Adaptively Unified Semi-supervised Learning for Cross-Modal Retrieval

Liang Zhang[1,3], **Bingpeng Ma**[1,2,3*], **Jianfeng He**[1,3], **Guorong Li**[1,3], **Qingming Huang**[1,2,3*], **Qi Tian**[4]
[1] University of Chinese Academy of Sciences, Beijing, 100049, China
[2] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, 100190, China
[3] Key Laboratory of Big Data Mining and Knowledge Management, CAS, Beijing, 100190, China
[4] Department of Computer Science, University of Texas at San Antonio, TX, 78249, USA
{zhangliang14, hejianfeng14}@mails.ucas.ac.cn
{bpma, liguorong, qmhuang}@ucas.ac.cn qitian@cs.utsa.edu

Abstract

Motivated by the fact that both relevancy of class labels and unlabeled data can help to strengthen multi-modal correlation, this paper proposes a novel method for cross-modal retrieval. To make each sample moving to the direction of its relevant label while far away from that of its irrelevant ones, a novel dragging technique is fused into a unified linear regression model. By this way, not only the relation between embedded features and relevant class labels but also the relation between embedded features and irrelevant class labels can be exploited. Moreover, considering that some unlabeled data contain specific semantic information, a weighted regression model is designed to adaptively enlarge their contribution while weaken that of the unlabeled data with non-specific semantic information. Hence, unlabeled data can supply semantic information to enhance discriminant ability of classifier. Finally, we integrate the constraints into a joint minimization formulation and develop an efficient optimization algorithm to learn a discriminative common subspace for different modalities. Experimental results on Wiki, Pascal and NUS-WIDE datasets show that the proposed method outperforms the state-of-the-art methods even when we set 20% samples without class labels.

1 Introduction

As a hot spot of big data era, multimedia data have vastly emerged in search engines and social media. Since multi-modal data may be assigned with the semantic association [Li and Tang, 2017], it is imperative to exploit the correlation among different modalities. Hence, the study of cross-modal retrieval has attracted increasing attention because it can support the similarity computation across different modalities.

Since different modalities lie on different feature spaces, the key idea of cross-modal retrieval is to develop techniques which can realize the calculation of heterogeneous similarity. A lot of works have been proposed to alleviate this problem by learning a common subspace. According to the fact whether the class labels are used or not, classical cross-modal learning methods can be categorized into unsupervised methods, semi-supervised methods and supervised methods.

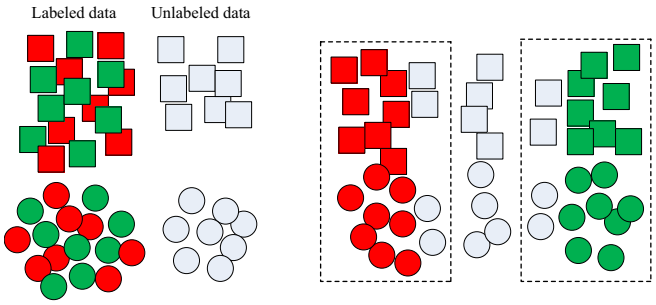
Unsupervised cross-modal methods learn the common subspace by uniting paired samples from two different modalities [Rasiwasia *et al.*, 2010; Andrew *et al.*, 2013]. Canonical correlation analysis (CCA) [Hardoon *et al.*, 2004] learns a common subspace by maximizing the correlation between the projected vectors of two different modalities. Rasiwasia *et al.*adopt CCA to match the heterogeneous samples. Deep CCA [Andrew *et al.*, 2013] learns a set of flexible nonlinear transformations by combining the autoencoder with CCA.

Semi-supervised cross-modal methods use labeled and unlabeled data to model multi-modal correlation. Since manually annotating data is expensive, a large amount of unlabeled data should be handled. Some methods have been proposed to exploit unlabeled data [Peng *et al.*, 2016; Zhai *et al.*, 2014; Li *et al.*, 2015; Gong *et al.*, 2016; Xu *et al.*, 2015]. Most of these methods construct labeled and unlabeled data into a unified graph model and then the label information can be propagated from labeled data to unlabeled data.

Supervised cross-modal methods utilize class information to learn a discriminative subspace. Since labels directly reveal the semantic of multi-modal data, a lot of methods use labels as interlinkage to model the correlations among different modalities [Sharma *et al.*, 2012; Wang *et al.*, 2013; Kang *et al.*, 2015; Rasiwasia *et al.*, 2014; Ranjan *et al.*, 2015]. For example, generalized multiview analysis (GMA) [Sharma *et al.*, 2012] applies the class information to learn a discriminant latent space. Wang *et al.*propose a half quadratic optimization to learn the coupled feature spaces for two modalities. Since multi-modal data may naturally be annotated with multiple labels, Ranjan *et al.*propose multi-label CCA (ml-CCA) by adopting the multi-labeled data to learn a shared subspace. Besides, many deep models have been developed to enhance correlations among the multimedia data using the class information. Multi-modal auto-encoders [Ngiam *et al.*, 2011] and deep boltzmann machines [Srivastava and Salakhutdinov, 2012] are designed to exploit the discriminative information of two modalities by learning the deep-based shared representation.

Though the above methods can achieve relatively good performances, they cannot exploit relation between embedded features and irrelevant class labels. In practice, we are usu-

*Corresponding Author.



(a) Original space (b) Common subspace

Figure 1: Working mechanism of AUSL. Different shapes denote different modalities. Red and green shapes indicate relevant labeled data. All gray shapes denote unlabeled data.

ally faced with abundant multi-labeled data because a single class label cannot sufficiently describe the content of multi-modal data. Some multi-label methods [Ranjan *et al.*, 2015; Wang *et al.*, 2013; Kang *et al.*, 2015] enhance the relation between embedded features and relevant class labels such that each sample is pushed toward the direction of its relevant class label. They neglect the relation between embedded features and irrelevant class labels. In fact, the irrelevance is also indispensable because it can make each sample far away from the directions of its irrelevant class labels.

Classical semi-supervised methods cannot consider the diversity among unlabeled data. In fact, compared with labeled data, some unlabeled data mainly focus on describing the specific semantic content. In this paper, we call these unlabeled data as specific points, like the gray shapes inside the dashed bounding boxes in Fig. 1(b). The specific points can easily research many relevant samples with class labels. Hence, these data can actually help to discriminate samples from the different classes. Conversely, some others usually reflect multiple aspects of semantic content. These unlabeled data can disturb the learned classifier since their content are relevant to multiple classes. We call these data as non-specific points, like the gray shapes outside the dashed bounding boxes in Fig. 1(b).

Besides, existing semi-supervised methods mainly focus on constructing graph models using labeled and unlabeled data from all modalities [Zhai *et al.*, 2014; Peng *et al.*, 2016; Li *et al.*, 2015]. When a new testing sample comes, they need to introduce this new datum into the existing dataset, and then reconstruct the graph model. Therefore, such strategy is very inefficient in processing the out-of-sample problem.

To overcome the aforementioned problems, this paper proposes a novel semi-supervised framework, named **A**daptively **U**nified **S**emi-supervised **L**earning (AUSL) for cross-modal retrieval. To utilize both the relevance and irrelevance between embedded features and class labels, a dragging technique is fused into a unified linear regression, by which the embedded feature of each sample will be close to its relevant class labels while far away from its irrelevant class labels. To enlarge the contribution of specific points while eliminate the interference of non-specific points, we design a weighted regression model to adaptively exploit the discriminative information from unlabeled data in learning the classifier. Meanwhile, our linear regression model can also deal with the out-of-sample problem. Finally, we integrate the different constraints into a joint minimization formulation and design an efficient optimization algorithm to learn a discriminative common subspace for different modalities. Experimental results on three cross-modal datasets show that AUSL outperforms the state-of-the-art methods even when AUSL uses part of unlabeled data but the others apply all labeled data.

In Fig. 1, we provide the working mechanism of the proposed AUSL. Fig. 1(a) shows the data distribution of the original features from two modalities. Each modality consists of labeled data from two classes (red and green colors with the same shapes) as well as some unlabeled data (shapes with gray color). Fig. 1(b) represents the learned common subspace. After learning, we hope that different shapes with same color are gathered together, while same shapes with different colors are far away from each other. Meanwhile, the specific points are correctly classified into their relevant classes, while non-specific points are located between the boundary of the two classes.

2 The Proposed Method

In this section, we first present the proposed AUSL to model the correlations among the multi-modal data. Then we introduce an iterative algorithm to optimize the objective function.

2.1 Preliminaries

Assume we have m modalities $\{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_m\}$, where $\mathbf{X}_r$ denotes the rth modality. The rth modality $\mathbf{X}_r$ consists of n_l labeled data $\mathbf{X}_{rl} = [\mathbf{x}_{r1}, \mathbf{x}_{r2}, ..., \mathbf{x}_{rn_l}]$ and n_u unlabeled data $\mathbf{X}_{ru} = [\mathbf{x}_{r(n_l+1)}, ..., \mathbf{x}_{rn}]$, where $\mathbf{x}_{ri} \in \mathbb{R}^{d_r \times 1}$ denotes the ith sample of the rth modality. Sample $\mathbf{x}_{ri}$ is assigned with a c-dimensional binary-valued vector $\mathbf{y}_i \in \mathbb{R}^{c\times 1}$, where c is the class number. If $\mathbf{x}_{ri}$ is classified into the kth class, y_{ik} is set to 1, otherwise 0. The class indicator matrix for the labeled data is constructed as $\mathbf{Y}_l = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{n_l}]^T$, which is applicable for all modalities. For unlabeled data, we learn the class probability matrices $\mathbf{P} = \{\mathbf{P}_1, \cdots, \mathbf{P}_m\}$ for multi-modal data, where $\mathbf{P}_r$ is learned for the rth modality.

2.2 Objective Function

Obviously, different modalities lie on different feature spaces, which lead to difficulty in measuring the similarity of two heterogeneous samples. Therefore, we focus on learning multiple projection matrices $\mathbf{U} = \{\mathbf{U}_1, \cdots, \mathbf{U}_m\}$ to transform the different features into a common subspace, where $\mathbf{U}_r \in \mathbb{R}^{d_r \times c}$ is the transformation for the rth modality.

Least square regression has achieved great success in learning transformations [Wang *et al.*, 2014; Kang *et al.*, 2015]. Taking the rth modality as example, traditional least square regression obtains the transformation $\mathbf{U}_r \in \mathbb{R}^{d_r \times c}$ and the bias $\mathbf{b}_r \in \mathbb{R}^{c\times 1}$ by solving the following formulation:

$$\min_{\mathbf{U}_r, \mathbf{b}_r} \sum_{i=1}^{n} \|\mathbf{U}_r^T \mathbf{x}_{ri} + \mathbf{b}_r - \mathbf{y}_i\|_F^2 \qquad (1)$$

It is interesting to note that $\mathbf{b}_r$ can be merged into $\mathbf{U}_r$ when we add the constant value 1 as an additional dimension for

each data point. Then we formulate the mapping as a linear transformation from the feature space to the label space, and optimize the labeling approximation error between the given data $\mathbf{x}_{ri}$ and the label vector $\mathbf{y}_i$:

$$\min_{\mathbf{U}_r,\mathbf{b}_r} \sum_{i=1}^{n} \|\mathbf{U}_r^T\mathbf{x}_{ri} - \mathbf{y}_i\|_F^2 \tag{2}$$

Henceforth, $\mathbf{U}_r$ and $\mathbf{x}_{ri}$ are augmented as $\mathbf{U}_r \in \mathbb{R}^{(d_r+1)\times c}$ and $\mathbf{x}_{ri} \in \mathbb{R}^{(d_r+1)\times 1}$.

However, the traditional regression model is suitable for handling single modality. To model the correlations among the multi-modal data, we extend it and propose an adaptively unified semi-supervised learning framework:

$$\min_{\mathbf{U},\mathbf{W},\mathbf{P}} L(\mathbf{U},\mathbf{W}) + \varphi(\mathbf{U},\mathbf{P}) + \lambda\Omega(\mathbf{U}) \tag{3}$$

where $L(\mathbf{U},\mathbf{W})$ is a linear regression with dragging to make each sample close to its true label and far away from its irrelevant labels after transforming. $\varphi(\mathbf{U},\mathbf{P})$ is a weighted regression which can effectively exploit unlabeled data with specific semantic information and weaken unlabeled data with non-specific semantic information. $\Omega(\mathbf{U})$ is a regularization to select discriminative features from the original features and avoid overfitting problem. $\lambda > 0$ is the tradeoff parameter.

Labeling loss: This constraint item is defined to embed the class label information into a unified linear regression model with dragging to force the regression targets of different classes moving along with the opposite directions.

Considering that class labels directly reveal the semantic information of multimedia data, it would be feasible using the label space as the linkage of different feature spaces. Hence, we embed the class label information into a unified linear regression model to link different modalities. Furthermore, the ε-dragging technique [Xiang *et al.*, 2012] is fused into the proposed model to enhance the relevance and irrelevance between embedded features and class labels:

$$L(\mathbf{U},\mathbf{W}) = \frac{1}{2}\sum_{r=1}^{m}\sum_{i=1}^{n_l} \|\mathbf{U}_r^T\mathbf{x}_{ri} - \mathbf{y}_i - \mathbf{z}_i \cdot \mathbf{w}_i\|_F^2 \tag{4}$$

where $\cdot$ denotes the Hadamard product operator between vectors, $\mathbf{w}_i \in \mathbb{R}^{c\times 1}$ is the ϵ-dragging factor, which is used to adaptively adjust the relevant and irrelevant degree between embedded features and class labels. $\mathbf{z}_i \in \mathbb{R}^{c\times 1}$ is defined as:

$$z_{ji} = \begin{cases} +1 & \text{if } y_i = j \\ -|\hat{\mathbf{y}}_i|/|\breve{\mathbf{y}}_i| & \text{otherwise} \end{cases} \tag{5}$$

where $|\hat{\mathbf{y}}_i|$ and $|\breve{\mathbf{y}}_i|$ represent the number of irrelevant and relevant labels of $\mathbf{x}_{ri}$. $|\hat{\mathbf{y}}_i|/|\breve{\mathbf{y}}_i|$ is used to adjust the imbalance between the number of relevant and irrelevant labels. If the number of irrelevant labels is greater than that of relevant labels, we give more penalty to the irrelevant labels so the embedded features can easily establish correlation with relevant labels. Specially, for the ith sample of arbitrary modalities, their $\mathbf{w}_i$ shared the same value. So does $\mathbf{z}_i$.

Unlabeling loss: We define this item to adaptively handle unlabeled data from multiple modalities such that specific points can contribute more than non-specific points in learning the classifier.

In practice, some unlabeled data actually contain discriminative information for they focus on describing specific semantic content, while some others usually involve multiple directions of description such that they will bring ambiguity in learning the classifier. Hence, we should pay more attention to the specific points and ignore the non-specific points. Based on this, we design a unified weighted regression model to learn a adaptive probability matrix for unlabeled data:

$$\varphi(\mathbf{U},\mathbf{P}) = \sum_{r=1}^{m}\sum_{i=n_l+1}^{n}\sum_{k=1}^{c} p_{rik}^s \|\mathbf{U}_r^T\mathbf{x}_{ri} - \mathbf{t}_k\|_F^2 \tag{6}$$

where p_{rik} is the probability of the ith unlabeled data of the rth modality belonging to the kth class, which should be a value between $[0, 1]$. $\mathbf{t}_k \in \mathbb{R}^{c\times 1}$ is the class indicator vector for the kth class, where the kth entry of $\mathbf{t}_j$ is set to 1, and the remaining entries are all zeros. When p_{rik} is a large value near to 1, it implies that $\mathbf{x}_{ri}$ is likely to be a specific point and contains discriminative information. $s \in [1, \infty]$ is the power exponent to control the effect of unlabel data. When r increases, p_{rik}^s decreases for specific points and non-specific points. Moreover, the weights of non-specific points are weakened much faster than those of specific points. Besides, compared with constructing graph model [Zhai *et al.*, 2014; Peng *et al.*, 2016; Li *et al.*, 2015], the computational complexity is greatly reduced.

Regularization: This constraint item aims to avoid overfitting problem induced by the sparse features $\mathbf{X}_r$ and select the discriminative features from the original feature. We utilize $\ell_{2,1}$-norm to formulate this item as follows:

$$\Omega(\mathbf{U}) = \sum_{r=1}^{m} \|\mathbf{U}_r\|_{2,1} \tag{7}$$

2.3 Optimization

From the above description, we know that the objective function is difficult to be directly solved since it is non-smooth. Hence, we design an efficient and iterative algorithm to optimize transformations $\mathbf{U}$, shared dragging matrix $\mathbf{W}$ and the class probability matrices $\mathbf{P}$.

Update U: We solve the transformations $\mathbf{U}$ by fixing $\mathbf{P}$ and $\mathbf{W}$. The second term in Eqn. (3) sums over n_u unlabeled points and c classes. If we directly solve the derivative with respect to $\mathbf{U}_r$, it would be slow since the resulting formula would need to iterate over n_u and c.

To simplify the optimization, we rewrite Eqn. (3) into the compact matrix representation in the following way:

$$\begin{aligned}\min_{\mathbf{U}_r} &\mathrm{Tr}((\mathbf{X}_{rl}^T\mathbf{U}_r - \mathbf{Y}_l - \mathbf{Z}\cdot\mathbf{W})^T(\mathbf{X}_{rl}^T\mathbf{U}_r - \mathbf{Y}_l - \mathbf{Z}\cdot\mathbf{W})) \\ &+\mathrm{Tr}(\mathbf{U}_r^T\mathbf{X}_{ru}\mathbf{S}_r\mathbf{X}_{ru}^T\mathbf{U}_r - 2\mathbf{P}_r\mathbf{U}_r^T\mathbf{X}_{ru}) + \lambda\mathrm{Tr}(\mathbf{U}_r^T\mathbf{Q}_r\mathbf{U}_r)\end{aligned} \tag{8}$$

where $\mathbf{P}_r \in \mathbb{R}^{n_u\times c}$ (each element of $\mathbf{P}_r$ is defined as p_{rik}^s), $\mathbf{S}_r \in R^{n_u\times n_u}$ is a diagonal matrix with $S_r^{ii} = \sum_{k=1}^{c} P_r^{ik}$. Based on the definition of $\ell_{2,1}$-norm, $\mathbf{Q}_r$ is a diagonal matrix

with $Q_r^{ii} = \frac{1}{2\|\mathbf{U}_r^i\|_2}, i = 1, ..., d_r$. Each row of $\mathbf{Z} \in \mathbb{R}^{n_l \times c}$ and $\mathbf{W} \in \mathbb{R}^{n_l \times c}$ is defined in Eqn. (4).

Setting the derivative of Eqn. (8) with respect to $\mathbf{U}_r$ to zero, we obtain the solution of $\mathbf{U}_r$ as follows:

$$\begin{aligned}\mathbf{U}_r =&(\mathbf{X}_{rl}\mathbf{X}_{rl}^T + \mathbf{X}_{ru}\mathbf{S}_r\mathbf{X}_{ru}^T + \lambda\mathbf{Q}_r)^{-1} \\ &(\mathbf{X}_{rl}(\mathbf{Y}_l - \mathbf{Z}\cdot\mathbf{W}) + \mathbf{X}_{ru}\mathbf{P}_r)\end{aligned} \tag{9}$$

Update $\mathbf{W}$: When $\mathbf{U}$ and $\mathbf{P}$ are fixed, we solve the shared dragging matrix $\mathbf{W}$. Let $\mathbf{T}_r = \mathbf{X}_{rl}^T\mathbf{U}_r - \mathbf{Y}_l$, the solution of $\mathbf{W}$ can be obtained from the following objective function:

$$\min_{\mathbf{W}} \frac{1}{2}\sum_{r=1}^{m}\|\mathbf{T}_r - \mathbf{Z}\cdot\mathbf{W}\|_F^2 \quad s.t.\ \mathbf{W} \geq 0 \tag{10}$$

Based on the Frobenius norm of matrix, Eqn. (10) can be separated into $n_l \times c$ subproblems, which have the unified optimization formulation:

$$\min_{\mathbf{W}^{ij}}(T_r^{ij} - Z^{ij}W^{ij})^2 \quad s.t.\ W^{ij} \geq 0 \tag{11}$$

where T_r^{ij} denotes the element locating in the i-th row and j-th column, so do W^{ij} and Z^{ij}. Considering the nonnegative constraint on Z^{ij}, the optimal solution of W^{ij} is:

$$W^{ij} = \max\{\sum_{r=1}^{m} Z^{ij}T_r^{ij}, 0\} \tag{12}$$

Therefore, the final optimal solution to Eqn. (12) is:

$$\mathbf{W} = \max\{\sum_{r=1}^{m}\mathbf{Z}\cdot\mathbf{T}_r, 0\} \tag{13}$$

Update $\mathbf{P}$: After solving $\mathbf{U}$ and $\mathbf{W}$, we turn to the solution of $\mathbf{P}$. Since the class probability matrices of different modalities can be solved independently, we give the derivation of $\mathbf{P}_r$ for the ith modality as follows:

$$\begin{aligned}\min_{\mathbf{P}_r} &\sum_{i=n_l+1}^{n}\sum_{k=1}^{c} p_{rik}^s\|\mathbf{U}_r^T\mathbf{x}_{ri} - \mathbf{t}_k\|_F^2 \\ &s.t.\ \forall i,\ p_{ik} \in [0,1],\ \sum_{i=1}^{c} p_{ik} = 1\end{aligned} \tag{14}$$

Denote $q_{rik} = \|\mathbf{U}_r x_{ri} - t_k\|_F^2$, the objective function in Eqn. (3) can be rewritten as:

$$\min_{\mathbf{P}_r}\sum_{i=n_l+1}^{n}\sum_{k=1}^{c} p_{rik}^s q_{rik} \quad s.t.\ p_{rik} \in [0,1],\ \sum_{k=1}^{c} p_{rik} = 1 \tag{15}$$

Eqn. (15) can be simplified to n_u subproblems, and the sub-objective function of each subproblem is defined as:

$$\min_{p_{ri}}\sum_{k=1}^{c} p_{rik}^s q_{rik} \tag{16}$$

Concretely, we should consider two cases for solving $\mathbf{P}_r$:

Case1: When $s = 1$, the optimal solution is:

$$\begin{cases} p_{rik} = 1, & \text{if } k = k^* \\ p_{rik} = 0, & \text{if } k \neq k^* \end{cases} \tag{17}$$

where $k^* = \arg\min_k q_{rik}$.

Case 2: When $s > 1$, we apply the *Lagrangian* function to solve the objective function:

$$\sum_{k=1}^{c} p_{rik}^s q_{rik} - \beta(\sum_{k=1}^{c} p_{rik} - 1) \tag{18}$$

where β is the Lagrangian multiplier. Differentiating Eqn. (18) with respect to p_{rik} and setting it to zero, we get:

$$p_{rik} = (\frac{\beta}{sq_{rik}})^{\frac{1}{s-1}} \tag{19}$$

Substituting Eqn. (19) into the constraint $\sum_{k=1}^{c} p_{rik} = 1$, we obtain the closed form solution of p_{rik} as follows:

$$p_{rik} = (\frac{1}{q_{rik}})^{\frac{1}{s-1}} / \sum_{k=1}^{c}(\frac{1}{q_{rik}})^{\frac{1}{s-1}} \tag{20}$$

To search an optimal solution, we alternatively optimize $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{P}$. To the best of our knowledge, it is novel to make three variables depending on each other for exploiting the multi-modal correlation. The convergence criterion of AUSL is that the change between two consecutive iterations is sufficiently small (0.001) or the maximal number (20) of iterations is reached. These convergence values are determined by experimental observation.

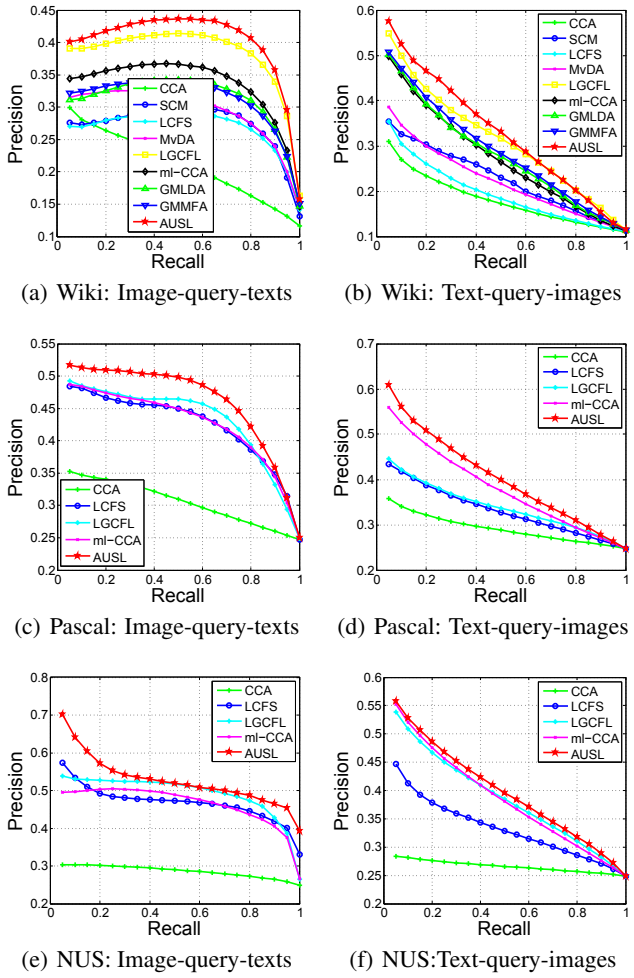
In each iteration, the proposed method achieves the minimum after updating $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{P}$. Hence, the value of objective function will decrease after each update. It is easy to understand that our method will converge because our objective function is constrained no less than zero. The major computational complexity of AUSL is to update $\mathbf{U}$. For the rth modality, the matrix inverse has the complexity of $O((d_r+1)^3)$, and the matrix multiplication has $O(n \times (d_r+1)^2 + c \times (d_r+1)^2 + n_u^2 \times (d_r+1) + c \times n_u \times (d_r+1))$. Considering that in practice c is much smaller than d_r and n_l. Hence, the total cost on calculating m modalities is $O(\sum_{r=1}^{m} N \times ((d_r+1)^3 + n \times (d_r+1)^2 + n_u^2 \times (d_r+1)))$, where N is the number of iterations.

3 Experimental Results

In this section, we present extensive experiments to demonstrate the effectiveness of the proposed method for text-image retrieval, *i.e.*, image-query-texts and text-query-images.

3.1 Datasets

Wiki dataset is collected from Wikipedia feature articles [Rasiwasia *et al.*, 2010]. It contains 2,866 image-text pairs belonging to 10 semantic classes, and each pair belongs to a unique class. Our image features are represented by 4,096 dimensional output from the fc7 layer of CNN [Jia *et al.*, 2014]. For text features, we first adopt word2vec model to learn the 100 dimensional skip-gram word vectors [Mikolov *et al.*, 2013]. Then we calculate a mean vector of the word



(a) Wiki: Image-query-texts (b) Wiki: Text-query-images

(c) Pascal: Image-query-texts (d) Pascal: Text-query-images

(e) NUS: Image-query-texts (f) NUS:Text-query-images

Figure 2: Precision-recall curves on the three datasets.

vectors of the words appearing in each text document. On this dataset, we randomly select 2,000 pairs of the data for training and 866 pairs for testing.

Pascal dataset consists of 5,011/4,952(training/testing) images-tag pairs [Everingham *et al.*, 2010]. All pairs belong to one or more of 20 semantic classes. We use the publicly available 512-dimensional GIST features for images. For texts, we use the 399-dimensional word frequency features. We use the original training-test split and remove some pairs since their text features are all zeros. Finally, 5,000 pairs are used for training and 4,919 pairs for testing.

NUS-WIDE dataset consists of 40,834/27,159 (training/testing) image-tag pairs, which are pruned from the training-test split of the NUS dataset [Chua *et al.*, 2009] by keeping the pairs belonging to one or more of the 10 largest classes. Each text is represented by an 1,000-dimensional word frequency vector based tag features, and each image is represented as an 500-dimensional SIFT feature.

3.2 Experimental Settings

We compare AUSL with several related methods, including CCA & SCM [Rasiwasia *et al.*, 2010], GMLDA & GMMFA [Sharma *et al.*, 2012], LCFS [Wang *et al.*, 2013], MvDA [Kan *et al.*, 2016], LGCFL [Kang *et al.*, 2015] and ml-CCA [Ranjan *et al.*, 2015]. For fairness, we ensure that the total number of training samples is equal for all methods. The training set of the compared methods consists of all labeled samples. But for AUSL, 20 percent samples in the training set are set to the unlabeled data. At the testing stage, we adopt the cosine distance to measure the similarity of features.

Table 1: MAP score comparison of text-image retrieval on three benchmark datasets. Note that AUSL sets 20% labeled samples from training set as unlabeled ones.

Method	Text query	Image query	Average	Dadaset
CCA	0.1872	0.2160	0.2016	Wiki
SCM	0.2336	0.2759	0.2548	
LCFS	0.2043	0.2711	0.2377	
MvDA	0.2319	0.2971	0.2645	
LGCFL	0.3160	0.3775	0.3467	
ml-CCA	0.2873	0.3527	0.3120	
GMLDA	0.2885	0.3159	0.3022	
GMMFA	0.2964	0.3155	0.3060	
AUSL	**0.3321**	**0.3965**	**0.3643**	
CCA	0.2945	0.3073	0.3009	Pascal
LCFS	0.3355	0.4278	0.3816	
LGCFL	0.3440	0.4362	0.3901	
ml-CCA	0.3885	0.4303	0.4094	
AUSL	**0.4653**	**0.4033**	**0.4343**	
CCA	0.2667	0.2869	0.2768	NUS
LCFS	0.3363	0.4742	0.4053	
LGCFL	0.3907	0.4972	0.4440	
ml-CCA	0.3908	0.4689	0.4299	
AUSL	**0.4128**	**0.5690**	**0.4909**	

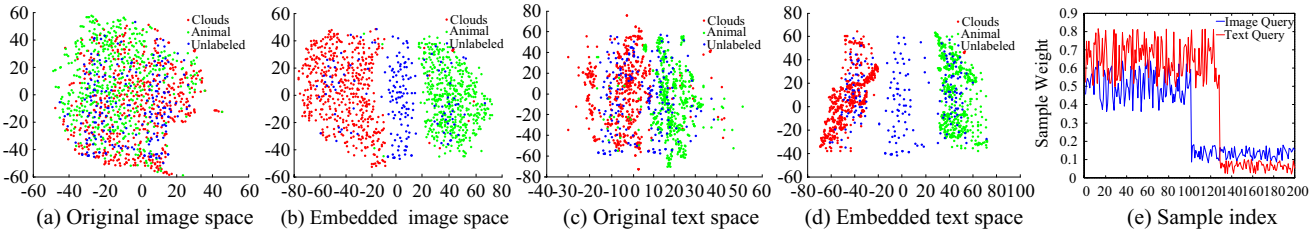
The mean average precision [Rasiwasia *et al.*, 2010] is used to evaluate performance. On Wiki dataset, we define that a retrieved sample is relevant to a query if they belong to the same semantic class. On Pascal and NUS-WIDE datasets, a retrieved sample is relevant if it shares at least one concept with a query. We also display the precision-recall curve [Rasiwasia *et al.*, 2010] for all the methods.

Considering that CCA, SCM, GMLDA, GMMFA and MvDA focus on learning the common subspace, principal component analysis is performed on the original features to remove redundant features, and 95% information energy is preserved. For all methods, parameters are set by 5-fold cross validation on the training set. After cross validation, the parameters s and λ of AUSL are set to 2 and 0.1 in all the experiments. The dimension of the common subspace is set to 10, 20 and 10 for Wiki, Pascal and NUS-WIDE, respectively.

3.3 Cross-Modal Retrieval

The MAP scores of all the methods are shown in Table 1. On Pascal and NUS-WIDE datasets, AUSL is compared with CCA, LCFS, ml-CCA and LGCFL because the other methods cannot handle the multi-labeled data. From Table 1, we can draw the following conclusions:

First, AUSL outperforms the compared methods which use all labeled training data. Since the training set of AUSL includes 20% unlabeled data, this improvement is more ap-



(a) Original image space (b) Embedded image space (c) Original text space (d) Embedded text space (e) Sample index

Figure 3: Demonstration on handling the unlabeled data: (a)-(d) denote two-dimensional t-SNE feature visualization on the NUS-WIDE dataset; (e) shows the probability weight of each unlabeled sample learned by AUSL.

plicable in practice. The reason is that the proposed model effectively discriminates the specific points and the non-specific points. Then it maximally exploits the discriminative information and eliminates the ambiguous semantic information from the unlabeled data. Hence, both labeled and unlabeled data can help to explore the correlations among different modalities in our framework.

Second, on Pascal dataset, the average MAP score of AUSL is 0.4343, which is about 6.08% higher than the second best result from ml-CCA. ml-CCA adopts the multi-label data to exploit the semantic correlation but they cannot enhance the relation between embedded features and irrelevant class labels. This phenomenon validates that the irrelevance can also help to enhance multi-modal correlation.

Third, AUSL achieves the best average MAP of 0.3643 and 0.4909 on Wiki and NUS-WIDE, respectively. These results are higher than the second best results from LGCFL (0.3467 and 0.4440). Since LGCFL has used the label space to link the image space and text space, this improvement of AUSL depends on exploiting the relevance and irrelevance between embedded features and class labels.

Finally, the precision-recall curves of the image-query and text-query are plotted in Fig. 2. From this figure, we observe that with the same recall rate, AUSL obtains the higher precision than all compared methods on three datasets.

3.4 Demonstration on Exploiting Unlabeled Data

We also show the working mechanism of AUSL on exploiting unlabeled data from different modalities. To demonstrate this, we construct a dataset using some data from the NUS-WIDE dataset. The constructed dataset consists of 500 paired samples from the 'clouds' class, 500 paired samples from the 'animal' class and 200 paired samples without class labels. Note that, to directly reflect discriminative ability of the unlabeled data, they are composed of 150 paired samples from the 'clouds' class and 'animal' class while the rest 50 paired samples come from the other classes for each modality, *i.e.*, the unlabeled data of each modality contains 150 specific semantic points and 50 non-specific points.

In Fig. 3, we adopt the t-SNE [Maaten and Hinton, 2008] to project original features and embedded features into a two-dimensional visualization space. Fig. 3(a)&(c) illustrate the two-dimensional distribution of original features showing that samples from different classes are mixed. We embed the original features into low-dimensional feature spaces by using the learned transformations. In Fig. 3(b)&(d), the embedded features are displayed in a two-dimensional coordinate plane by t-SNE. From it, we conclude that AUSL can separate the labeled samples from different classes and most of the specific points are also classified into their relevant categories. Besides, the distance between different classes of text query is larger than that of image query. We note that AUSL directly optimizes the labeling error between given data and class labels. Since class labels apply more directly on textual features than image features, this phenomenon is reasonable.

Fig. 3(e) shows the probability weight of each unlabeled sample from two modalities. The weight of the ith sample is defined as $\sum_{k=1}^{c} p_{rik}^{s}$. If the weight of a sample is large, it will contribute more to the learned classifier. Experimental results show that image query and text query correctly recover 99 and 128 specific points, respectively. We observe that the weights of those specific points are much larger than the others. In conclusion, our model can adaptively control the weight for each sample, and effectively exploit the discriminative information from unlabeled data.

4 Conclusion

In this paper, we propose AUSL for cross-modal retrieval. We first propose a unified linear regression model with dragging factor to maximally differentiate samples from various categories. Then, we design a weighted regression to adaptively exploit the unlabeled data with specific semantic information and weaken the unlabeled data with non-specific semantic information. Finally, we combine two constraints with $\ell_{2,1}$-norm based regularization and design an efficient optimization algorithm, by which the subspace learning, feature selection and label prediction can be simultaneously realized. Extensive experiments demonstrate that the proposed method outperforms the state-of-the-art methods on the three cross-modal datasets.

Acknowledgements

This work was supported in part by National Basic Research Program of China (973 Program): 2015CB351800, in part by National Natural Science Foundation of China: 61572465, 61332016, 61429201, 61620106009 and U1636214, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013. This work was also supported in part to Dr. Qi Tian by ARO grant W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar.

References

[Andrew *et al.*, 2013] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[Chua *et al.*, 2009] 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 *Proceedings of the 8th ACM International Conference on Image and Video Retrieval*, 2009.

[Everingham *et al.*, 2010] Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[Gong *et al.*, 2016] Chen Gong, Dacheng Tao, Stephen J. Maybank, Wei Liu, Guoliang Kang, and Jie Yang. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Trans. Image Processing*, 25(7):3249–3260, 2016.

[Hardoon *et al.*, 2004] David R. Hardoon, Sándor Szedmák, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, 2014.

[Kan *et al.*, 2016] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):188–194, 2016.

[Kang *et al.*, 2015] Cuicui Kang, Shiming Xiang, Shengcai Liao, Changsheng Xu, and Chunhong Pan. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE Trans. Multimedia*, 17(3):370–381, 2015.

[Li and Tang, 2017] Zechao Li and Jinhui Tang. Weakly supervised deep matrix factorization for social image understanding. *IEEE Trans. Image Processing*, 26(1):276–288, 2017.

[Li *et al.*, 2015] Zechao Li, Jing Liu, Jinhui Tang, and Hanqing Lu. Robust structured subspace learning for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(10):2085–2098, 2015.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

[Ngiam *et al.*, 2011] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[Peng *et al.*, 2016] Yuxin Peng, Xiaohua Zhai, Yunzhen Zhao, and Xin Huang. Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Trans. Circuits Syst. Video Techn.*, 26(3):583–596, 2016.

[Ranjan *et al.*, 2015] Viresh Ranjan, Nikhil Rasiwasia, and C. V. Jawahar. Multi-label cross-modal retrieval. In *IEEE International Conference on Computer Vision*, 2015.

[Rasiwasia *et al.*, 2010] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th International Conference on Multimedia 2010*, 2010.

[Rasiwasia *et al.*, 2014] Nikhil Rasiwasia, Dhruv Mahajan, Vijay Mahadevan, and Gaurav Aggarwal. Cluster canonical correlation analysis. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 2014.

[Sharma *et al.*, 2012] Abhishek Sharma, Abhishek Kumar, Hal Daumé III, and David W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[Srivastava and Salakhutdinov, 2012] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*, 2012.

[Wang *et al.*, 2013] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. Learning coupled feature spaces for cross-modal matching. In *IEEE International Conference on Computer Vision*, 2013.

[Wang *et al.*, 2014] De Wang, Feiping Nie, and Heng Huang. Large-scale adaptive semi-supervised learning via unified inductive and transductive model. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

[Xiang *et al.*, 2012] Shiming Xiang, Feiping Nie, Gaofeng Meng, Chunhong Pan, and Changshui Zhang. Discriminative least squares regression for multiclass classification and feature selection. *IEEE Trans. Neural Netw. Learning Syst.*, 23(11):1738–1754, 2012.

[Xu *et al.*, 2015] Xing Xu, Yang Yang, Atsushi Shimada, Rin-ichiro Taniguchi, and Li He. Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, 2015.

[Zhai *et al.*, 2014] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Trans. Circuits Syst. Video Techn.*, 24(6):965–978, 2014.