An LLE based Heterogeneous Metric Learning for Cross-media Retrieval

Peng Zhou [*†] Liang Du [*] Mingyu Fan [‡] Yi-Dong Shen [*§]

Abstract

With unstructured heterogeneous multimedia data such as texts, images being more and more widely used on the web, cross-media retrieval has become an increasingly important task. One of the key techniques in cross-media retrieval is how to compute distances or similarities among different types of media data. In this paper, we propose a novel heterogeneous metric learning method to compute distances between images and texts. We extend Locally Linear Embedding (LLE) to deal with heterogeneous data, so that we can not only preserve homogeneous local information but also capture heterogeneous constraints. In order to handle the out-of-sample problem, we learn two map functions from the embedding, and use them to transform heterogeneous data into a homogeneous space and do the retrieval in the new space. The experimental results on two real-world datasets show the effectiveness of our approach.

1 Introduction

With the rapid increasing of unstructured heterogeneous multimedia data such as image, text, video on the Internet, cross-media retrieval has become an important task [25, 26, 17, 8]. Given a query, cross-media retrieval aims to find related data with different data types. For example, when the query is a picture, the related data could be texts, images, videos, and so on. Different from traditional single-media retrieval used in search engines, cross-media retrieval makes full use of cross media relations among different modalities. Such relations play important roles in understanding multimedia contents.

One key problem in cross-media retrieval is how to measure distances or similarities between heterogeneous multimedia objects. One commonly used way to measure such distances is via metric learning, which learns a linear transformation from the source data space into a new space; then the distance is computed over the new space using traditional distance metric such as the Euclidean distance. Typical approaches such as [23, 2, 1, 6, 10, 16, 9] learn a distance metric that keeps all the data points with the same label close, while separating data points with different labels far apart. These methods can only deal with homogeneous data, and are difficult to be used in cross-media retrieval directly because heterogeneous multimedia objects are represented in different feature spaces.

Heterogeneous metric learning is a relatively new problem. To the best of our knowledge, the state-of-the-art heterogeneous metric learning methods include [18, 14, 22, 25, 20]; they use some supervised heterogeneous information as the bridge to connect different types of data. For example, in [18, 14], they have a coupled constraint which is a one-to-one relation between texts and images. More specifically, each couple of image and text expresses that the image and text are about the same thing. In [25, 22], they use must-link and cannot-link constraints. If the image and the text are related, there is a must-link constraint between them; otherwise, there is a cannot-link constraint. With the heterogeneous constraints, they learn two linear functions to transform all data into a homogeneous space and measure distances in this homogeneous space.

However, it is quite often that there is no such linear relation available between source data spaces and target data spaces, so that the above mentioned metric learning cannot be applied; then non-linear embedding is introduced. Instead of learning a linear transformation function, non-linear embedding such as Linear Locally Embedding (LLE) [19] and Isomap [21] directly learns the embedding results which meet some specific properties. For example, LLE learns the vectors in the new space to recover global non-linear structure from locally linear fits so that it can preserve neighborhood information. Like homogeneous metric learning, these methods are also difficult to handle heterogeneous data. In addition, although non-linear methods are more general and offer greater separation ability in theory [9], since they learn embedding results

[*] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China $\{zhoup, duliang, ydshen\}$@*ios.ac.cn*

[†] University of Chinese Academy of Sciences, Beijing 100049, China

[‡] Institute of Intelligent System and Decision, Wenzhou University, Zhejiang 325035, China *fanmingyu@amss.ac.cn*

[§] Corresponding author

directly rather than a map function, such methods are quite inefficient to deal with new data (the out-of-sample problem). Therefore, to measure distances between heterogeneous data by non-linear embedding, two questions remain to be investigated: (I) How to extend it to handle heterogeneous data? (II) How to deal with the out-of-sample problem?

To answer the above two questions, in this paper we propose a novel heterogeneous metric learning method which is based on non-linear embedding so that it can capture some non-linear relation between data. This approach extends the well-known LLE method to heterogeneous data, thus is named Locally Linear Embedding based Heterogeneous Metric Learning (LLEHML). For simplicity, we consider two data types: image and text, and it is easy to extend our method to other multimedia data. We assume that images and texts come from a unified "original" homogeneous space and in this space, images and texts are sampled from their respective underlying manifold. Because images and texts are in different "description" spaces (images consist of pixels and texts consist of words), we need to map them from "description" spaces into the unified "original" space. As LLE is a neighborhood-preserved embedding method and can capture manifold information well, we use it to compute and preserve locally linear reconstruction weight of each media data so that we can recover global non-linear structure from these local information. Additionally, to handle heterogeneous data, we force the embedding results to preserve the heterogeneous constraints. In a summary, our method can capture both homogeneous local information and heterogeneous constraints. At last, LLE is an embedding method rather than metric learning method as introduced before. Once we get new data, we have to run the whole algorithm again and it is very expensive. To handle the out-of-sample problem, we use a linear approximation to transform the embedding method into a metric learning method.

Experiments on two real-world datasets show that our proposed approach outperforms the current state-of-the-art methods.

2 Related Work

Heterogeneous metric learning, which is most relevant to our work, enables us to measure distances from different types of data. Existing methods of heterogeneous metric learning focus on learning linear transformations from heterogeneous data spaces to a homogeneous space. For example, [18] applied Canonical Correlation Analysis (CCA) to cross-media retrieval. CCA [11] is a data analysis and dimensionality reduction method similar to Principal Component Analysis (PCA) [12] . While PCA deals with only one data space, CCA is a technique for joint dimensionality reduction across two (or more) spaces. It attempts to maximize the correlation between same labeled objects in the transformed space. Based on the learning results of CCA, [18] further learned a high-level semantic metric by logistic regression. This method considers not only correlation analysis but also semantic abstraction for different data types.

Li et al. [14] used Cross-modal Factor Analysis (CFA) to find the optimal transformations that can best represent (or identify) the coupled patterns between features of two different subsets. Unlike CCA, CFA adopts a criterion of minimizing the Frobenius norm between same labeled objects in the transformed space.

Both CCA and CFA consider only pairs of the same labeled objects. They do not explicitly separate different labeled objects. To overcome this problem, [22] applied Partial Least Square (PLS) to learn two orthogonal transformation matrices by minimizing the distances between same labeled objects and maximizing the distances between different labeled objects.

Some researches show that hash function can be used to map different types of data into a new Hamming space [20, 3, 13]. [20] is one of the current state-of-the-art cross-media hash methods. It proposed an inter-media hashing mode to explore the correlation among multiple media types from different data sources and tackle the scalability issue. It learned two linear hash function to map data into a common Hamming space, in which fast search can be easily implemented by XOR and bit-count operations.

Zhai et al. [25] proposed a new approach which integrates joint graph regularization into the objective function to preserve the similarity constraints in both modalities. It used must-link and cannot-link constraints and made objects in embedding space closer if they had a must-link and further if they had a cannot-link. Then a manifold ranking based algorithm was used to get a high level semantic metric.

3 Heterogeneous Metric Learning

In this section, we elaborate the details of the proposed heterogeneous metric learning method. First, we briefly introduce the whole framework. Then we depict three steps of our method in detail.

3.1 Framework Overview In this subsection, we introduce the framework of our heterogeneous metric learning. Let $\mathbb{D}^x = \{x_1, \ldots, x_p, x_{p+1}, \ldots, x_n\}$ be an image dataset, where $x_i \in \mathcal{X}$ denotes image data, and $\mathbb{D}^y = \{y_1, \ldots, y_p, y_{p+1}, \ldots, y_m\}$ be a text dataset, where $y_i \in \mathcal{Y}$ denotes text data. To bridge these

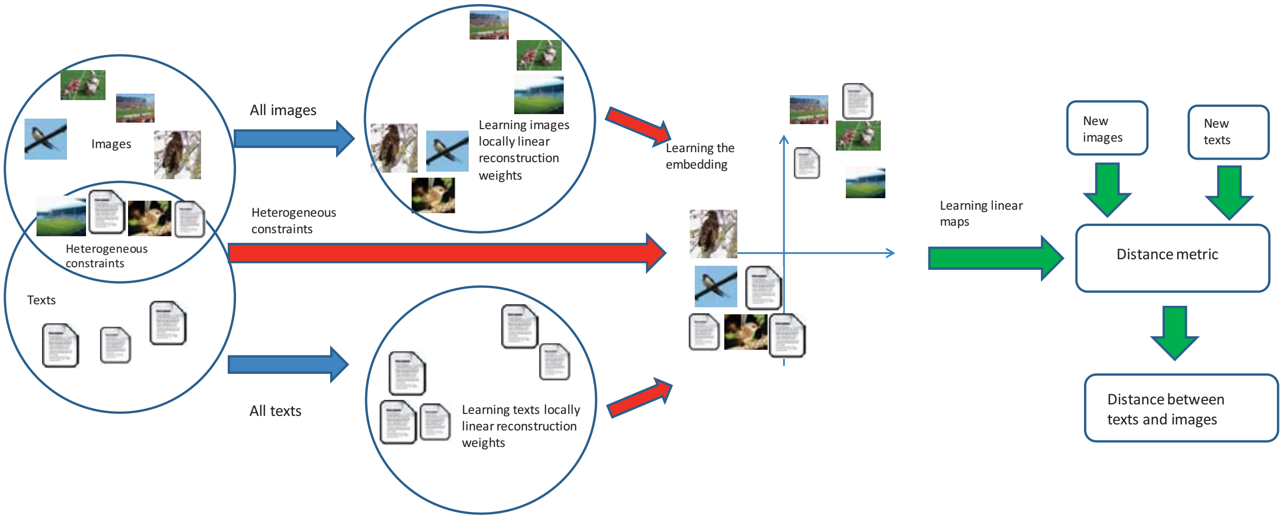

Images
Heterogeneous constraints
Texts
All images
Learning images locally linear reconstruction weights
Heterogeneous constraints
All texts
Learning texts locally linear reconstruction weights
Learning the embedding
Learning linear maps
New images
New texts
Distance metric
Distance between texts and images

Figure 1: The overview of our heterogeneous metric learning framework.

Table 1: Notations and descriptions used in this section

Notation	Description
n	number of images
m	number of texts
d	dimension of new representations of image and text
d^x	dimension of original image's feature
d^y	dimension of original text's feature
$\mathcal{X}$, $\mathcal{Y}$,$\mathcal{Z}$	image data space, text data space, and embedding space
α, β, γ_1, γ_2	balancing parameter
k	number of nearest neighbour
$\mathbf{X}$	$d^x \times n$ matrix, each column represents an image data
$\mathbf{Y}$	$d^y \times n$ matrix, each column represents an text data
x_i, y_i	the i-th column of $\mathbf{X}$ or $\mathbf{Y}$
$\mathbf{Z}^x$	$d \times n$ matrix, each column represents an image in new space
$\mathbf{Z}^y$	$d \times n$ matrix, each column represents an text in new space
z_i^x, z_i^y	i-th column of $\mathbf{Z}^x$ or $\mathbf{Z}^y$
w_i^x, w_i^y	the reconstruction weight vectors
w_{ij}^x, w_{ij}^y	the j-th element in vector w_i^x or w_i^y
$\mathbf{V}^x$, $\mathbf{V}^y$	linear map matrix
$\mathcal{S}$	must-link constraint set
$\mathcal{D}$	cannot-link constraint set

image and text data for heterogeneous matching, we carefully investigate and utilize additional supervised context information; these information can be collected in various ways such as tags and category information. Thus, $\{x_1, \ldots, x_p\}$ and $\{y_1, \ldots, y_p\}$ are used to denote data with these heterogeneous connections while other data points are not directly related. This context information can be used in different ways, for example, in CCA, they use coupled constraints to ensure that x_i and y_i are two descriptions of one same object for any $1 \leqslant i \leqslant p$. In our work, we use must-link and cannot-link constraints as [22, 25] did. It means that for any $1 \leqslant i, j \leqslant p$ if x_i and y_j is similar, there is a must-link constraint between them and otherwise there is a cannot-link constraint. We use must-link and cannot-link constraints because they are more informative than coupled constraints and are not difficult to get. We will introduce in detail how to get the constraints in Section 4, because it depends on datasets.

The goal of heterogeneous metric learning is to learn two linear functions to transform the dataset $\mathbb{D}^x$ and $\mathbb{D}^y$ into a new dataset $\mathbb{Z} = \{z_1, \ldots, z_{n+m}\}$, where $z_i \in \mathcal{Z}$ is in a homogeneous space $\mathcal{Z}$. To achieve a good performance for heterogeneous metric learning method, on the one hand, it should well characterize the intrinsic homogeneous data structure (geometrical or discriminative) for each description separately; on the other hand, the heterogeneous relation should be well preserved. Thus, our proposed method considers not only constraints between texts and images but also local information in each data type. It has three steps: firstly, we learn locally linear reconstruction weights in each data space; secondly, we use the reconstruction weights and heterogeneous constraints to learn the embedding of the data; finally, we learn two linear approximation functions which transform the embedding to metric learning. After getting the linear functions, we use these functions to map images and texts into new space, and measure their Euclidean distances in the new space. Figure 1 shows the whole framework of our method. Table 1 shows the notations and descriptions used in this section.

3.2 Locally Linear Reconstruction In this stage, we deal with homogeneous information and have the same assumptions of LLE [19]. We suppose that the text and image data are sampled from their underlying manifold respectively. We expect each data point (image and text) and its neighbors to lie on or close to a locally linear patch of the manifold and characterize the local geometry of these patches by linear coefficients that reconstruct each data point from its neighbors. Therefore, we learn the locally linear reconstruction weights of images and texts respectively.

Take images as example, to learn locally linear reconstruction weights, we first compute k nearest neighbors of each image. We use Euclidean distance to find k nearest neighbor. Here we follow the assumption: for neighboring points, Euclidean distance provides a good approximation to real distance. Therefore, finding k nearest neighbor by Euclidean distance is feasible.

After getting k nearest neighbor, we minimize the following cost function for each image x_i to learn the reconstruction weight of x_i

$$(3.1)\qquad J_1{}^i(\mathbf{w_i^x}) = \|\mathbf{x_i} - \sum_{j=1}^{k} w_{ij}^x \mathbf{x_{N(j)}^i}\|^2 + \frac{\alpha}{2}\|\mathbf{w_i^x}\|^2$$

where $\mathbf{x_{N(j)}^i}$ is the j-th nearest neighbor of x_i, $\mathbf{w_i^x}$ is a k-dimension vector which contains w_{ij}^x as its j-th element, and w_{ij}^x means the weight of j-th nearest neighbor of image x_i. The first term is to minimize the reconstruction error and the second term is to avoid over-fitting as [7] did. For each $\mathbf{x_i}$, the weights w_{ij}^x sum up to 1. Then the above equation is equal to:

$$(3.2)\qquad \begin{aligned} J_1{}^i(\mathbf{w_i^x}) &= \|\sum_{j=1}^{k} w_{ij}^x(\mathbf{x_i} - \mathbf{x_{N(j)}^i})\|^2 + \frac{\alpha}{2}\|\mathbf{w_i^x}\|^2 \\ &= \sum_{j=1}^{k}\sum_{l=1}^{k} w_{ij}^x w_{il}^x Q_{jl}^i + \frac{\alpha}{2}\|\mathbf{w_i^x}\|^2 \end{aligned}$$

where Q_{jl}^i is the (j,l)-th element of a $k \times k$ matrix $\mathbf{Q}^i$:

$$(3.3)\qquad Q_{jl}^i = (\mathbf{x_i} - \mathbf{x_{N(j)}^i})^T(\mathbf{x_i} - \mathbf{x_{N(l)}^i})$$

Now we need to solve the equality constrained minimization problem:

$$\min_{w_i^x} \sum_{j=1}^{k}\sum_{l=1}^{k} w_{ij}^x w_{il}^x Q_{jl}^i + \frac{\alpha}{2}\|w_i^x\|^2 \quad s.t.\ \sum_{j=1}^{k} w_{ij}^x = 1$$

By introducing a Lagrange multiplier λ, we reformulate it to an unconstrained minimization:

$$(3.4)\quad L^i = \sum_{j=1}^{k}\sum_{l=1}^{k} w_{ij}^x w_{il}^x Q_{jl}^i + \frac{\alpha}{2}\|w_i^x\|^2 + \lambda(\sum_{j=1}^{k} w_{ij}^x - 1)$$

Setting the partial derivative with respect to w_i^x to zero and considering $\sum_{j=1}^{k} w_{ij}^x = 1$ in addition, we can get:

$$(3.5)\qquad w_{ij}^x = \frac{\sum_{l=1}^{k} R_{jl}^i}{\sum_{p=1}^{k}\sum_{q=1}^{k} R_{pq}^i}$$

where $\mathbf{R}^i = (\mathbf{Q}^i + \alpha\mathbf{I})^{-1}$ and R_{jl}^i is the (j,l)-th element in $\mathbf{R}^i$. Empirically set $\alpha = 0.001 \times tr(\mathbf{Q}^i)$.

For all images, we use Eq.(3.5) to compute their reconstruction weights. Then we get reconstruction weight matrix $\mathbf{U}^x$ of all images:

$$U_{il}^x = \begin{cases} w_{ij}^x, & \textit{If the } l\textit{-th image is the } j\textit{-th} \\ & \textit{nearest neighbor of the } i\textit{-th image} \\ 0, & \textit{Otherwise} \end{cases}$$

where U_{il}^x is the (i,l)-th element of $\mathbf{U}^x$.

We deal with the texts in the same way as images and get $\mathbf{U}^y$ similarly.

Rewrite $\mathbf{U}^x$ and $\mathbf{U}^y$ together to get reconstruction weight matrix $\tilde{\mathbf{U}}$ of all data (images and texts):

$$\tilde{\mathbf{U}} = \begin{pmatrix} \mathbf{U}^x & \mathbf{0} \\ \mathbf{0} & \mathbf{U}^y \end{pmatrix}$$

3.3 Heterogeneous Embedding After getting reconstruction weights of data, we expect to learn new d-dimension vectors $z_i^x, z_j^y \in \mathcal{Z}$ to preserve the reconstruction information in embedding space. Additionally, this embedding vectors should also capture the heterogeneous constraints.

To minimize the image reconstruction error in the embedding space, we minimize the loss function:

$$(3.6)\qquad J_2(\mathbf{Z}^x) = \sum_{i=1}^{n}\|z_i^x - \sum_{j=1}^{k} w_{ij}^x z_{i,N(j)}^x\|^2$$

where z_i^x is new representation of x_i in the embedding space $\mathcal{Z}$, $\mathbf{Z}$ contains z_i^x as the i-th column, and $z_{i,N(j)}^x$ is the new representation of j-th nearest neighbor of image

x_i. w_{ij}^x is locally linear reconstruction weights learned in previous subsection.

Similarly, we minimize the loss function of text reconstruction error:

$$J_3(\mathbf{Z}^y) = \sum_{i=1}^{n} \|z_i^y - \sum_{j=1}^{k} w_{ij}^y z_{i,N(j)}^y\|^2 \tag{3.7}$$

where z_i^y, $\mathbf{Z}^y$, and $z_{i,N(j)}^y$ are defined similar to z_i^x, $\mathbf{Z}^x$, and $z_{i,N(j)}^x$.

To make the similar image and text compact while dissimilar image and text diverse in the embedding space, we define the loss function as follows:

$$J_4(\mathbf{Z}^x, \mathbf{Z}^y) = \sum_{(x_i,y_j)\in\mathcal{S}} \|z_i^x - z_j^y\|^2 - \sum_{(x_i,y_j)\in\mathcal{D}} \|z_i^x - z_j^y\|^2 \tag{3.8}$$

To simplify Eq.(3.8), we define the heterogeneous constraints. Here we use must-link and cannot-link constraints c_{ij} of the i-th image and the j-th text defined as follows:

$$c_{ij} = \begin{cases} 1, & (x_i, y_j) \in \mathcal{S} \\ -1, & (x_i, y_j) \in \mathcal{D} \\ 0, & otherwise \end{cases} \tag{3.9}$$

The c_{ij}'s can be stored in an $n \times m$ matrix $\mathbf{C}$, where c_{ij} is the (i,j)-th element of $\mathbf{C}$. Matrix $\mathbf{C}$ is just about heterogeneous data, and we can construct an $(m+n)\times(m+n)$ matrix $\mathbf{E}$ with $\mathbf{C}$:

$$\mathbf{E} = \begin{pmatrix} \mathbf{0} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{0} \end{pmatrix}$$

By introducing must-link and cannot-link constraints $\mathbf{C}$, we rewrite Eq.(3.8) as follows:

$$J_4(\mathbf{Z}^x, \mathbf{Z}^y) = \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} \|z_i^x - z_j^y\|^2 \tag{3.10}$$

By integrating J_2, J_3 and J_4 to a unified formulation, we get the final loss function:

$$\begin{aligned} J_5(\mathbf{Z}^x, \mathbf{Z}^y) &= \sum_{i=1}^{n} \|z_i^x - \sum_{j=1}^{k} w_{ij}^x z_{i,N(j)}^x\|^2 \\ &+ \sum_{i=1}^{m} \|z_i^y - \sum_{j=1}^{k} w_{ij}^y z_{i,N(j)}^y\|^2 + \beta \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} \|z_i^x - z_j^y\|^2 \end{aligned} \tag{3.11}$$

where β is a balancing parameter.

To minimize the loss function Eq.(3.11), we rewrite it first. The first term can be written as follows:

$$\sum_{i=1}^{n} \|z_i^x - \sum_{j=1}^{k} w_{ij}^x z_{i,N(j)}^x\|^2 = tr(\mathbf{Z}^x(\mathbf{I}-\mathbf{U}^x)^T(\mathbf{I}-\mathbf{U}^x)\mathbf{Z}^{xT})$$

The second term of Eq.(3.11) can be rewritten similarly:

$$\sum_{i=1}^{m} \|z_i^y - \sum_{j=1}^{k} w_{ij}^y z_{i,N(j)}^y\|^2 = tr(\mathbf{Z}^y(\mathbf{I}-\mathbf{U}^y)^T(\mathbf{I}-\mathbf{U}^y)\mathbf{Z}^{yT})$$

Note that z_i^x and z_j^y are in the same space, we can write $\mathbf{Z}^x$ and $\mathbf{Z}^y$ together as $\mathbf{Z} = [\mathbf{Z}^x, \mathbf{Z}^y]$. Let e_{ij} be the (i, j)-th element of $\mathbf{E}$ and z_{ij} be the (i, j)-th element of $\mathbf{Z}$. Let $\mathbf{D}$ be a diagonal matrix with its (i,i)-element $\mathbf{D}_{ii}$ equals to the sum of i-th row of $\mathbf{E}$. We rewrite the third term of Eq.(3.11) as follows:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} \|z_i^x - z_j^y\|^2 = tr(\mathbf{Z}(\mathbf{D}-\mathbf{E})\mathbf{Z}^T)$$

Then Eq.(3.11) can be re-written as:

$$J_5 = tr(\mathbf{Z}\mathbf{M}\mathbf{Z}^T) + \beta tr(\mathbf{Z}\mathbf{N}\mathbf{Z}^T) = tr(\mathbf{Z}(\mathbf{M}+\beta\mathbf{N})\mathbf{Z}^T)$$

where $\mathbf{M}$ is an $(m+n)\times(m+n)$ matrix found as $\mathbf{M} = (\mathbf{I}-\tilde{\mathbf{U}})^T(\mathbf{I}-\tilde{\mathbf{U}})$. $\mathbf{N}$ is also an $(m+n)\times(m+n)$ matrix found as $\mathbf{N} = \mathbf{D} - \mathbf{E}$. To be able to solve this problem, we can add a well-known orthogonal constraint: $\mathbf{Z}\mathbf{Z}^T = \mathbf{I}$.

Now we need to solve a minimization problem with orthogonal constraint as follows:

$$\min_{\mathbf{Z}} \; tr(\mathbf{Z}\mathbf{T}\mathbf{Z}^T) \quad s.t.\ \mathbf{Z}\mathbf{Z}^T = \mathbf{I} \tag{3.12}$$

where $\mathbf{T} = \mathbf{M} + \beta\mathbf{N}$.

According to Ky Fan theorem [24], the solution of Eq.(3.12) is $\mathbf{Z} = [\mathbf{z_1}, \mathbf{z_2}, ..., \mathbf{z_d}]$ where $\mathbf{z_1}...\mathbf{z_d}$ are the eigenvectors corresponding to the smallest d eigenvalues of $\mathbf{T}$. As result, the original heterogeneous data have been transformed into a new homogeneous space $\mathcal{Z}$. In this space, we can directly use Euclidean distance to measure the distance between any two objects.

3.4 Learning Linear Map

Although we can measure distances between images and texts after heterogeneous embedding, we can hardly handle out-of-sample data. If we get new images and texts, we have to compute reconstruction weights and do embedding once again. It is impractical obviously. Therefore, we should learn a map function instead. In heterogeneous data, x and y are not in a unified space, thus we need to learn two linear matrices $\mathbf{V}^x$ and $\mathbf{V}^y$ and rewrite the Mahalanobis distance as follows:

$$d(x, y) = \sqrt{(\mathbf{V}^x x - \mathbf{V}^y y)^T(\mathbf{V}^x x - \mathbf{V}^y y)} \tag{3.13}$$

In our method, we use two linear function to approximate the embedding method introduced in previous subsection. More specifically, we view $\mathbf{V}^x$ ($\mathbf{V}^y$) as a linear map matrix which transform original image (text) data $\mathbf{X}$ ($\mathbf{Y}$) to embedded data $\mathbf{Z}^x$ ($\mathbf{Z}^y$). Thus we compute $\mathbf{V}^x$ and $\mathbf{V}^y$ as follows:

$$
\begin{aligned}
&\min_{\mathbf{V}^x} \|\mathbf{Z}^x - \mathbf{V}^x\mathbf{X}\|_F^2 + \gamma_1\|\mathbf{V}^x\|_F^2 \\
&\min_{\mathbf{V}^y} \|\mathbf{Z}^y - \mathbf{V}^y\mathbf{Y}\|_F^2 + \gamma_2\|\mathbf{V}^y\|_F^2
\end{aligned} \tag{3.14}
$$

where $\|\cdot\|_F$ is the Frobenius norm and $\mathbf{Z}^x$, $\mathbf{Z}^y$ are computed by solving Eq.(3.12). Eq.(3.14) is classical ridge regression problem, thus we can get $\mathbf{V}^x$ and $\mathbf{V}^y$ as:

$$
\begin{aligned}
\mathbf{V}^x &= \mathbf{Z}^x\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \gamma_1\mathbf{I})^{-1} \\
\mathbf{V}^y &= \mathbf{Z}^y\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T + \gamma_2\mathbf{I})^{-1}
\end{aligned} \tag{3.15}
$$

After getting $\mathbf{V}^x$ and $\mathbf{V}^y$, when a new text or image comes, we use $\mathbf{V}^x$ or $\mathbf{V}^y$ to transform it into the new space, and compute Euclidean distance in the new space.

Algorithm 1 summarizes the whole process:

Algorithm 1 HLLE Algorithm Description

```
Input: Image dataset $\mathbf{X} \in \mathbb{R}^{d^x*n}$, text dataset $\mathbf{Y} \in \mathbb{R}^{d^y*m}$, heterogeneous constraint matrix $\mathbf{C} \in \mathbb{Z}^{n*m}$
Output: Linear map matrix $\mathbf{V}^x \in \mathbb{R}^{d*d^x}$ and $\mathbf{V}^y \in \mathbb{R}^{d*d^y}$
Compute locally linear reconstruction weights $w_{ij}^x$ and $w_{ij}^y$ by Eq.(3.5);
Compute embedding vectors of all the images and texts by optimizing Eq.(3.12);
Compute map matrix $\mathbf{V}^x$ and $\mathbf{V}^y$ by Eq.(3.15).
```

3.5 Discussion In the *Locally Linear Reconstruction* stage, we need to compute matrix inverse ($(\mathbf{Q}^i + \gamma\mathbf{I})^{-1}$) in Eq.(3.5). Since $(\mathbf{Q}^i + \gamma\mathbf{I})$ is a $k \times k$ matrix, the time complexity is $O(k^3)$. While $k \ll n + m$ in practice, the inverse is not the computationally heaviest step. Similarly, in Eq.(3.15), we need to compute $d^x \times d^x$ and $d^y \times d^y$ matrices inverse, and $d^x, d^y \ll n + m$, thus this step is also not very expensive. In the *Heterogeneous Embedding* stage, we need do an eigenvalue decomposition of the $(m + n) \times (m + n)$ matrix $\mathbf{T}$, whose complexity is $O((m + n)^3)$. This is the computationally heaviest problem in our method. One of our future work is to further reduce the time complexity. However, due to *Learning Linear Map*, all these three stages can be done offline. In offline process, we learn the map function, and in online process, we just need to use the function to map data into new space and retrieve in new space. Therefore, the efficiency of our method is acceptable.

Table 2: Cross-media retrieval on Wikipedia dataset(MAP scores). I → T means use Image as query to retrieve Text. T → I means use Text as query to retrieve Image.

Method	I → T	T → I	Average
RANDOM	0.1179	0.1179	0.1179
CCA	0.1719	0.1904	0.1812
CFA	0.1552	0.1835	0.1694
PLS	0.2622	0.1804	0.2213
CCA+SMN	0.2439	0.1964	0.2202
IMH	0.2220	0.1740	0.1980
JGRHML	0.2768	0.1880	0.2324
LLEHML	**0.2930**	**0.2236**	**0.2583**

4 Experiments

We conduct experiments on two publicly available real-world datasets to verify the effectiveness of our method.

4.1 Datasets Wikipedia dataset[*] [18] is chosen from the Wikipedia's "featured articles". The dataset contains 2866 documents which belong to 10 categories, which are text-image pairs, and it is randomly split into a training set of 2173 documents and a test of 693 documents. In this dataset texts are represented using a histogram of a 10-topic latent Dirichlet allocation (LDA) model and images are represented using a histogram of a 128-codeword SIFT codebook [5]. In the training set, we randomly choose 1000 images and texts to construct constraints. More specifically, in the chosen image and text sets, if the image and text belong to the same category, there is a must-link constraint between them, otherwise, there is a cannot-link constraint.

NUS-WIDE dataset[†] [4] is a web image dataset containing 269648 images downloaded from Flickr. Each image has a short text description. Tagging ground-truth for 81 semantic concepts is provided for evaluation. Since some of the concepts are very scarce, we considered only the 10 most populated ones. Then we randomly choose 4000 images and texts which have at least one of the 10 concepts as its tag. We use 3000 as the training set and 1000 as the test set. In training set, we randomly choose 1000 to construct constraints. If the text and image have at least one same concept, there is a must-link between them, otherwise there is a cannot-link constraint. The texts are represented with 1000-dimension word vector, The images are represented with 500-dimension bag of visual words [15].

[*] http://www.svcl.ucsd.edu/projects/crossmodal

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

4.2 Compared Methods, Experimental Setting and Evaluation Metrics Several different heterogeneous metric learning methods are compared:

- **Random**: Randomly retrieving the results.
- **CCA**: Canonical Correlation Analysis (CCA)[11] is used in [18] to learn two linear transformation matrices which maximize the correlation between two sets of heterogeneous objects.
- **CFA**: Cross-modal Factor Analysis (CFA) [14] also learns two linear transformation matrices. It adopts a criterion of minimizing the Frobenius norm between pairwise data in the new space.
- **PLS**: [22] uses Partial Least Square (PLS) to learn two orthogonal transformation matrices by minimizing the distances between relevant objects and maximizing them between irrelevant objects.
- **CCA+SMN**: CCA+SMN [18] considers not only correlation analysis but also semantic abstraction for different modalities.
- **IMH**: Inter-Media Hashing (IMH) [20] is current state-of-the-art cross-media hash method. It learns two linear hash function to transform texts and images into a unified Hamming space.
- **JGRHML**: Joint Graph Regularized Heterogeneous Metric Learning (JGRHML) [25] learns two transformation which minimizing (maximizing) the distances between the objects with the similar (dissimilar) constraints with joint graph regularization.

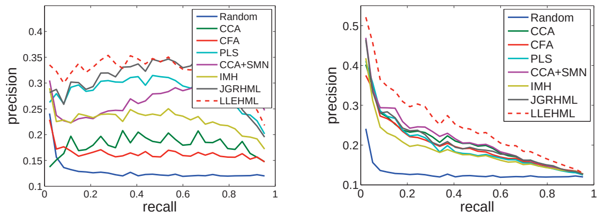
To run all of these compared methods, we use their own heterogeneous constraints introduced in related literatures, and we use the distance suggested in their respect related literatures to retrieval in the new space.

On Wikipedia dataset, we build ground-truth with category information as [25] did. If the query data and the result data belong to the same category, this result is correct. On Nus-Wide dataset, similarly, ground-truth is built based on concept information. If the query data and result data have same concept, the result is correct.

On both datasets, we perform two cross-media retrieval tasks: using texts to retrieve images and using images to retrieve texts. We evaluate the retrieving results with precision-recall (PR) curves and mean average precision (MAP), which are widely used in the image retrieval literature.

Table 3: Cross-media retrieval on NUS-WIDE dataset(MAP scores).

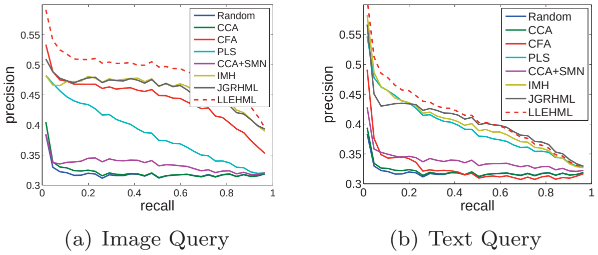
Method	I → T	T → I	Average
RANDOM	0.2456	0.2456	0.2456
CCA	0.2471	0.2466	0.2469
CFA	0.3784	0.2498	0.3141
PLS	0.3334	0.3285	0.3310
CCA+SMN	0.2599	0.2601	0.2600
IMH	0.3632	0.3145	0.3389
JGRHML	0.3658	0.3171	0.3415
LLEHML	**0.4334**	**0.3419**	**0.3877**



(a) Image Query (b) Text Query

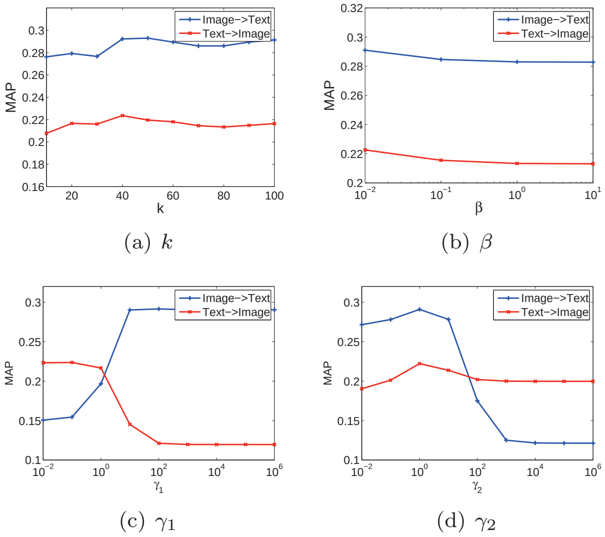
Figure 2: Precision recall curves on Wikipedia dataset

4.3 Experimental Results Table 2 shows the MAP scores of our method LLEHML and compared methods on Wikipedia dataset. LLEHML outperforms all of these compared methods in two query tasks on this dataset. From the table, we can see that CCA+SMN, in addition to using CCA, also uses high level semantic information and it can improve CCA to some extent. CCA, CFA, and IMH use coupled constraints, but IMH also considers homogeneous graphs of texts and images and it can get a better result. Different from CCA and CFA, PLS uses must-link and cannot-link constraints, which is more informative than coupled constraints and works better than CCA and CFA. Although IMH considers more on homogeneous graphs, it does not work better than PLS on this dataset. It shows that, on Wikipedia dataset, heterogeneous constraints play a more important role, i.e. in order to get a better result, we should get better supervised information. JGRHML not only uses must-link and cannot-link constraints, but also considers joint graph regularization, which can make the solution smoother for both media, and this makes it outperform other methods. However, their graph regularization is completely from label information and it does not consider data feature itself, i.e. if the text or image does not have constraint, this method will not use any information from this text or image. Different from JGRHML, LLEHML uses locally linear embedding to deal with homogeneous data, and this method uses all data whether it has constraint or not, i.e. we use more information from the dataset. The results show that the information hidden in the uncon-



(a) Image Query (b) Text Query

Figure 3: Precision recall curves on NUS-WIDE dataset



(a) k (b) β

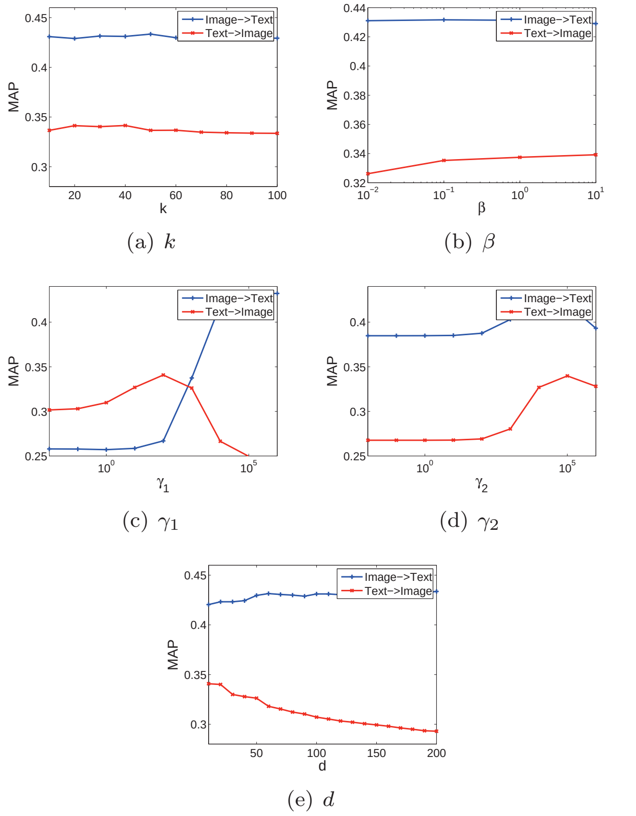
(c) γ_1 (d) γ_2

Figure 4: Study of parameter sensitivity on Wikipedia dataset

straint data is helpful to learn the metric.

Table 3 shows the MAP scores of LLEHML and compared methods on NUS-WIDE dataset. LLEHML also outperforms all compared methods in two query tasks. On this dataset, all methods work better than on Wikipedia dataset. It shows that, cross-media retrieval on this dataset is easier. There is no wonder why IMH can outperform PLS on this dataset. Because retrieval on this dataset is easier, we do not need too much supervised information to get a better result, and the advantage of a more informative heterogeneous constraints is not so significant. CCA does not work well on this dataset, and is just comparable to randomly search. CCA+SMN also does not work well. It shows that, when CCA does not work well, CCA+SMN's improvement may be limited.

Figure 2 and Figure 3 show the precision recall curve of the above methods on wikipedia dataset and NUS-WIDE dataset respectively. It can be seen that LLEHML also attains higher precision at most levels of recall, outperforming those compared methods.



(a) k (b) β

(c) γ_1 (d) γ_2

(e) d

Figure 5: Study of parameter sensitivity on Nus-Wide dataset

4.4 Parameter Sensitivity

We test different parameter settings for LLEHML to see the performance variation. We test k-NN parameter in the range $[10, 100]$, the balance parameter β in the range $[10^{-2}, 10^{1}]$, and test γ_1, γ_2 in the range $[10^{-2}, 10^{6}]$. On wikipedia dataset, text data only have 10 dimensions, therefore the dimension of the embedding space d is between 1 to 10. It is a small range and d is insensitive, thus we fix d to be 5. On NUS-WIDE dataset, we test d in the range $[10, 200]$. Figure 4 and 5 show the results.

5 Conclusion

We have proposed a novel heterogeneous metric learning algorithm LLEHML to measure distances between heterogeneous media data. It first obtains homogeneous local information using locally linear reconstruction, and then learns a heterogeneous embedding which can preserve both the constraint of heterogeneous data and the local information in homogeneous data. To overcome the out-of-sample problem in LLE, two linear functions are learned to approximate the original embedding method.

Several questions remain to be investigated in future work. The first one is scalability issue with large scale datasets. The second one is how to get more informative constraints with less cost. Although must-link and cannot-link are more informative than coupled constraints, in practical application, most often the relevance of two objects are in between.

6 Acknowledgments

This work is supported in part by China National 973 program 2014CB340301 and NSFC grant 61379043. Mingyu Fan is supported in part by the NSFC grants 61203241, and 61473212.

References

[1] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6(6):937–965, 2005.

[2] John Blitzer, Kilian Q Weinberger, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2005.

[3] Michael M. Bronstein, Alexander M. Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, pages 3594–3601, 2010.

[4] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR '09*, pages 48:1–48:9, 2009.

[5] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, page 22, 2004.

[6] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216. ACM, 2007.

[7] Dick de Ridder and Robert P.W. Duin. Locally linear embedding for classification. 2002.

[8] Hayit Greenspan, Jacob Goldberger, and Arnaldo Mayer. Probabilistic space-time video modeling via a piecewise gmm. *IEEE Transactions on PAMI*, 26(3):384–396, 2004.

[9] Yujie He, Wenlin Chen, and Yixin Chen. Kernel density metric learning. In *ICDM*, pages 271–280. IEEE, 2013.

[10] Steven CH Hoi, Wei Liu, and Shih-Fu Chang. Semi-supervised distance metric learning for collaborative image retrieval. In *CVPR*, pages 1–7. IEEE, 2008.

[11] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[12] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[13] Shaishav Kumar and Raghavendra Udupa. Learning hash functions for cross-view similarity search. In *IJCAI*, pages 1360–1365, 2011.

[14] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K Sethi. Multimedia content processing through cross-modal association. In *ACM MM*, pages 604–611. ACM, 2003.

[15] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.

[16] Kyoungup Park, Chunhua Shen, Zhihui Hao, and Junae Kim. Efficiently learning a distance metric for large margin nearest neighbor classification. In *AAAI*, 2011.

[17] Yuxin Peng and Chong-Wah Ngo. Clip-based similarity measure for query-dependent clip retrieval and video summarization. *IEEE Transactions on CSVT*, 16(5):612–627, 2006.

[18] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, pages 251–260. ACM, 2010.

[19] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[20] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *SIGMOD Conference*, pages 785–796, 2013.

[21] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[22] Wei Wu, Jun Xu, and Hang Li. Learning similarity function between objects in heterogeneous spaces.

[23] Eric P Xing, Michael I Jordan, Stuart Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. In *NIPS*, pages 505–512, 2002.

[24] Hongyuan Zha, Xiaofeng He, Chris Ding, Ming Gu, and Horst D Simon. Spectral relaxation for k-means clustering. In *NIPS*, pages 1057–1064, 2001.

[25] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In *AAAI*, 2013.

[26] Hong Zhang and Li Chen. Learning optimal data representation for cross-media retrieval. In *ICIP*, pages 1925–1928. IEEE, 2012.