Multi-Modal Distance Metric Learning*

Pengtao Xie
Department of Computer Science
Tsinghua University
Beijing 10086, China

Eric P. Xing
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Abstract

Multi-modal data is dramatically increasing with the fast growth of social media. Learning a good distance measure for data with multiple modalities is of vital importance for many applications, including retrieval, clustering, classification and recommendation. In this paper, we propose an effective and scalable multi-modal distance metric learning framework. Based on the multi-wing harmonium model, our method provides a principled way to embed data of arbitrary modalities into a single latent space, of which an optimal distance metric can be learned under proper supervision, i.e., by minimizing the distance between similar pairs whereas maximizing the distance between dissimilar pairs. The parameters are learned by jointly optimizing the data likelihood under the latent space model and the loss induced by distance supervision, thereby our method seeks a balance between explaining the data and providing an effective distance metric, which naturally avoids overfitting. We apply our general framework to text/image data and present empirical results on retrieval and classification to demonstrate the effectiveness and scalability.

1 Introduction

Multi-modal data is ubiquitous in web. Especially, with the vast prosperity of social media, data with multiple modalities is enjoying explosive growth. For example, in user-centric social networks (e.g., Facebook, Google Plus), users possess blogs, photos, friends circles. In photo sharing websites (e.g., Flickr, Pinterest), photos can be described by image contents, text tags and meta information like albums and groups. In video sharing website (e.g., Youtube), videos can be characterized by image frames, audio, and user comments. In music social network (e.g., iTunes Ping), songs are accompanied by acoustic features (e.g., rhythm and timbre), semantic features (e.g., tags, lyrics) and social features (e.g., artist reviews) [McFee and Lanckriet, 2011]. Information from different sources (text, image, video, audio, meta and social information) jointly reveal the fundamental characteristics of the study subjects from different views.

Choosing a proper distance function or similarity measure for multi-modal data is crucial for many applications, including retrieval [Zhang *et al.*, 2011; Zhen and Yeung, 2012], clustering [Bekkerman and Jeon, 2007; Qi *et al.*, 2012], classification [Nishida *et al.*, 2012] and recommendation [Aizenberg *et al.*, 2012; Baluja *et al.*, 2008]. While various metric learning methods [Xing *et al.*, 2002; Globerson and Roweis, 2006; Weinberger *et al.*, 2006; Davis *et al.*, 2007] defined on single data modality have been proposed, learning distance in the presence of multiple modalities remains largely unexploited. McFee and Lanckriet [2011] applied the multiple kernel learning technique for integrating heterogeneous feature modalities into a single unified similarity space. An ensemble of kernel transformations are learned given the labeled relative similarity comparisons. To our best knowledge, this is the only work regarding multi-modal distance metric learning. However, their method enjoys very limited scalability.

In this paper, we propose a multi-modal distance metric learning framework based on the multi-wing harmonium (MWH) model [Xing *et al.*, 2005] for multi-model integration and the metric learning method by [Xing *et al.*, 2002] for distance supervision. Our method provides a principled way to embed data of arbitrary modalities into a single latent space where distance supervision is leveraged. This MWH is a two-layer random field that jointly models the visible feature modalities and their latent semantic embeddings. Given labeled "similar" and "dissimilar" pairs, we aim to minimize the distance of similar pairs while separating dissimilar pairs with a certain margin in the latent space. The embedding from feature modalities to semantic space is learned by simultaneously maximizing random field induced data likelihood and minimizing distance induced loss, with the goal to reasonably explain data and provide an effective distance metric. The data likelihood provides a natural regularizer to avoid overfitting commonly suffered by most metric learning algorithms. Unlike existing distance metric learning approaches which requires costly semi-definite programming or kernel embedding, our method is highly efficient and scalable.

The rest of the paper is organized as follows. Section 2 introduces related work. In section 3, we propose the multi-

*The work is done while Pengtao Xie is visiting Carnegie Mellon University.

modal distance metric learning framework based on multi-wing harmonium model and present the optimization technique. Section 4 gives experimental results. Section 5 concludes the paper.

2 Related Work

Metric learning on single modality has been widely studied in [Xing *et al.*, 2002; Globerson and Roweis, 2006; Weinberger *et al.*, 2006; Davis *et al.*, 2007]. Xing *et al* [2002] used semidefinite programming to learn a Mahalanobis distance metric for clustering under similarity and dissimilarity constraints. They aim to minimize the distance of similar pairs while separating dissimilar pairs with a certain margin. Weinberger *et al* [2006] employed a similar semidefinite formulation for k-nearest neighbor classification. The metric is trained with the goal that the k-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin. Globerson and Roweis [2006] proposed a formulation aiming to collapse all examples in the same class to a single point and push examples in other classes infinitely far away. Davis *et al* [2007] proposed information theoretic metric learning which minimizes the differential relative entropy between two multivariate Gaussians under the constraints on the distance function. These algorithms are primarily designed for single feature representations and are not suitable for multi-modal data. One naive way is to concatenate features from different modalities into a single representation and subsequently apply single-modal metric learning techniques to learn a distance measure. This strategy ignores the incompatibility of heterogeneous information sources and fails to consider the dependency and complementarity relationships among different modalities, thereby leading to suboptimal performance.

To our best knowledge, the only work of distance metric learning on multiple modalities is proposed by McFee and Lanckriet [2011]. They integrate heterogeneous data to learn a holistic similarity measure based on multiple kernel learning, where each kernel encodes a different modality of the data. Their method learns a Mahalanobis distance metric for each modality over the reproducing kernel Hilbert space, which can be solved by semidefinite programming. This method is computationally costly both in training and testing. The training phase involves optimizing over multiple high-dimensional positive semi-definite matrices, which is hard to scale to large data set. In the test phase, to embed query into the learned target space, it needs to evaluate the kernel functions at query against the entire training set, which can not meet real-time requirements when the training set is large.

One closely related work is proposed in [Chen *et al.*, 2010], which integrates multi-wing harmonium model [Xing *et al.*, 2005] and large margin learning for the purpose of predictively learning a latent subspace for multi-view data. In their work, the available supervised information is class labels and the major goal of their approach is classification. Our work focuses on distance metric learning and the available supervision is "similar" and "dissimilar" pairs. Li *et al* [2009] proposed a unified framework to estimate similarities by exploiting the interactions between objects of different modality. In their work, similarity measure is learned in an unsupervised manner. Our work exploits the problem of distance metric learning which leverages supervised "similar" and "dissimilar" pairs to learn a similarity measure.

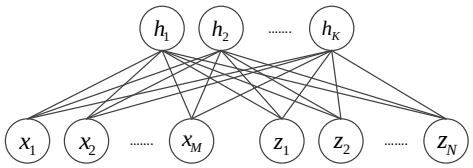

Figure 1: Dual-wing harmonium model

3 Multi-Modal Distance Metric Learning

In this section, we first briefly describe the multi-wing harmonium model (MWH). Based on MWH model, we introduce the multi-modal distance metric learning (MMDML) framework and the optimization technique.

3.1 Multi-Wing Harmonium Model

For simplicity, we start with dual-wing harmonium model, which is a special case of multi-wing harmonium model and can be easily extended to multi-wing case. The dual-wing harmonium model (DWH) [Xing *et al.*, 2005] is shown in Figure 1, which consists of two modalities of input units $\mathbf{x} = \{x_i\}$, $\mathbf{z} = \{z_j\}$ and a set of hidden units $\mathbf{h} = \{h_k\}$. In this undirected graphic model, there exist no connections between two input modalities. Each modality of input units and the hidden units form a complete bipartite graph where units in the same set have no connections and are fully connected to units in the other set. This topology induces three conditional independence assumptions: given latent variables $\mathbf{h}$, the two modalities $\mathbf{x}$ and $\mathbf{z}$ are independent, $p(\mathbf{x}, \mathbf{z}|\mathbf{h}) = p(\mathbf{x}|\mathbf{h})p(\mathbf{z}|\mathbf{h})$; given $\mathbf{x}$ and $\mathbf{z}$, each unit in $\mathbf{h}$ is independent from each other $p(\mathbf{h}|\mathbf{x}, \mathbf{z}) = \prod_k p(h_k|\mathbf{x}, \mathbf{z})$; given $\mathbf{h}$, units within each modality are independent $p(\mathbf{x}|\mathbf{h}) = \prod_i p(x_i|\mathbf{h})$, $p(\mathbf{z}|\mathbf{h}) = \prod_j p(z_j|\mathbf{h})$. Consider the case where all observed and hidden variables are from exponential family, we have

$$\begin{array}{rcl} p(x_i) & = & \exp\{\theta_i^\mathsf{T}\phi(x_i) - A(\theta_i)\} \\ p(z_j) & = & \exp\{\eta_j^\mathsf{T}\varphi(z_j) - B(\eta_j)\} \\ p(h_k) & = & \exp\{\lambda_k^\mathsf{T}\psi(h_k) - C(\lambda_k)\} \end{array} \quad (1)$$

where $\theta_i, \eta_j, \lambda_k$ are natural parameters, $\phi(\cdot), \varphi(\cdot), \psi(\cdot)$ are sufficient statistics and $A(\cdot), B(\cdot), C(\cdot)$ are log partition functions.

We couple the random variables in the log-domain by introducing an additional term and get the joint distribution

$$\begin{array}{rcl} p(\mathbf{x}, \mathbf{z}, \mathbf{h}) & \propto & \exp\{\sum_i \theta_i^\mathsf{T}\phi(x_i) + \sum_j \eta_j^\mathsf{T}\varphi(z_j) \\ & + & \sum_k \lambda_k^\mathsf{T}\psi(h_k) + \sum_{ik} \phi(x_i)^\mathsf{T}\mathbf{W}_i^k\psi(h_k) \\ & + & \sum_{jk} \varphi(z_j)^\mathsf{T}\mathbf{U}_j^k\psi(h_k)\} \end{array} \quad (2)$$

the log-partition function of the joint probability is not explicitly shown to emphasize the difficulty of its estimation. If we examine the joint distribution from the random field

viewpoint, $\phi(x_i), \varphi(z_j), \psi(h_k)$ define potentials over cliques formed by individual nodes, $\phi(x_i)\psi(h_k)$,$\varphi(z_j)\psi(h_k)$ define potentials over cliques consisting of pairwise linked nodes and $\theta_i, \eta_j, \lambda_k, \mathbf{W}_i^k, \mathbf{U}_j^k$ are the associated weights of potential functions. We use Θ to denote all the parameters $(\theta, \eta, \lambda, \mathbf{W}, \mathbf{U})$. Θ is learned by maximum likelihood method. From the joint distribution, we derive the conditional distributions

$$\begin{array}{lcl} p(x_i|\mathbf{h}) & \propto & \exp\{\hat{\theta}_i^{\mathsf{T}}\phi(x_i) - A(\hat{\theta}_i)\} \\ p(z_j|\mathbf{h}) & \propto & \exp\{\hat{\eta}_j^{\mathsf{T}}\varphi(z_j) - B(\hat{\eta}_j)\} \\ p(h_k|\mathbf{x},\mathbf{z}) & \propto & \exp\{\hat{\lambda}_k^{\mathsf{T}}\psi_k(h_k) - C(\hat{\lambda}_k)\} \end{array} \tag{3}$$

with shifted parameters $\hat{\theta}_i = \theta_i + \sum_k \mathbf{W}_i^k\psi(h_k)$,$\hat{\eta}_j = \eta_j + \sum_k \mathbf{U}_j^k\psi(h_k)$, $\hat{\lambda}_k = \lambda_k + \sum_i \mathbf{W}_i^k\phi(x_i) + \sum_j \mathbf{U}_j^k\varphi(z_j)$, where the shifts are induced by the coupling between the observed and hidden units.

Reversely, we can firstly specify the local conditional distributions in Eq.(3) according to specific applications, then write the joint distribution in Eq.(2). This is called bottom-up construction of the dual-wing harmonium model.

The dual-wing harmonium model can be readily extended to multi-wing when multi-modal input feature set are observed.

3.2 Multi-Modal Distance Metric Learning

In this section, we present how to employ MWH model to learn a distance measure on multiple modalities. Given a data point $\mathbf{y} = (\mathbf{x}, \mathbf{z})$ with two feature modalities $\mathbf{x}$ and $\mathbf{z}$, under the dual-wing harmonium model framework, we can embed $\mathbf{y}$ into the shared latent space and obtain its latent representation $\mathbf{t}$:

$$\mathbf{t} = \mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z};\Theta)}[\mathbf{h}] \tag{4}$$

Note that $\mathbf{t}$ is a function of Θ. Semantically, hidden units $\mathbf{h}$ can be viewed as a set of latent topics. Observations of different sources reflect the central theme from different perspectives and are generated from the shared topics.

Given a set of pairs labeled as "similar" or "dissimilar", we enforce similar pairs to be close to each other and dissimilar pairs to be far from each other in the latent space. We simply use Euclidean distance as distance measure for embedded points in this latent space. Let $\mathcal{S} = \{(\mathbf{y}^{(i)}, \mathbf{y}^{(j)})\}$ denote the set of similar pairs and $\mathcal{D} = \{(\mathbf{y}^{(i)}, \mathbf{y}^{(j)})\}$ denote the set of dissimilar pairs, we formulate the following optimization problem:

$$\begin{array}{ll} \min_\Theta & \sum_{(\mathbf{y}^{(i)},\mathbf{y}^{(j)})\in\mathcal{S}} \|\mathbf{t}^{(i)} - \mathbf{t}^{(j)}\|^2 \\ \text{s.t.} & \forall(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}) \in \mathcal{D}, \|\mathbf{t}^{(i)} - \mathbf{t}^{(j)}\|^2 \geq 1 \end{array} \tag{5}$$

where $\mathbf{t}^{(i)}$ is the representation of $\mathbf{y}^{(i)}$ in the latent space. The goal is to minimize the distance between points labeled as "similar" while keeping the "dissimilar" points separated by a margin of 1 in the latent space. The parameter to be learned is Θ which is explicitly embedded in $\mathbf{t}$.

Let $\mathcal{Y}$ denote all data instances appearing in $\mathcal{S}$ or $\mathcal{D}$. The unsupervised dual wing harmonium model learns parameter Θ by maximizing the likelihood of data observations $\mathcal{Y}$. Θ is learned in the sense of best explaining the data. In supervised multi-modal distance learning metric framework, we combine distance metric learning and maximum likelihood learning together and learn Θ by jointly maximizing data likelihood, minimizing distances of similar pairs and maximizing distances of dissimilar pairs. The learned Θ not only explains the data reasonably, but also facilitates good similarity comparison between data points. Specifically, we define the optimization problem as

$$\begin{array}{ll} \min_\Theta & \frac{1}{|\mathcal{Y}|}\mathcal{L}(\mathcal{Y};\Theta) + \lambda\frac{1}{|\mathcal{S}|}\sum_{(\mathbf{y}^{(i)},\mathbf{y}^{(j)})\in\mathcal{S}} \|\mathbf{t}^{(i)} - \mathbf{t}^{(j)}\|^2 \\ \text{s.t.} & \forall(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}) \in \mathcal{D}, \|\mathbf{t}^{(i)} - \mathbf{t}^{(j)}\|^2 \geq 1 \end{array} \tag{6}$$

where $\mathcal{L}(\mathcal{Y};\Theta)$ is the negative log-likelihood of data $\mathcal{Y}$ parameterized by Θ and λ is the trade-off parameter. $|\cdot|$ denotes the cardinality of a set.

3.3 Optimization

In this section, we present an efficient solver of the problem defined in Eq.(6). The strategy is to use hinge loss to eliminate constraints, obtaining an unconstrained problem and subsequently employ subgradient method to do optimization.

We use hinge loss to eliminate the constrains in Eq.(6) and obtain:

$$\begin{array}{ll} \min_\Theta & \frac{1}{|\mathcal{Y}|}\mathcal{L}(\mathcal{Y};\Theta) + \lambda_1\frac{1}{|\mathcal{S}|}\sum_{(\mathbf{y}^{(i)},\mathbf{y}^{(j)})\in\mathcal{S}} \|\mathbf{t}^{(i)} - \mathbf{t}^{(j)}\|^2 + \\ & \lambda_2\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{y}^{(i)},\mathbf{y}^{(j)})\in\mathcal{D}} \max(0, 1 - \|\mathbf{t}^{(i)} - \mathbf{t}^{(j)}\|^2) \end{array} \tag{7}$$

where λ_1 and λ_2 are trade-off parameters. The problem defined in Eq.(7) is a relaxed version of that in Eq.(6). When constraints in Eq.(6) are satisfied, hinge loss in Eq.(7) is zero. Otherwise, hinge loss is nonzero and is minimized to enforce the constraints to be satisfied. Since hinge loss is non-differential, we adopt sub-gradient method to do optimization.

We use contrastive divergence [Hinton, 2002] method to approximate the gradient of Θ w.r.t the negative log-likelihood $\frac{1}{|\mathcal{Y}|}\mathcal{L}(\mathcal{Y};\Theta)$. Gradients (or sub-gradients) w.r.t the distance loss induced by similar pairs and distance loss induced by dissimilar pairs can be easily derived. In summary, the sub-gradients of parameters w.r.t the objective function defined in Eq.(7) can be computed as

$$\begin{array}{lcl} \nabla\theta_i & = & \mathbb{E}_p[\phi(x_i)] - \mathbb{E}_{\hat{p}}[\phi(x_i)] \\ \nabla\eta_j & = & \mathbb{E}_p[\varphi(z_j)] - \mathbb{E}_{\hat{p}}[\varphi(z_j)] \\ \nabla\lambda_k & = & \mathbb{E}_p[\psi(h_k)] - \mathbb{E}_{\hat{p}}[\psi(h_k)] \end{array} \tag{8}$$

$$\begin{aligned} &\nabla W_i^k = \mathbb{E}_p[\phi(x_i)\psi(h_k)^{\mathsf{T}}] - \mathbb{E}_{\hat{p}}[\phi(x_i)\psi(h_k)^{\mathsf{T}}] \\ &+\lambda_1\frac{2}{|\mathcal{S}|}\sum_{(\mathbf{y}^{(m)},\mathbf{y}^{(n)})\in\mathcal{S}}(t_k^{(m)} - t_k^{(n)})(\frac{\partial t_k^{(m)}}{\partial W_i^k} - \frac{\partial t_k^{(n)}}{\partial W_i^k}) \\ &+\lambda_2\frac{2}{|\mathcal{D}|}\sum_{(\mathbf{y}^{(m)},\mathbf{y}^{(n)})\in\mathcal{D}}\mathbb{I}(\|\mathbf{t}^{(m)} - \mathbf{t}^{(n)}\|^2 < 1) \\ &(t_k^{(n)} - t_k^{(m)})(\frac{\partial t_k^{(m)}}{\partial W_i^k} - \frac{\partial t_k^{(n)}}{\partial W_i^k}) \end{aligned} \tag{9}$$

$$\begin{aligned} &\nabla U_j^k = \mathbb{E}_p[\varphi(z_j)\psi(h_k)^{\mathsf{T}}] - \mathbb{E}_{\hat{p}}[\varphi(z_j)\psi(h_k)^{\mathsf{T}}] \\ &+\lambda_1\frac{2}{|\mathcal{S}|}\sum_{(\mathbf{y}^{(m)},\mathbf{y}^{(n)})\in\mathcal{S}}(t_k^{(m)} - t_k^{(n)})(\frac{\partial t_k^{(m)}}{\partial U_j^k} - \frac{\partial t_k^{(n)}}{\partial U_j^k}) \\ &+\lambda_2\frac{2}{|\mathcal{D}|}\sum_{(\mathbf{y}^{(m)},\mathbf{y}^{(n)})\in\mathcal{D}}\mathbb{I}(\|\mathbf{t}^{(m)} - \mathbf{t}^{(n)}\|^2 < 1) \\ &(t_k^{(n)} - t_k^{(m)})(\frac{\partial t_k^{(m)}}{\partial U_j^k} - \frac{\partial t_k^{(n)}}{\partial U_j^k}) \end{aligned} \tag{10}$$

where $\mathbb{E}_p[\cdot]$ is the expectation w.r.t the true distribution and $\mathbb{E}_{\hat{p}}[\cdot]$ is the expectation w.r.t the empirical distribution. Exact computation of $\mathbb{E}_p[\cdot]$ is intractable. $\mathbb{E}_p[\cdot]$ is approximated by running a few step of Gibbs sampling starting from $\mathbb{E}_{\hat{p}}[\cdot]$. The sampling can be iteratively done as follows

$$\begin{aligned} \mathbb{E}[h_k^l] &= \mathbb{E}_{p(h_k|\mathbf{x},\mathbf{z})}[h_k|\mathbb{E}[\mathbf{x}^{l-1}], \mathbb{E}[\mathbf{z}^{l-1}]] \\ \mathbb{E}[x_i^l] &= \mathbb{E}_{p(x_i|\mathbf{h})}[x_i|\mathbb{E}[\mathbf{h}^{l-1}]] \\ \mathbb{E}[z_j^l] &= \mathbb{E}_{p(z_j|\mathbf{h})}[z_j|\mathbb{E}[\mathbf{h}^{l-1}]] \end{aligned} \tag{11}$$

where l is the index of iterations.

4 Experiments

We have developed the general framework of large scale multi-modal distance metric learning. To corroborate the effectiveness and efficiency of our method, we evaluate it on tagged images data. Images tagged with textual tags are quite prevalent in photo sharing websites (like Flickr, Pinterest, Instagram), where each image is associated with user tags, title, description and comments.

4.1 MMDML on Tagged Images

To specialize the general MMDML framework to tagged images data, we need to specify the local conditionals defined in Eq.(3) to concrete exponential family distributions according to data characteristics. Specially, we consider two modalities: a discrete bag-of-words representation $\mathbf{x}$ of text and a normalized bag-of-words representation $\mathbf{z}$ of image based on SIFT [Lowe, 2004] feature. We assume each x_i is a Bernoulli variable denoting whether the ith term of a tag dictionary appears or not around an image. Each z_j is a Gaussian variable denoting the normalized bag-of-words representation based on SIFT feature.

We assume each hidden variable h_k follows a Gaussian distribution conditioned on both text modality $\mathbf{x}$ and image modality $\mathbf{z}$. We can define a dual-wing harmonium model in a bottom-up manner by specifying the local conditionals

$$\begin{aligned} p(x_i = 1|\mathbf{h}) &= \frac{1}{1+\exp(-(\theta_i + \sum_k \mathbf{W}_{ik} h_k))} \\ p(z_j|\mathbf{h}) &= \mathcal{N}(z_j|\eta_j + \sum_k \mathbf{U}_{jk} h_k, 1) \\ p(h_k|\mathbf{x},\mathbf{z}) &= \mathcal{N}(h_k|\sum_i \mathbf{W}_{ik} x_i + \sum_j \mathbf{U}_{jk} z_j, 1) \end{aligned} \tag{12}$$

From the definition of the conditional distribution of $\mathbf{h}$ over the observations $\mathbf{x}, \mathbf{z}$, we can easily infer the latent representation $\mathbf{t}$

$$t_k = \mathbb{E}_{p(\mathbf{h_k}|\mathbf{x},\mathbf{z};\Theta)}[\mathbf{h_k}] = \sum_i \mathbf{W}_{ik} x_i + \sum_j \mathbf{U}_{jk} z_j \tag{13}$$

Accordingly, the sub-gradients defined in Eq.(8-10) can be specialized to

$$\nabla\theta_i = \mathbb{E}_p[x_i] - \mathbb{E}_{\hat{p}}[x_i], \nabla\eta_j = \mathbb{E}_p[z_j] - \mathbb{E}_{\hat{p}}[z_j] \tag{14}$$

$$\begin{aligned} \nabla W_i^k &= \mathbb{E}_p[x_i t_k] - \mathbb{E}_{\hat{p}}[x_i t_k] \\ &+ \lambda_1 \tfrac{2}{|\mathcal{S}|} \sum_{(\mathbf{y}^{(m)},\mathbf{y}^{(n)})\in\mathcal{S}} (t_k^{(m)} - t_k^{(n)})(x_i^{(m)} - x_i^{(n)}) \\ &+ \lambda_2 \tfrac{2}{|\mathcal{D}|} \sum_{(\mathbf{y}^{(m)},\mathbf{y}^{(n)})\in\mathcal{D}} \mathbb{I}(\|\mathbf{t}^{(m)} - \mathbf{t}^{(n)}\|^2 < 1) \\ &\quad (t_k^{(n)} - t_k^{(m)})(x_i^{(m)} - x_i^{(n)}) \end{aligned} \tag{15}$$

$$\begin{aligned} \nabla U_j^k &= \mathbb{E}_p[z_j t_k] - \mathbb{E}_{\hat{p}}[z_j t_k] \\ &+ \lambda_1 \tfrac{2}{|\mathcal{S}|} \sum_{(\mathbf{y}^{(m)},\mathbf{y}^{(n)})\in\mathcal{S}} (t_k^{(m)} - t_k^{(n)})(z_j^{(m)} - z_j^{(n)}) \\ &+ \lambda_2 \tfrac{2}{|\mathcal{D}|} \sum_{(\mathbf{y}^{(m)},\mathbf{y}^{(n)})\in\mathcal{D}} \mathbb{I}(\|\mathbf{t}^{(m)} - \mathbf{t}^{(n)}\|^2 < 1) \\ &\quad (t_k^{(n)} - t_k^{(m)})(z_j^{(m)} - z_j^{(n)}) \end{aligned} \tag{16}$$

The Gibbs sampling process defined in Eq.(11) can be specialized to

$$\begin{aligned} \mathbb{E}[h_k^l] &= \textstyle\sum_i \mathbf{W}_{ik} \mathbb{E}[x_i^{l-1}] + \sum_j \mathbf{U}_{jk} \mathbb{E}[z_j^{l-1}] \\ \mathbb{E}[x_i^l] &= \frac{1}{1+\exp(-(\theta_i + \sum_k \mathbf{W}_{ik} \mathbb{E}[h_k^{l-1}]))} \\ \mathbb{E}[z_j^l] &= \eta_j + \textstyle\sum_k \mathbf{U}_{jk} \mathbb{E}[h_k^{l-1}] \end{aligned} \tag{17}$$

4.2 Dataset

The dataset used in our experiments is NUS-WIDE-1.5K: a subset selected from NUS-WIDE dataset [Chua *et al.*, July 8 10 2009]. Images are downloaded from Flickr and each image is associated with more than one user tags. We choose 30 classes and select about 50 images for each class. The total number of images is 1536. The 30 classes are actor, airplane, bicycle, bridge, buddha, building, butterfly, camels, car, cathedral, cliff, clouds, coast, computers, desert, flag, flowers, food, forest, glacier, hills, lake, leaf, monks, moon, motorcycle, mushrooms, ocean, police, pyramid. We randomly choose half of the images for training and the other half for testing. For text modality, 1000 tags with top frequency are selected to form the tag dictionary. For image modality, we extract SIFT based bag-of-words representation with a codebook of size 1024.

Following [Davis *et al.*, 2007; Xing *et al.*, 2002], we sample "similar" pairs by picking up two instances from the same class and "dissimilar" pairs by choosing two instances from different classes. We randomly sample about 10K "similar" pairs and 10K "dissimilar" pairs from the training set.

4.3 Experiment Setup

We compare with the following baselines:

- Xing+Original. We concatenate original feature vectors of text modality and image modality into a single representation and subsequently learn a Mahalanobis distance using the metric learning method proposed in [Xing *et al.*, 2002].
- ITML+Original. We combine features of text and image into a whole and feed it to the ITML [Davis *et al.*, 2007] method.
- Xing+MWH. We use the unsupervised MWH model to embed data from text and image modalities to the latent space and learn distance measure on the latent representations using the method proposed in [Xing *et al.*, 2002].
- ITML+MWH. We use ITML [Davis *et al.*, 2007] to learn distance on the latent feature vectors obtained from MWH model.
- MKE. We compare with the multiple kernel embedding method proposed in [McFee and Lanckriet, 2011].

Table 1: Average precision (AP) of image retrieval on NUS-WIDE-1.5K dataset

Method	Xing+Original	ITML+Original	Xing+MWH	ITML+MWH	MKE	Our method
AP	0.5482	0.4078	0.7673	0.6907	0.4972	**0.8409**

Table 2: k-NN classification accuracy on NUS-WIDE-1.5K dataset

Method	Xing+Original	ITML+Original	Xing+MWH	ITML+MWH	MKE	Our method
1-NN	0.8995	0.8995	0.8995	0.9286	0.8056	**0.9352**
3-NN	0.8108	0.6653	0.8849	0.8929	0.6944	**0.9021**
5-NN	0.6971	0.4868	0.8426	0.8519	0.5860	**0.8849**
10-NN	0.4775	0.2394	0.7646	0.7394	0.4405	**0.8333**
20-NN	0.1548	0.0450	0.6230	0.4841	0.1746	**0.7130**

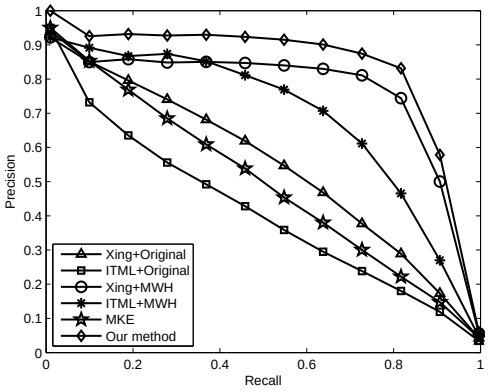

Figure 2: Precision-recall curve for image retrieval on NUS-WIDE-1.5K dataset

All parameters are tuned by 5-fold cross validation. In our method MMDML, the dimension K of latent variables is set to 100. Trade-off parameters λ_1, λ_2 are set to 100. Coupling matrices $\mathbf{W}$ and $\mathbf{U}$ defined in Eq.(12) are initialized using SVD. Partial SVDs are performed on the design matrices over text features and image features respectively. Singular vectors corresponding to the top K singular values are selected as initial values of $\mathbf{W}$ and $\mathbf{U}$. We run stochastic gradient descent with a fixed step size of 0.000001 and 200 iterations. The Gibbs sampling iterations in contrastive divergence is set to 1. For unsupervised MWH used in two baseline methods Xing+MWH and ITML+MWH, the parameters are the same as those in MMDML. For ITML, the trade-off parameter is set to 10. For MKE, we use Gaussian kernel with bandwidth 1 to compute the kernel matrices. The trade-off parameter is set to 10000.

After obtaining the learned parameters, we infer the latent representation for each test image using Eq.(13). We use Euclidean distance computed on latent representations as distance measure. Next, we report experimental results on image retrieval and classification.

4.4 Retrieval and Classification

For image retrieval, each test image is treated as query and the other images in the test set are ranked according to their distances with the given query. An image is considered relevant to query if it share the same class label with query image. We use average precision (AP) [Smeaton and Over, 2003] and precision-recall curve to measure retrieval performance.

The AP result is summarized in Table 1. The precision-recall curve is shown in Figure 2. As can be seen from Table 1 and Figure 2, our method MMDML significantly outperforms baseline methods. MMDML achieves an average precision score of 0.8409, which is greatly higher than the second best method Xing+MWH. Note that, Xing+MWH and ITML+MWH both achieve substantial improvements compared with Xing+Original and ITML+Original. The reasons are two folds. First, Xing+MWH and ITML+MWH operate in a space with much lower dimension than the original feature space. Thereby the risk of overfitting is reduced. Second, MWH model maps two different modalities into a single latent space, which captures the correlation and complementary relationships between modalities. Xing+Original and ITML+Original naively concatenate heterogeneous modalities into a whole, where the hidden structure of modalities is not explored. However, compared with our method, Xing+MWH and ITML+MWH are suboptimal. In these two methods, multi-modal embedding and distance metric learning are performed separately. MWH is employed to project different modalities into the latent space and subsequently distance measure is learned in this latent space. In our method, multi-modal embedding and distance metric learning are jointly performed to achieve the overall optimality, thus yielding much better performance.

We also apply the learned distance measure for k-nearest neighbor (k-NN) classification. Table 2 summarizes the classification accuracy for k=1, 3, 5, 10, 20. Under varying k, our method consistently outperforms other methods.

4.5 Parameter Sensitivity

We test the sensitivity of MMDML to different choices of the dimension K of latent space and the tradeoff parameter λ_1, λ_2. Throughout the experiments, we set λ_1 and λ_2 to the same value to indicate equal importance of "similar" pairs and "dissimilar" pairs. When evaluating one parameter, the other parameters are fixed to values reported in Section 4.3. Figure 3(a) shows the variation of average precision (AP) with varying K. As can be seen, the retrieval performance is robust to the choice of K. For K larger than 10, the average precision

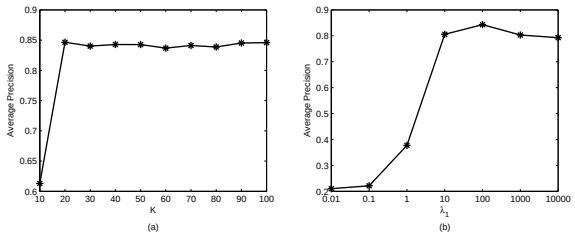

Figure 3: Retrieval performance sensitivity with respect to K and λ_1

almost remains the same. This suggest that we can choose a sufficiently small K to reduce computational cost and storage cost without degrading retrieval performance. Figure 3(b) shows how AP varies with different λ_1. λ_1 has a strong influence on MMDML. A larger λ_1 means more emphasis over distance metric learning loss. For small λ_1 (0.01, 0.1, 1), the average precision is very low. As λ_1 increases to 10 and 100, AP is dramatically improved. However, further increasing λ_1 results in performance decreasing. When λ_1 is increased from 10 to 10000, the average score drops from 0.8433 to 0.7928. The possible reason is that too large λ_1 makes the model overfitted to training data.

4.6 Computational Efficiency

To evaluate the efficiency of MMDML, we compare its running time with Xing+Original, ITML+Original, MKE. The comparison is performed on a machine with 3.3GHz quadcore Intel processor and 8GB memory. Table 3 summarizes the results. ITML is substantially time-consuming. It takes about 6.5 hours to converge. Considering its computational complexity grows quadratically with feature dimension and linearly with number of constraints, we can conclude this method is not applicable for large scale problems. Xing+Original takes reasonable time to converge. This method requires eigen-decomposition over the Mahalanobis matrix, which is of size 2024 in NUS-WIDE-1.5K dataset. However, in problems of high feature dimension, eigen-decomposition will be extremely computationally demanding, if possible. MKE takes the least time in this experiment. However, this does not imply that this method is scalable. In MKE, each iteration of the projected gradient solver requires eigen-decompositions, whose complexity is $O(mn^3)$, where m is the number of modalities and n is the number of training examples. Clearly, MKE is not scalable for large dataset whose n can be very huge. Our method takes about 6.5 minutes for learning. It requires no eigen-decompositions and facilitates efficient stochastic gradient descent optimization. In our method, we mainly need to compute Eq.(8-10) which involves simple arithmetic operations and one step Gibbs sampling which turns out to be very efficient [Hinton, 2002].

Table 3: Computational time (in seconds) on NUS-WIDE-1.5K dataset

Method	Time(s)
Xing+Original	859
ITML+Original	23460
MKE	350
Our method	387

5 Conclusions

We introduce a general framework of multi-modal distance metric learning based on multi-wing harmonium model. The framework can flexibly embed arbitrary number of feature modalities into a shared latent space where distance supervision is encoded. We apply the method to tagged image retrieval and experiments demonstrate the effectiveness and scalability of our method.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments and suggestions and thank Jun Zhu and Brian McFee for providing their code.

References

[Aizenberg *et al.*, 2012] N. Aizenberg, Y. Koren, and O. Somekh. Build your own music recommender by modeling internet radio streams. In *Proceedings of the 21st international conference on World Wide Web*, pages 1–10. ACM, 2012.

[Baluja *et al.*, 2008] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th international conference on World Wide Web*, pages 895–904. ACM, 2008.

[Bekkerman and Jeon, 2007] R. Bekkerman and J. Jeon. Multi-modal clustering for multimedia collections. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[Chen *et al.*, 2010] N. Chen, J. Zhu, and E.P. Xing. Predictive subspace learning for multi-view data: A large margin approach. *Advances in Neural Information Processing Systems*, 24, 2010.

[Chua *et al.*, July 8 10 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009.

[Davis *et al.*, 2007] J.V. Davis, B. Kulis, P. Jain, S. Sra, and I.S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

[Globerson and Roweis, 2006] A. Globerson and S. Roweis. Metric learning by collapsing classes. *Advances in neural information processing systems*, 18:451, 2006.

[Hinton, 2002] G.E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[Li *et al.*, 2009] Ming Li, Xiao-Bing Xue, and Zhi-Hua Zhou. Exploiting multi-modal interactions: A unified framework. In *Proeeedings of the 21st International Joint Conference on Artificial Intelligence. Pasadena, CA*, pages 1120–1125, 2009.

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

[McFee and Lanckriet, 2011] B. McFee and G. Lanckriet. Learning multi-modal similarity. *The Journal of Machine Learning Research*, 12:491–523, 2011.

[Nishida *et al.*, 2012] K. Nishida, T. Hoshide, and K. Fujimura. Improving tweet stream classification by detecting changes in word probability. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 971–980. ACM, 2012.

[Qi *et al.*, 2012] G.J. Qi, C.C. Aggarwal, and T.S. Huang. On clustering heterogeneous social media objects with outlier links. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 553–562. ACM, 2012.

[Smeaton and Over, 2003] Alan F Smeaton and Paul Over. Trecvid: Benchmarking the effectiveness of information retrieval tasks on digital video. In *Image and Video Retrieval*, pages 19–27. Springer, 2003.

[Weinberger *et al.*, 2006] K.Q. Weinberger, J. Blitzer, and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. In *In NIPS*. Citeseer, 2006.

[Xing *et al.*, 2002] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. *Advances in neural information processing systems*, 15:505–512, 2002.

[Xing *et al.*, 2005] E.P. Xing, R. Yan, and A. Hauptmann. Mining associated text and images with dual-wing harmoniums. 2005.

[Zhang *et al.*, 2011] D. Zhang, F. Wang, and L. Si. Composite hashing with multiple information sources. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011.

[Zhen and Yeung, 2012] Y. Zhen and D.Y. Yeung. A probabilistic model for multimodal hash function learning. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012.