verify the effectiveness of the proposed method, we take bimodal data as a case study, and perform classification and retrieval experiments on two publicly available datasets.

A. Datasets

MIR Flickr Dataset [14]: The dataset contains 1 million images retrieved from the photography website Flickr. 25,000 images have their associated tags and classes, each of which may belong to multiple classes, 38 classes in total. The classes include object and scene concepts such as *bird*, *flower*, *lake* and

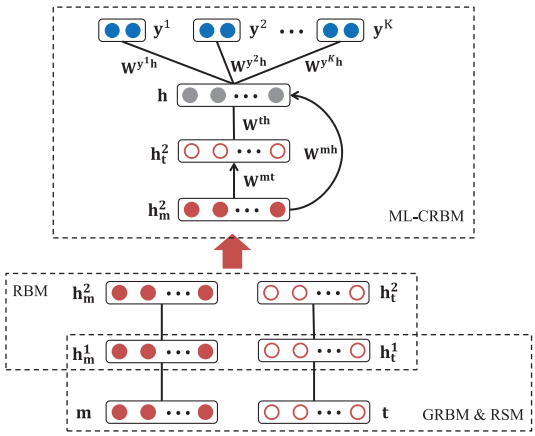


Fig. 3. Deep multi-label conditional restricted Boltzmann machine.

night. Similar to the settings in [37], we use 3,857-dimensional features[1] to describe images, which consist of Pyramid Histogram of Words (PHOW) [3], Gist [26] and MPEG-7 descriptors [21]. The PHOW features are obtained by first extracting dense SIFT features over multi-scale images and then clustering them. We use 2,000-dimensional word count vectors to describe the associated tags. Among the 25,000 images, 10,000/5,000/10,000 are used as the training/validation/testing set. We randomly split the training, validation and testing sets for 5 times and compute their average performance as the final result.

NUS-WIDE Dataset [5]: The dataset contains 269,648 web images and the associated tags, each of which could belong to one or more cases of 81 classes. In our experiments, we use a lite version as NUS-WIDE-LITE which consists of 55,615 images. Images are described by 634-dimensional features[2] which consist of color histogram (LAB) [32], color auto-correlogram (HSV) [12], edge direction histogram [27], wavelet texture [22] and block-wise color moments (LAB) [39]. Tags are represented by 1,001-dimensional word count vectors. We follow the public protocol in [5], which uses 27,807 images for training and 27,808 images for testing.

B. Experimental Settings

Different modalities have various modality-specific properties, which will have a great impact on the quality of data fusion. To reduce such impact, we combine Gaussian RBM (GRBM) [46], Replicated Softmax Model (RSM) [9], standard RBM and the proposed ML-CRBM together as a deep model, as shown in Fig 3. The model first learns less modality-specific representations for image and tag, and then performs modality-free fusion based on the learned representations. The deep model can be learned in a similar way as DBN [8], by separately training all the RBM variants from the bottom up. The architecture of the deep model (e.g., the number of layers, and the number of variables in each layer) on the MIR Flickr and the NUS-WIDE datasets is shown in Table I. Note that we simply exploit a similar model architecture as [37] rather than vastly tune it, because such deep models are found to be insensitive to the choice of these hyperparameters [36]. In our experiments, we

[1][Online]. Available: http://www.cs.toronto.edu/nitish/multimodal/index.html

[2][Online]. Available: http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

TABLE I
MODEL ARCHITECTURE ON THE TWO DATASETS

Layer	MIR Flickr	NUS-WIDE
$\mathbf{y}$	76	162
$\mathbf{h}$	2000	1000
$\mathbf{h}_{\mathbf{t}}^{\mathbf{2}}$	1000	800
$\mathbf{h}_{\mathbf{t}}^{\mathbf{1}}$	1000	800
$\mathbf{t}$	2000	1001
$\mathbf{h}_{\mathbf{m}}^{\mathbf{2}}$	1000	800
$\mathbf{h}_{\mathbf{m}}^{\mathbf{1}}$	1000	800
$\mathbf{m}$	3857	634

TABLE II
AVERAGE PRECISIONS FOR MULTIMODAL MULTI-LABEL CLASSIFICATION ON THE MIR FLICKR DATASET

Model	MAP
LDA	0.492
SVM	0.475
DAE	0.600
DBN	0.599
DBM	0.609
MKL	0.623
TagP	0.640
CRBM	0.563
ML-CRBM-dis	0.618
ML-CRBM-mar	0.655
ML-CRBM-drop	**0.661**

tune the learning rate from $1e$-5 to $1e$-1, and empirically find that using a small learning rate ($1e$-4) for $\mathbf{W^{th}}$ and a large learning rate ($1e$-2) for the rest of parameters can lead to the satisfying performance.

We study four variants of ML-CRBM: 1) ML-CRBM-dis and 2) ML-CRBM-mar utilize two different learning objectives corresponding to (14) and (16), respectively. In contrast to ML-CRBM-dis, 3) ML-CRBM-zero does not generate the missing tag inputs but keeps them clamped at zero. Based on ML-CRBM-mar, 4) ML-CRBM-drop employs the popular Dropout [35] to reduce the over-fitting.

C. Multimodal Multi-Label Classification

Although our method is designed for unconstrained multimodal multi-label learning, it can be applied to the general multimodal multi-label classification. We compare the proposed ML-CRBM with several baseline and state-of-the-art methods, including SVM [15], LDA [15], multimodal Deep AutoEncoder (DAE) [24], multimodal Deep Belief Net (DBN) [36], multimodal Deep Boltzmann Machine (DBM) [37], Multiple Kernel Learning (MKL) [6], TagP [44] and Classification RBM (CRBM) [19]. It should be noted that we implement a multi-label conditional version of CRBM, by using a multi-hot encoding for class labels. Due to the multi-label setting, we compute the Average Precision (AP) for each single-label assignment, and average over all the APs to obtain the Mean Average Precision (MAP) as the evaluation criterion. For the experiments on the MIR Flickr dataset, similar to [37], we use the unlabeled 975,000 images to pretrain all the deep models, which aims to provide a better initialization for model parameters.

The results of the compared methods on the MIR Flickr dataset (in Table II) come from already published papers [37], [6], [44], while the results on the NUS-WIDE dataset (in

TABLE III
MEAN AVERAGE PRECISIONS FOR MULTIMODAL MULTI-LABEL CLASSIFICATION ON THE NUS-WIDE DATASET

Model	MAP
SVM	0.517
DAE	0.538
DBN	0.544
DBM	0.568
ML-CRBM-dis	0.602
ML-CRBM-mar	0.613
ML-CRBM-drop	**0.625**

TABLE IV
AVERAGE PRECISIONS FOR UNCONSTRAINED MULTIMODAL MULTI-LABEL CLASSIFICATION ON THE MIR FLICKR DATASET

Model	MAP
Image SVM	0.375
Image DBN	0.413
Image DBM	0.452
DBN	0.492
DBM-zero	0.490
DBM	0.513
CRBM-zero	0.433
ML-CRBM-zero	0.508
ML-CRBM-dis	0.530
ML-CRBM-mar	0.598
ML-CRBM-drop	**0.611**

Table III) are from our own re-implementations. Note that for all the methods, we report their best performance after parameter tunings. Due to the space limitation, we do not present class-wise APs but only MAPs on the NUS-WIDE dataset. From the two tables we can see that, all the ML-CRBM variants consistently perform better than the compared deep models including DAE, DBN and DBM. In particular, ML-CRBM-dis greatly outperforms CRBM by 0.055 on the MIR Flickr dataset, which indicates that the multi-task encoding for class labels is more effective than multi-hot encoding. Compared with ML-CRBM-dis, the max-margin principle aids ML-CRBM-mar to learn more discriminative fused representations, which results in the improvements by 0.037 and 0.011 on the two datasets, respectively. By randomly dropping a subset of variables in the shared hidden layer, ML-CRBM-drop alleviates the over-fitting problem and further improves the accuracies by 0.013 and 0.012 on the two datasets, respectively. Both ML-CRBM-mar (0.655) and ML-CRBM-drop (0.668) perform better than the state-of-the-arts including MKL (0.623), TagP (0.640) and tree priors based DBM (0.651) [38] on the MIR Flickr dataset.

D. Unconstrained Multimodal Multi-Label Classification

We will next perform the experiments of unconstrained multimodal multi-label classification. In our experiments, we assume that the tag modality is missing. For the multimodal methods, such as DBN and DBM, we sample the missing tag features from the conditional distribution with a Gibbs sampler [37]. For the proposed ML-CRBM, we perform the mean-field updates in (18) to get the estimations $\{\tau_j\}$, and then perform sampling to generate tag features.

We compute MAPs of all the methods in Tables IV and V. The prefix "Image" denotes unimodal methods including Image SVM, Image DBN, and Image DBM, which do not use multimodal data but only images for learning. So they

TABLE V
MEAN AVERAGE PRECISIONS FOR UNCONSTRAINED MULTIMODAL MULTI-LABEL CLASSIFICATION ON THE NUS-WIDE DATASET

Model	MAP
Image SVM	0.249
Image DBN	0.410
Image DBM	0.422
DBN	0.450
DBM	0.489
ML-CRBM-zero	0.470
ML-CRBM-dis	0.507
ML-CRBM-mar	0.532
ML-CRBM-drop	**0.545**

perform much worse than other multimodal methods. The suffix "-zero" denotes multimodal methods which do not generate missing tags but keep them clamped at zero. Similar to the previous observations, the max-margin principle and the Dropout strategy can largely improve the performance of ML-CRBM-mar and ML-CRBM-drop, respectively. But without generating the missing tags, ML-CRBM-zero performs much worse than ML-CRBM-dis by 0.022 and 0.037 on the two datasets, respectively, which shows that the generated tag features by ML-CRBM-dis are very useful for classification. By comparing ML-CRBM-zero with CRBM-zero, our multi-task encoding once again surpasses multi-hot encoding by 0.075 on the MIR Flickr dataset. It should be noted that, even though our method performs slightly worse than multimodal Deep Recurrent Neural Networks [34] (0.661 vs. 0.686) for multimodal multi-label classification in Section IV-C, ML-CRBM-drop can achieve better results (0.611 vs. 0.607) under the unconstrained setting due to the particularly designed model architecture for modality generation.

As shown in Fig 4, we draw class-wise improvement curves [37] of some comparison methods over SVM on both two datasets. Note that without particular statements, we abbreviate ML-CRBM-drop as ML-CRBM in the following. From the figures we can observe that several class-wise APs by ML-CRBM are greatly improved. Specifically, on the MIR Flickr dataset, due to the fewer training samples, some classes have very low APs by SVM such as $Baby^*$ (0.088), $River^*$ (0.102) and Sea^* (0.126), as shown in Table IV. By modelling the label co-occurrence with the proposed ML-CRBM, these classes are able to transfer knowledge from their frequently co-occurred classes, which leads to the great improvements by 0.446, 0.431 and 0.409, respectively, as illustrated in Fig. 4(a). Similarly, we can find from Tables II and IV that, by slightly sacrificing some higher APs as a trade-off, our method can significantly promote the APs of some classes which have fewer training samples.

Previous experiments are all performed by assuming that the tag modality is fully missing. Noticing that the tag modality may be partially instead of fully missing in some real-world scenarios, so we study the performance of the compared methods given partially missing tag modality on the MIR Flickr dataset. In particular, we vary the percentage of missing tag modality from 100% to 0%, and separately perform the experiment of unconstrained multi-label classification. All the MAPs by DBM and the proposed ML-CRBM are illustrated in Table VI. As we can see, the results of both two methods can be progressively improved when reducing the percentage of missing tag. ML-CRBM can outperform DBM given various percentages of

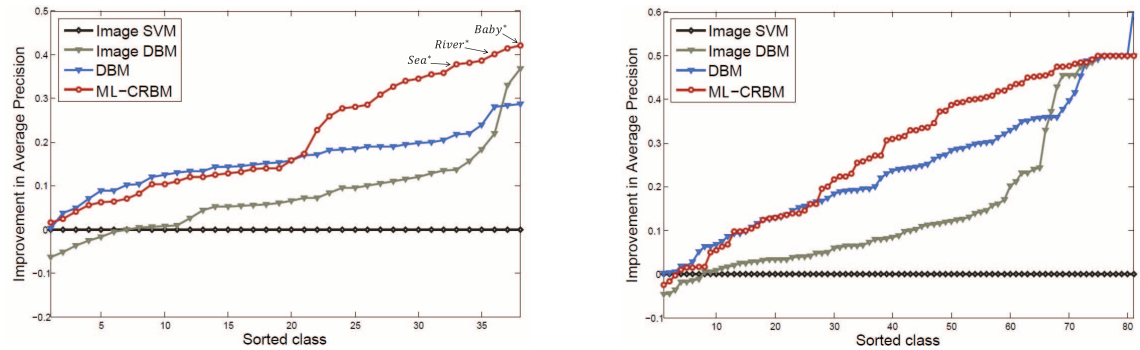


Fig. 4. Class-wise improvement curves for unconstrained multimodal multi-label classification on the MIR Flickr and the NUS-WIDE datasets. (a) MIR Flickr. (b) NUS-WIDE.

TABLE VI
MEAN AVERAGE PRECISIONS FOR UNCONSTRAINED MULTIMODAL MULTI-LABEL CLASSIFICATION ON THE MIR FLICKR DATASET, GIVEN VARYING PERCENTAGE OF MISSING TAG

Percentage of missing tag	MAP	
	DBM	ML-CRBM
100%	0.513	0.611
90%	0.524	0.615
70%	0.545	0.627
50%	0.572	0.642
30%	0.593	0.652
10%	0.605	0.659
0%	0.609	0.661

missing tags, which demonstrates its effectiveness on modality generation.

E. Label Co-occurrence Visualization

To further verify the effectiveness of modelling the label co-occurrence by ML-CRBM, we visualize the incorrect label co-occurrence matrix which measures the prediction error of pairwise labels. The co-occurrence matrix $\mathbf{C} = (C_{mn})_{K\times K}$ has the size of $K \times K$, where K is the number of labels. Each element C_{mn} is the number of incorrect pairwise assignments (labels c_m and c_n), the larger the worse. We compute the incorrect label co-occurrence matrices for the three representative methods: SVM, DBM and ML-CRBM.

The matrix visualizations on the two datasets are illustrated in Figs. 5 and 6, where the redder the color is, the larger the value is. Since each matrix is symmetric, we only present the values in the upper triangular for simplicity. In both Figs, we can observe that the co-occurrence matrices of the ML-CRBM are much more sparse than those of SVM and DBM, which clearly shows that ML-CRBM can make fewer mistakes on the prediction of the label co-occurrence.

In fact, when handling each single-label assignment, SVM and DBM do not consider the relationships of labels, so they make a great number of incorrect predictions of hardly correlated pairwise labels, e.g., *Car-Indoor*. On the other hand, because the prediction error of each single-label assignment is high, after independent combination among these incorrect assigned labels, the models make many false positive assignments such as *Indoor-People*, *Plant-Tree* and *Female-Portrait*. However, by regarding the label co-occurrence as high-order relationship constraints, ML-CRBM considerably suppresses those prediction errors. Note that the matrix visualization only measures the second-order co-occurrence, while for the three-order case, we can obtain the similar observations by visualizing the co-occurrence cubes in Fig 7.

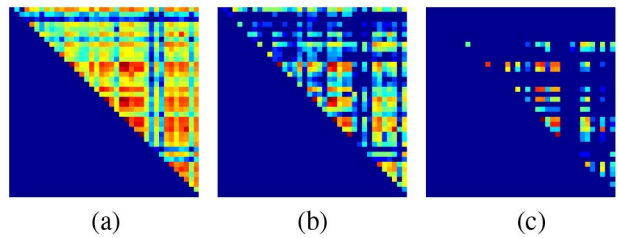


Fig. 5. Visualization of incorrect label co-occurrence matrices on the MIR Flickr dataset. (a) SVM. (b) DBM. (c) ML-CRBM.

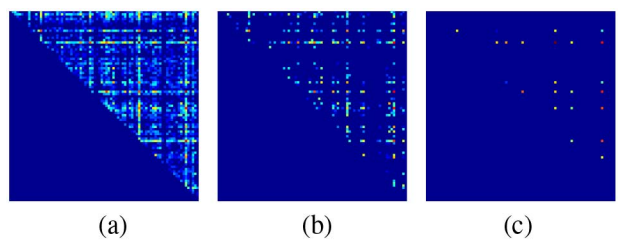


Fig. 6. Visualization of incorrect label co-occurrence matrices on the NUS-WIDE dataset. (a) SVM. (b) DBM. (c) ML-CRBM.

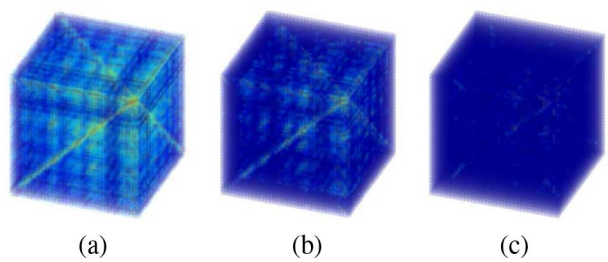


Fig. 7. Visualization of incorrect label co-occurrence cubes on the MIR Flickr dataset. (a) SVM. (b) DBM. (c) ML-CRBM.

F. Multimodal Multi-Label Retrieval

To demonstrate the discriminative power of the fused representations, we perform experiments of multimodal multi-label

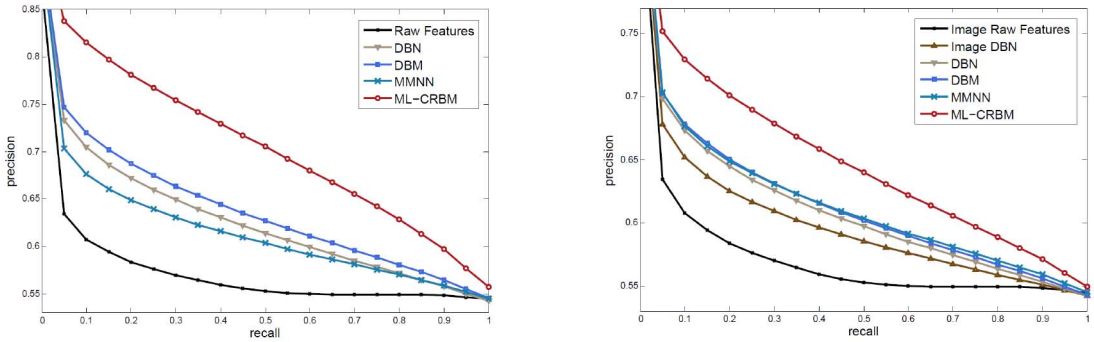


Fig. 8. Precision-recall curves for (unconstrained) multimodal multi-label retrieval on the MIR Flickr dataset. (a) Multimodal multi-label retrieval. (b) Unconstrained multimodal multi-label retrieval.

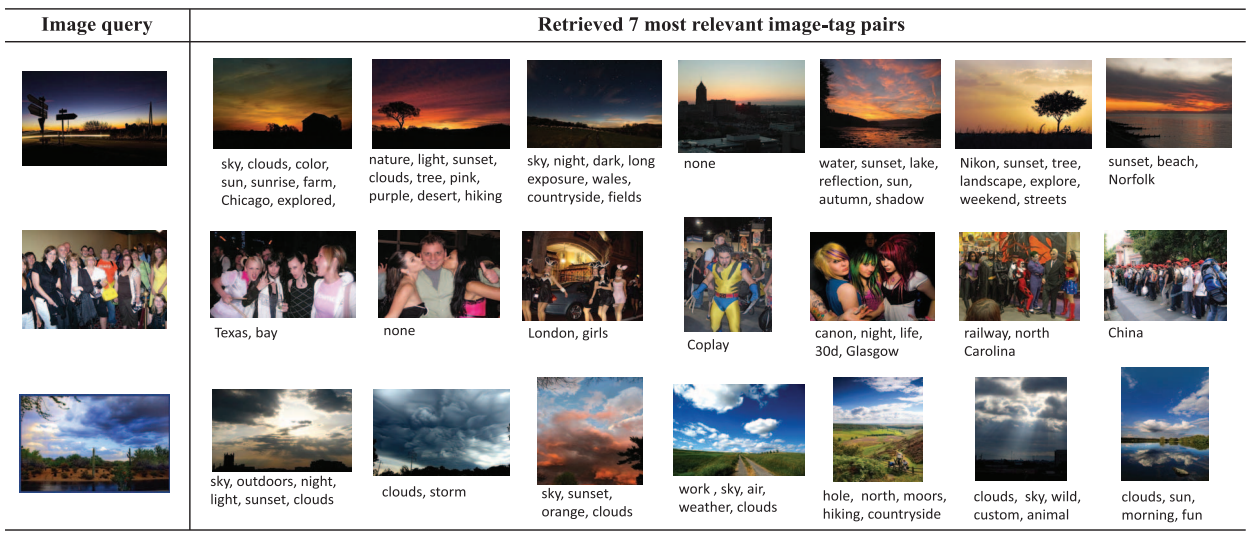


Fig. 9. Results of unconstrained multimodal multi-label retrieval (image query) by ML-CRBM.

retrieval on the MIR Flickr dataset. We randomly select 1000/5000 pairs of image and tag as the query/target set from the testing set. Given a query pair of image and tag, we compare it with each target by computing the cosine distance between the fused representations, and then sort relevant terms by the distances. Since each pair of image and tag could belong to multiple classes, similar to [36], we regard a query and a target as relevant if their class labels are overlapped.

We compare the ML-CRBM with DBN, DBM, a widely recognized multimodal retrieval method namely Multi-Modal Neural Networks (MMNN) [23] and a baseline method (Raw Features) which uses the concatenation of raw image and tag features as fused representations. The precision-recall curves of all the methods are illustrated in Fig. 8(a), from which we can see that, ML-CRBM clearly outperforms other methods, which demonstrates that our learned fused representations are much more discriminative. In particular, when the recall is low, our model is able to achieve much higher precision.

G. Unconstrained Multimodal Multi-Label Retrieval

We also apply our model to the problem of unconstrained multimodal multi-label retrieval. The only difference from the standard multimodal retrieval is that the query set contains only images but without tags. Now the goal is to use image to retrieve pairs of image and tag.

We compare the proposed ML-CRBM with four methods including DBN, DBM, Image DBN and Image Raw Features. Similar to Section IV-D, the unimodal methods (Image DBN and Image Raw Features) utilize only image features while the multimodal methods (DBN, DBM and ML-CRBM) can generate the missing tags before modality fusion.

The precision-recall curves are shown in Fig. 8(b). We can see that all the multimodal methods consistently outperform the unimodal methods. By optimizing an explicit objective for modality generation, ML-CRBM achieves much better performance than the two deep learning methods DBN and DBM. We also present several retrieval examples by ML-CRBM in Fig. 9 (or Fig. 10), where each row presents an image (or a tag) query and its top 7 retrieved results. We can see that, even though given only images (or tags), the model is able to accurately find similar images and tags.

H. Discussion

In the following, we will qualitatively compare the proposed multi-task encoding for class labels with three other encodings, including one-hot [18], one-vs-all [37] and multi-hot [19]. One-hot encoding is shown in Fig. 11(a), where $\mathbf{h}$ and $\mathbf{y}$ are two

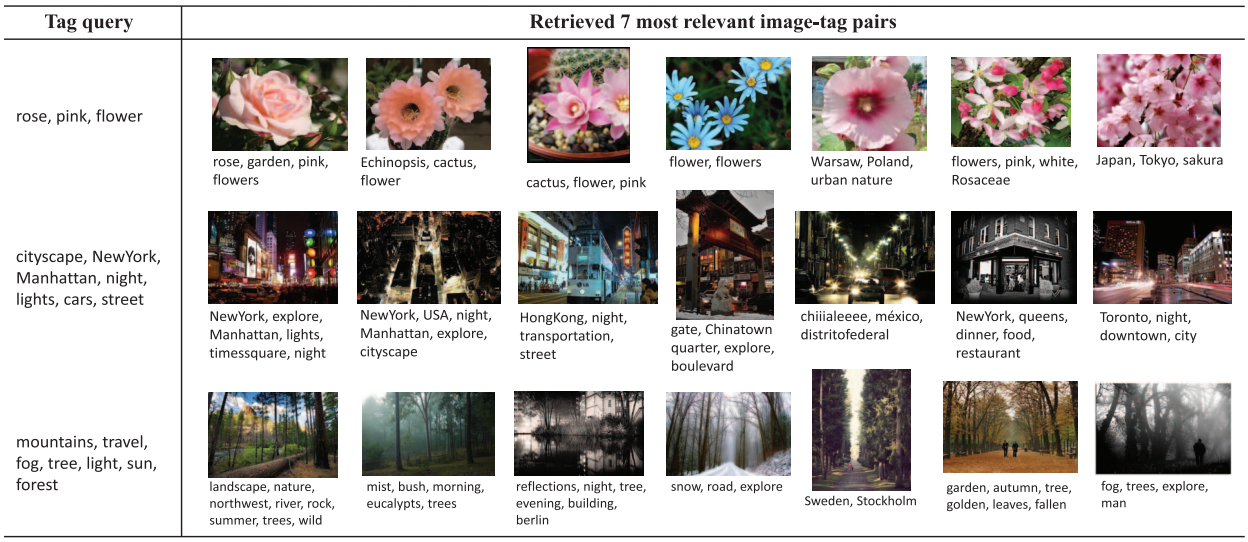


Fig. 10. Results of unconstrained multimodal multi-label retrieval (tag query) by ML-CRBM.

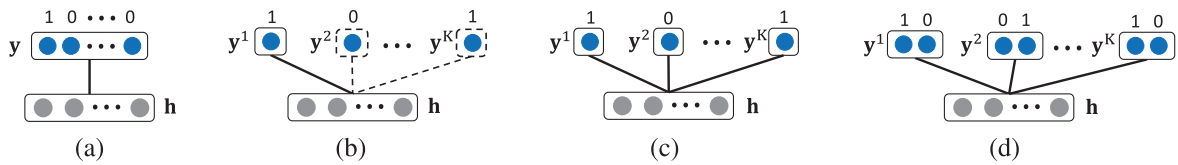


Fig. 11. Comparison of four encoding methods for class labels. (a) One-hot. (b) One-versus-all. (c) Multi-hot. (d) Multi-task.

sets of variables representing fused representations and class labels, respectively. The encoding is often used in multi-class learning, where a single label is assigned by activating the corresponding variable in $\mathbf{y}$ as 1. When performing multi-label learning, i.e., assigning more than one label, a common way is to transform the problem to multiple single-label assignments in a one-vs-all manner [37], i.e., splitting positive and negative samples for each label, and separately performing multiple binary classifications with logistic regressions. As illustrated in Fig. 11(b), each label layer $\mathbf{y}^k$ contains only one logistic variable representing the probability of the positive case. The dotted lines indicate that the multiple binary classifications are independently implemented, so such one-vs-all encoding can not model dependency relationships of labels [49], [38]. To consider the label relationships, it is straightforward to directly extend the one-hot encoding to a multi-hot version as shown in Fig. 11(c), where multiple variables can be simultaneously activated as 1. Multi-hot encoding can also be regarded as a synchronized version of the one-vs-all encoding.

Compared with the multi-hot encoding, our proposed multi-task encoding handles each single-label assignment with a binary softmax regression rather than the logistic regression. As shown in Fig. 11(d), it uses two variables for each single-label assignment corresponding to the positive case and the negative case, respectively. For common binary classification, softmax regression is equivalent to logistic regression, but different from that, our case here contains multiple binary classifications. And more importantly, the regression here is not independently linear but in conjunction with the nonlinear data fusion. By using additional variables to model the negative case, multi-task encoding enables the model to perform negative regulation to suppress the incorrect prediction. Accordingly, multi-task encoding achieves much better performance than multi-hot encoding as shown in Section IV-C and Section IV-D.

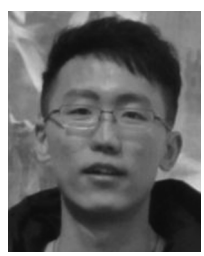
V. Conclusion and Future Work

In this paper, we have proposed the Multi-Label Conditional Restricted Boltzmann Machine to handle the problem of unconstrained multimodal multi-label learning. By jointly modelling the conditional distribution over missing modalities and considering label co-occurrence as a multi-task learning problem, the proposed model can handle modality completion, fusion and multi-label prediction in a unified framework. The experimental results of unconstrained multimodal multi-label classification, retrieval and visualization have demonstrated the effectiveness of our model. In the future, we will validate our model on other multimodal data where more than one modalities are missing.

References

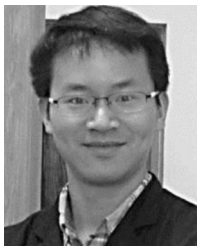
[1] Y. Bengio, *Learning Deep Architectures for AI (Foundations and Trends in Machine Learning)*. Delft, The Netherlands: Now Publishers, 2009, pp. 1–127.
[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[3] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[4] N. Chen, J. Zhu, and E. P. Xing, "Large-margin predictive latent subspace learning for multi-view data analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2365–2378, Dec. 2012.
[5] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world web image database from national university of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, pp. 1–9, Art. ID 48.

[6] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 902–909.
[7] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.
[8] 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.
[9] G. E. Hinton and R. Salakhutdinov, "Replicated softmax: An undirected topic model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1607–1614.
[10] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[11] Y. Hua, S. Wang, S. Liu, Q. Huang, and A. Cai, "TINA: Cross-modal correlation learning by adaptive hierarchical semantic aggregation," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 190–199.
[12] J. Huang, S. Kumar, M. Mitra, W. J. Zhu, and R. Zabih, "Image indexing using color correlogram," in *Proc. Comput. Vis. Pattern Recog.*, 1997, pp. 762–768.
[13] Y. Huang, W. Wang, L. Wang, and T. Tan, "Multi-task deep neural network for multi-label learning," in *Proc. 20th IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 2897–2900.
[14] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 39–43.
[15] M. J. Huiskes, B. Thomee, and M. S. Lew, "New trends and ideas in visual concept detection: The MIR Flickr retrieval evaluation initiative," in *Proc. ACM Int. Conf. Multimedia Inf. Retrieval*, 2010, pp. 527–536.
[16] R. Kiros and C. Szepesvari, "Deep representations and codes for image auto-annotation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 908–916.
[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.
[18] H. Larochelle and Y. Bengio, "Classification using discriminative restricted Boltzmann machines," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 536–543.
[19] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio, "Learning algorithms for the classification restricted Boltzmann machine," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 643–669, 2012.
[20] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. 26th Int. Conf. Mach. Learn.*, 2009, pp. 609–616.
[21] B. Manjunath, J. Ohm, V. Vasudevan, and A. Ya-mada, "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, Jun. 2001.
[22] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, Aug. 1996.
[23] J. Masci, M. M. Bronstein, A. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 824–830, Apr. 2014.
[24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.
[25] C.-T. Nguyen, D.-C. Zhan, and Z.-H. Zhou, "Multi-modal image annotation with multi-instance multi-label LDA," in *Proc. 23rd Int. Joint Conf. Artificial Intell.*, 2013, pp. 1558–1564.
[26] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[27] D. K. Park, Y. S. Jeon, and C. S. Won, "Efficient use of local edge histogram descriptor," in *Proc. MM Workshops*, 2000, pp. 51–54.
[28] G.-J. Qi, C. Aggarwal, and T. Huang, "Towards semantic knowledge propagation from text corpus to web images," in *Proc. WWW*, 2011, pp. 297–306.
[29] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *Proc. 15th Int. Conf. Multimedia*, 2007, pp. 17–26.
[30] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang, "Two-dimensional multilabel active learning with an efficient online adaptation model for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1880–1897, Oct. 2009.
[31] R. Salakhutdinov, A. Mnih, and G. E. Hinton, "Restricted Boltzmann machines for collaborative filtering," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 791–798.
[32] L. G. Shapiro and G. C. Stockman, *Computer Vision*. Englewood Cliffs, NJ, USA: Prentice Hall, 2003.
[33] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA, USA: MIT Press, 1986, vol. 1, pp. 194–281.
[34] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2141–2149.
[35] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *JMLR*, vol. 15, no. 1, pp. 1929–1958, 2014.
[36] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *Proc. Int. Conf. Mach. Learn. Workshop*, 2012, pp. 1–8.
[37] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2231–2239.
[38] N. Srivastava and R. Salakhutdinov, "Discriminative transfer learning with tree-based priors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2094–2102.
[39] M. Stricker and M. Orengo, "Similarity of color images," in *Proc. IS&T/SPIE's Symp. Electron. Imaging: Sci. Technol.*, 1995, pp. 381–392.
[40] I. Sutskever and G. E. Hinton, "Learning multilevel distributed representations for high-dimensional sequences," in *Proc. AISTATS*, 2007, pp. 548–555.
[41] Y. Tang, "Deep learning using linear support vector machines," *CoRR*, vol. abs/1306.0239, Jul. 2013 [Online]. Available: http://arxiv.org/abs/1306.0239
[42] G. Taylor, G. Hinton, and S. Roweis, "Modeling human motion using binary latent variables," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1345–1352.
[43] G. W. Taylor and G. E. Hinton, "Factored conditional restricted Boltzmann machines for modeling motion style," in *Proc. 26th Int. Conf. Mach. Learn.*, 2009, pp. 1025–1032.
[44] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid, "Image annotation with tagprop on the MIRFLICKR set," in *Proc. Int. Conf. Multimedia Inf. Retrieval*, 2010, pp. 537–546.
[45] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, "Effective multi-modal retrieval based on stacked auto-encoders," in *Proc. VLDB Endowment*, Apr. 2014, vol. 7, no. 6, pp. 649–660.
[46] M. Welling, M. Rosen-Zvi, and G. Hinton, "Exponential family harmoniums with an application to information retrieval," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 1481–1488.
[47] P. Xie and E. P. Xing, "Multimodal distance metric learning," in *Proc. Int. Joint Conf. Artificial Intell.*, 2013, pp. 1806–1812.
[48] E. P. Xing, R. Yan, and A. G. Hauptmann, "Mining associated text and images with dual-wing harmoniums," in *Proc. 21st Conf. Uncertainty Artificial Intell.*, 2005, pp. 633–641.
[49] J. Xu, V. Jagadeesh, and B. S. Manjunath, "Multi-label learning with fused multimodal bi-relational graph," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 403–412, Feb. 2014.
[50] J. Yang, S. Safar, and M.-H. Yang, "Max-margin Boltzmann machines for object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 320–327.
[51] M. L. Zhang and Z. H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1338–1351, Oct. 2006.
[52] M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 18, pp. 1819–1837, Aug. 2014.



Yan Huang received the B.Sc. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2012, and is currently working toward the Ph.D. degree at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China.

His research interests include machine learning and pattern recognition.



Wei Wang received the B.E. degree from Wuhan University, Wuhan, China, in 2005, and the Ph.D. degree in information science and engineering from the Graduate University of Chinese Academy of Sciences (GUCAS), Beijing, China, in 2011.

Since July 2011, he has been an Assistant Professor with NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than ten papers that have appeared in leading international conferences such as CVPR and ICCV. His research interests include computer vision, pattern recognition, and machine learning, particularly on the computational modeling of visual attention, deep learning, and multimodal data analysis.

Liang Wang (M'09–SM'09) received the B.Eng. and M.Eng. degrees from Anhui University, Hefei, China, in 1997 and 2000, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2004.

From 2004 to 2010, he was a Research Assistant with Imperial College London, London, U.K., and Monash University, Melbourne, Australia, a Research Fellow with the University of Melbourne, Melbourne, Australia, and a Lecturer with the University of Bath, Bath, U.K. He is currently a Full Professor of the Hundred Talents Program with the National Lab of Pattern Recognition, CASIA. He has authored or coauthored papers that have appeared in highly ranked international journals such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the IEEE TRANSACTIONS ON IMAGE PROCESSING, and leading international conferences such as CVPR, ICCV, and ICDM. His major research interests include machine learning, pattern recognition, and computer vision.

Dr. Wang is an IAPR Fellow.