Scalable multi-label canonical correlation analysis for cross-modal retrieval

Xin Shu[a,b], Guoying Zhao[b]

[a] *College of Information Science and Technology*
Nanjing Agricultural University, Nanjing, China
[b] *Center for Machine Vision and Signal Analysis*
University of Oulu, Finland

Abstract

Multi-label canonical correlation analysis (ml-CCA) has been developed for cross-modal retrieval. However, the computation of ml-CCA involves dense matrices eigendecomposition, which can be computationally expensive. In addition, ml-CCA only takes semantic correlation into account which ignores the cross-modal feature correlation. In this paper, we propose a novel framework to simultaneously integrate the semantic correlation and feature correlation for cross-modal retrieval. By using the semantic transformation, we show that our model can avoid computing the covariance matrix explicitly which is a huge save of computational cost. Further analysis shows that our proposed method can be solved via singular value decomposition which has linear time complexity. Experimental results on three multi-label datasets have demonstrated the accuracy and efficiency of our proposed method.

Keywords: canonical correlation analysis, semantic transformation, cross-modal retrieval, singular value decomposition

1. Introduction

With the huge surge in multimodal data such as image, text, video, cross modal retrieval has recently attained considerable attention [1, 2]. Cross-modal retrieval aims to take one type of data as the query and return the relevant data in other modalities. However, cross-modal retrieval is a challenging task since different modalities have inconsistent representations. One popular solution to deal with this problem is to learn a common space where multimodal data can be projected into such common space and compare their similarity [3]. Canonical correlation analysis(CCA) [4] is one of the most popular unsupervised space learning methods to achieve this goal. C-CA aims to find the common space by maximizing the correlation between the projections of the two modalities. It has shown promising performance for cross-media retrieval [5, 6, 7].

It should be aware that existing algorithms are designed for single-label dataset. However, many large scale datasets such as Imagenet [8] and the MIRFlickr [9] are labeled with multiple labels. Therefore, it is important to design cross-modal algorithms that naturally take multiple labels into account. Recently, a multi-label canonical correlation analysis (ml-CCA) has been developed in [10]. ml-CCA needs to compute and store the similarity matrix between the corresponding multi-label vectors, which requires quadratic time complexity. In addition, solving the eigen-value problem of ml-CCA requires cubic time complexity even with the fast implementation proposed in [10].

In this paper, we propose a scalable multi-label canonical correlation analysis (sml-CCA) for cross-modal retrieval. Our method build upon ml-CCA.

Unlike ml-CCA, we incorporate the feature correlation in the formulation of ml-CCA to boost the performance. In addition, we introduce a semantic transformation to compute the label similarity matrix which avoids quadratic time complexity. Furthermore, we propose an efficient algorithm for solving the optimization problem. In summary, the main contributions of this paper are outlined as follows:

- We develop a novel framework that can integrate the feature correlation and semantic correlation to boost retrieval performance.
- A semantic transformation is developed to effectively approximate the label similarity matrix without explicitly computing the pairwise semantic similarity matrix. Hence, the $\mathcal{O}(n^2)$ computation cost and storage cost can be avoided, which makes our proposed method suitable for large-scale applications with quick response.
- The optimal solution of the new formulation can be efficiently solved by singular value decomposition which has linear time complexity in terms of the number of samples.
- Experimental results on several multi-label datasets have demonstrated the effectiveness of our proposed method.

2. Related work

Due to the ever-increasing amount of multimedia data on the web, cross-modal retrieval has attained considerable attention in relevant research areas.

One commonly used technique for cross-modal retrieval is learning a common space for different modalities. Canonical correlation analysis (CCA) [4]

is the main technique for learning a couple of mappings to maximize the correlations between two variables. Rasiwasia et al. [5] integrate labels with CCA to perform semantic matching. Gong et.al [11] proposed a multi-view embedding to learn a common space with images, tags and the corresponding semantics. Generalized multiview analysis (GMA), which is a supervised extension of CCA, has been developed in [12]. Cao et al. [13] developed a unified framework with graph embedding for multi-view embedding. Yuan et al. [14] developed a fractional-order embedding canonical correlation analysis (FECCA) for multi-modal data feature extraction. Similar to CCA, cross-modal factor analysis [15] has been proposed to process multimedia information. Besides, graph based cross-modal retrieval methods have also been investigated in the literature. A joint graph regularized heterogeneous metric learning (JGRHML) algorithm has been developed in [16] to integrate the structure of different media. Zhai et al. [17] proposed a joint learning framework to explore the correlation and semantic information with sparse and semisupervised regularization. Peng et al. [18] further proposed a unified patch graph regularization based semi-supervised cross-media feature learning method. Zhang et al. [19] proposed the generalized semi-supervised structured subspace learning method for cross-modal retrieval. Specifically, they use a label graph constraint to ensure the intrinsic geometric structures of different modalities consistent with the label space. Dictionary learning has also been introduced for cross-modal retrieval. Zhuang et al. [20] proposed a supervised coupled dictionary learning with group structure for cross-modal retrieval. A cross-modality submodular dictionary learning which incorporates the maximum mean discrepancy has been proposed in [21]. Shang et

al. [22] proposed a dictionary learning based adversarial cross-modal retrieval to explore the complex statistical properties of multimodal data.

The above methods are usually shallow learning methods which mainly learn linear projections for cross-modal retrieval. Recently, due to the successful application of deep neural networks (DNN) in single-modal tasks such as image classification, DNN based cross-modal retrieval methods have been investigated in the literature. Some works attempt to extend traditional models to deep learning models. Typical examples include deep canonical correlation analysis (DCCA) [23], deep canonically correlated autoencoders (DCCAE) [24]. DCCA [23] employs two deep networks to learn a pair of highly correlated representation. DCCAE [24] is an extension of DCCA with an extra auto-encoder regularization term in the formulation of DCCA. To further preserve the discrimination among the samples from different semantic categories and eliminate the cross-modal discrepancy, deep supervised cross-modal retrieval (DSCMR) has been developed in [25]. The work in [26] utilize the inter- and intra- modality correlation to learn a more representative common subspace. Recently, multi-modal semantic autoencoder [27] has been developed for cross-modal retrieval.

Unfortunately, the above studies on cross-modal retrieval mainly focus on single-label dataset, which means they cannot capture the multi-label semantic information. Recently, by taking into account the high level semantic information in the form of multi-label annotations, multi-label CCA (ml-CCA) has been developed in [10]. However, ml-CCA involves huge computational cost which limits their application.

3. Brief review of ml-CCA

Let $X = [x_1, x_2, \cdots, x_{n_x}] \in \mathbf{R}^{d_x \times n_x}$ and $Z_x = [z_1, z_2, \cdots, z_{n_x}] \in \mathbf{R}^{c \times n_x}$ be the first modality data matrix and corresponding label matrix respectively. Similarly, $Y = [y_1, y_2, \cdots, y_{n_y}] \in \mathbf{R}^{d_y \times n_y}$ and $Z_y = [z_1, z_2, \cdots, z_{n_y}] \in \mathbf{R}^{c \times n_y}$ be the second modality data matrix and the corresponding label matrix, respectively. Let f be a similarity function which assigns a high value to the similar label pair (z_i, z_j), and a low value to the dissimilar label pair (z_i, z_j). In particular, the similarity function f is usually defined as

$$f(z_i, z_j) = e^{-\frac{\|z_i - z_j\|_2^2}{\sigma}} \tag{1}$$

where σ is a constant factor.

The formulation of ml-CCA [10] can be defined as

$$\rho = \max_{w,v} \frac{w^T C_{xy} v}{\sqrt{w^T C_{xx} w}\sqrt{v^T C_{yy} v}} \tag{2}$$

where

$$\begin{aligned} C_{xy} &= \tfrac{1}{N} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} f(z_i, z_j) x_i y_j^T \\ C_{xx} &= \tfrac{1}{N} \sum_{i=1}^{n_x} \alpha_i x_i x_i^T \\ C_{yy} &= \tfrac{1}{N} \sum_{j=1}^{n_y} \beta_j y_j y_j^T \end{aligned}$$

are the weighted covariance matrices, $N = n_x \times n_y$ is the total number of pairs across the two modalities, $\alpha_i = \sum_{j=1}^{n_y} f(z_i, z_j)$ and $\beta_i = \sum_{i=1}^{n_x} f(z_i, z_j)$.

Let $F \in \mathbf{R}^{n_x \times n_y}$, where $F_{ij} = f(z_i, z_j)$. With simple linear algebraic manipulation, we can rewrite C_{xy}, C_{xx}, C_{yy} in the following matrix form

$$\begin{aligned} C_{xy} &= \tfrac{1}{N} X F Y^T \\ C_{xx} &= \tfrac{1}{N} X D_x X^T \\ C_{yy} &= \tfrac{1}{N} Y D_y Y^T \end{aligned} \tag{3}$$

where D_x, D_y are diagonal matrices whose diagonal elements are given by $D_x(i,i) = \sum_{j=1}^{n_y} F_{ij}, D_y(j,j) = \sum_{i=1}^{n_x} F_{ij}$ respectively.

Since ρ is invariant to the scaling of w and v, ml-CCA can be formulated equivalently as the following optimization problem

$$\begin{aligned} \max_{w,v} \quad & w^T C_{xy} v \\ s.t. \quad & w^T C_{xx} w = 1,\ v^T C_{yy} v = 1 \end{aligned} \tag{4}$$

According to the Lagrange dual function in optimization theory, it can be easily shown that w is the eigen-vector corresponding to the largest eigenvalue of the following eigenvalue problem:

$$C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{xy} w = \lambda^2 w \tag{5}$$

For projection vector v, we have

$$v = \frac{C_{yy}^{-1} C_{yx} w}{\lambda^2} \tag{6}$$

However, the above approach suffers from several limitations. First, ml-CCA needs to precompute and store the similarity function $f(z_i, z_j)$ which requires huge storage for large number samples. Specifically, both computing and storing the similarity matrix require $\mathcal{O}(n^2)$. Second, $C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{xy}$ involves matrix inversion which has cubic complexity. In addition, $C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{xy}$ is usually not symmetric, thus many efficient methods for symmetric eigenvalue problems cannot be applied [28].

4. Proposed method

In this section, we present our scalable multi-label cross-modal model, dubbed sml-CCA. We first introduce an extra feature correlation term in our

framework to boost the performance. A semantic transformation is further introduced to avoid computing the semantic similarity matrix and the weighted covariance matrix explicitly. Finally, an efficient SVD based optimization algorithm is employed to solve the model.

4.1. Formulation

First, to further boost the performance of ml-CCA, we integrate the cross-modal feature correlation in the formulation of ml-CCA. Mathematically, our proposed model can be formulated as follow:

$$\begin{aligned} &\max_{w,v} w^T C_{xy} v \\ &\text{s.t. } w^T X D_x X^T w = 1,\ v^T Y D_y Y^T v = 1 \end{aligned} \tag{7}$$

where

$$C_{xy} = XFY^T + \eta XY^T \tag{8}$$

and η is a parameter to control the relative importance of semantic correlation part XFY^T and feature correlation part XY^T. As shown in ml-CCA, the optimization problem (7) can be solved by the following generalized problem

$$C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{xy} w = \lambda^2 w \tag{9}$$

However, this approach involves the explicit computation of the semantic similarity matrix F and dense matrix inversion leading to expensive computational cost.

4.2. Semantic transformation

In this section, we propose a semantic transformation to approximately compute the pairwise semantic similarity without explicitly formulating F.

We first define $p(z)$ as follow:

$$p(z) = \left[\sqrt{\frac{2(e^2-1)}{e\sigma}}e^{-\frac{\|z\|^2}{\sigma}}z; \sqrt{\frac{e^2+1}{e}}e^{-\frac{\|z\|^2}{\sigma}}\right] \tag{10}$$

where $z \in \mathbf{R}^c$. Given two vectors $z_i \in \mathbf{R}^c, z_j \in \mathbf{R}^c$, it is easy to show that

$$\begin{aligned} p(z_i)^T p(z_j) &= 2\left[\frac{e^2-1}{2e} \times \frac{2z_i^T z_j}{\sigma} + \frac{e^2+1}{2e}\right] e^{-\frac{\|z_i\|^2+\|z_j\|^2}{\sigma}} \\ &\approx 2e^{-\frac{\|z_i\|^2+\|z_j\|^2-2z_i^T z_j}{\sigma}} \\ &= 2e^{-\frac{\|z_i-z_j\|^2}{\sigma}} \end{aligned} \tag{11}$$

Here we use an approximation $\frac{e^2-1}{2e}a + \frac{e^2+1}{2e} \approx e^a$ for $a \in [-1, 1]$, which is shown in Figure 1. To make the approximation sensible, we require $-1 \leq \frac{2}{\sigma}z_i^T z_j \leq 1$. This can be achieved by setting $\sigma = 2\max\{\|z_i\|_2^2\}_{i=1}^n$.

It should be pointed out that the semantic transformation is inspired by SGH [29], which is proposed for binary codes learning with single-labels. Unlike SGH, we use semantic transformation to compute the similarity for multi-label vectors.

By using this semantic transformation, it is easy to show that the similarity matrix F and the two diagonal matrices can be computed as follows:

$$\begin{aligned} F &= P(Z)^T P(Z) \\ D_x &= \mathrm{diag}(P(Z)^T P(Z)e) \\ D_y &= \mathrm{diag}(e^T P(Z)^T P(Z)) \end{aligned} \tag{12}$$

where $P(Z) = [\frac{1}{\sqrt{2}}p(z_1), \frac{1}{\sqrt{2}}p(z_2), \cdots, \frac{1}{\sqrt{2}}p(z_n)] \in \mathbf{R}^{c' \times n}$, $c' = c + 1$ and $e \in \mathbf{R}^{n \times 1}$ is a vector with elements are all ones.

Please note that the time complexity is still $\mathcal{O}(n^2)$ if we explicitly compute F via (12). However, we use only $P(Z)$ for computation, but do not explicitly formulating F in our following learning algorithm. Hence $\mathcal{O}(n^2)$ complexity can be elegantly avoided in our learning algorithm.

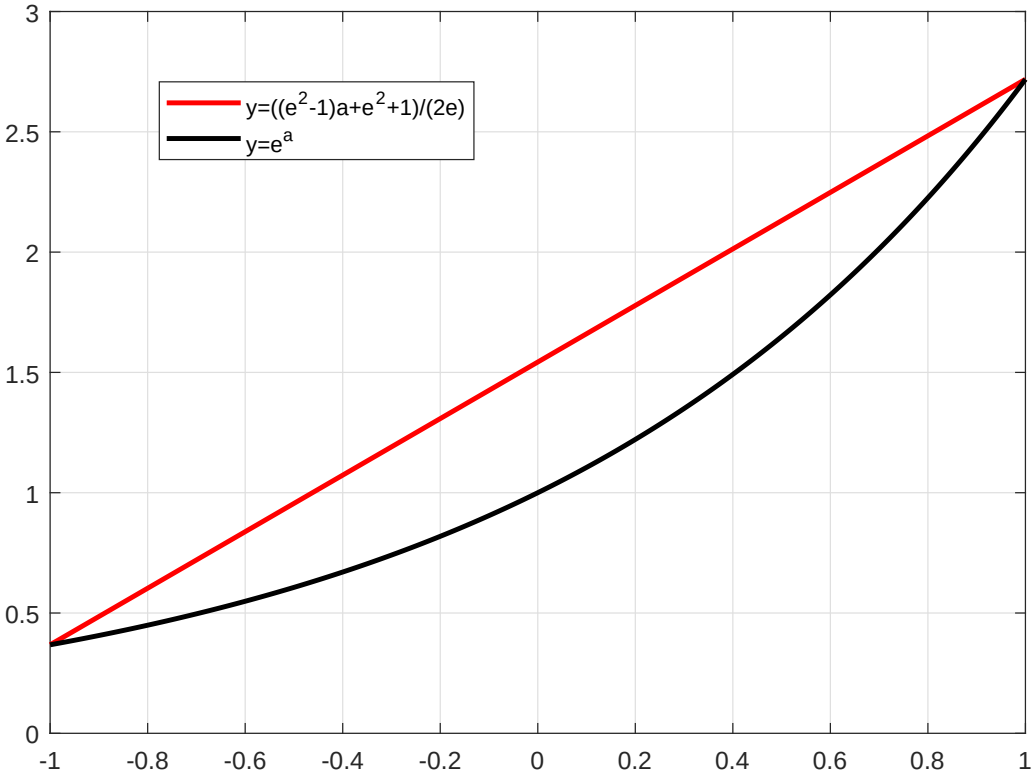

Figure 1: Approximation of semantic transformation

4.3. Optimization

We are now ready to present our learning algorithm. Let $\widetilde{X} = XD_x^{\frac{1}{2}}, \widetilde{Y} = YD_y^{\frac{1}{2}}$, Eq.(7) is equivalent to the following expression

$$\begin{aligned} &\max_{w,v} w^T \widetilde{C}_{xy} v \\ &\text{s.t. } w^T \widetilde{C}_{xx} w = 1,\ v^T \widetilde{C}_{yy} v = 1 \end{aligned} \tag{13}$$

where

$$\widetilde{C}_{xy} = \widetilde{X} D_x^{-\frac{1}{2}} \widetilde{P}^T \widetilde{P} D_y^{-\frac{1}{2}} \widetilde{Y}^T$$
$$\widetilde{C}_{xx} = \widetilde{X} \widetilde{X}^T$$
$$\widetilde{C}_{yy} = \widetilde{Y} \widetilde{Y}^T$$
$$\widetilde{P} = \begin{bmatrix} P \\ \sqrt{\eta} I \end{bmatrix}$$

Similar to ml-CCA [10], we can solve the following eigenvalue problem to get the optimal projection vector w

$$\widetilde{C}_{xx}^{-1} \widetilde{C}_{xy} \widetilde{C}_{yy}^{-1} \widetilde{C}_{yx} w = \lambda w \tag{14}$$

To make the computational process stable, we modify $\widetilde{C}_{xx}$ as $\widetilde{C}_{xx} = \widetilde{X}\widetilde{X}^T + \lambda_x I$, where λ_x is a regularization parameter to prevent singularity. Similarly, we have $\widetilde{C}_{yy} = \widetilde{Y}\widetilde{Y}^T + \lambda_y I$. Direct solving (14) is infeasible since we cannot take full advantage of the symmetric matrix. In addition, matrix inversion is usually computational expensive which requires $\mathcal{O}(d_x^3)$ time complexity. In the following, we develop an efficient approach to solve the eigenvalue problem (14).

Let

$$\widetilde{X} = U \Sigma V^T \tag{15}$$

be the singular value decompositon (SVD) [28] of $\widetilde{X}$, where $U \in \mathbf{R}^{d_x \times d_x}$ and $V \in \mathbf{R}^{n \times n}$ are orthogonal, $\Sigma = \mathrm{diag}(\Sigma_t, 0) \in \mathbf{R}^{d_x \times n}$ is diagonal, and $t = \mathrm{rank}(\widetilde{X})$. Let $U = [U_1, U_2]$, where $U_1 \in \mathbf{R}^{d_x \times t}, U_2 \in \mathbf{R}^{d_x \times (d_x - t)}$, $V = [V_1, V_2]$,$V_1 \in \mathbf{R}^{n \times t}, V_2 \in \mathbf{R}^{n \times (n-t)}$, and Σ_t consists of the first t rows and the first t columns of Σ. Similarly, Let

$$\widetilde{Y} = \widetilde{U} \widetilde{\Sigma} \widetilde{V}^T \tag{16}$$

be the singular value decompositon (SVD) of $\widetilde{Y}$, where $\widetilde{U} \in \mathbf{R}^{d_y \times d_y}$ and $\widetilde{V} \in \mathbf{R}^{n \times n}$ are orthogonal, $\widetilde{\Sigma} = \mathrm{diag}(\widetilde{\Sigma}_{\widetilde{t}}, 0) \in \mathbf{R}^{d_y \times n}$ is diagonal, and $\widetilde{t} = \mathrm{rank}(\widetilde{Y})$. Let $\widetilde{U} = [\widetilde{U}_1, \widetilde{U}_2]$, where $\widetilde{U}_1 \in \mathbf{R}^{d_y \times \widetilde{t}}, \widetilde{U}_2 \in \mathbf{R}^{d_y \times (d_y - \widetilde{t})}$, $\widetilde{V} = [\widetilde{V}_1, \widetilde{V}_2]$,$\widetilde{V}_1 \in \mathbf{R}^{n \times \widetilde{t}}, \widetilde{V}_2 \in \mathbf{R}^{n \times (n-\widetilde{t})}$, and $\widetilde{\Sigma}_t$ consists of the first $\widetilde{t}$ rows and the first $\widetilde{t}$ columns of $\widetilde{\Sigma}$.

Hence we have

$$\begin{aligned}
&\widetilde{C}_{xx}^{-1}\widetilde{C}_{xy}\widetilde{C}_{yy}^{-1}\widetilde{C}_{yx} \\
&= (\widetilde{X}\widetilde{X}^T + \lambda_x I)^{-1}\widetilde{X}D_x^{-\frac{1}{2}}\widetilde{P}^T\widetilde{P}D_y^{-\frac{1}{2}}\widetilde{Y}^T(\widetilde{Y}\widetilde{Y}^T + \lambda_y I)^{-1}\widetilde{Y}D_y^{-\frac{1}{2}}\widetilde{P}^T\widetilde{P}D_x^{-\frac{1}{2}}\widetilde{X}^T \\
&= U_1(\Sigma_t^2 + \lambda_x I)^{-1}\Sigma_t V_1^T D_x^{-\frac{1}{2}}\widetilde{P}^T\widetilde{P}D_y^{-\frac{1}{2}}\widetilde{Y}^T(\widetilde{Y}\widetilde{Y}^T + \lambda_y I)^{-1}\widetilde{Y}D_y^{-\frac{1}{2}}\widetilde{P}^T\widetilde{P}D_x^{-\frac{1}{2}}\widetilde{X}^T \\
&= U_1(\Sigma_t^2 + \lambda_x I)^{-1}\Sigma_t V_1^T D_x^{-\frac{1}{2}}\widetilde{P}^T\widetilde{P}D_y^{-\frac{1}{2}}\widetilde{V}_1\widetilde{\Sigma_{\widetilde{t}}}(\widetilde{\Sigma}_{\widetilde{t}}^2 + \lambda_y I)^{-1}\widetilde{\Sigma}_{\widetilde{t}}\widetilde{V}_1^T D_y^{-\frac{1}{2}}\widetilde{P}^T\widetilde{P}D_x^{-\frac{1}{2}}\widetilde{X}^T \\
&= U_1(\Sigma_t^2 + \lambda_x I)^{-1}\Sigma_t V_1^T D_x^{-\frac{1}{2}}\widetilde{P}^T BB^T\widetilde{P}D_x^{-\frac{1}{2}}V_1\Sigma_t U_1^T
\end{aligned} \tag{17}$$

where

$$B = \widetilde{P}D_y^{-\frac{1}{2}}\widetilde{V}_1\widetilde{\Sigma_{\widetilde{t}}}(\widetilde{\Sigma}_{\widetilde{t}}^2 + \lambda_y I)^{-0.5} \in \mathbf{R}^{(n+c') \times \widetilde{t}} \tag{18}$$

and the second and third equalities follow from the fact that $U_2^T\widetilde{X} = 0$ and $\widetilde{U}_2^T\widetilde{Y} = 0$ respectively.

Let

$$B = U_b\Sigma_b V_b^T \tag{19}$$

be the SVD of B, where $U_b \in \mathbf{R}^{(n+c') \times s}$, $V_b \in \mathbf{R}^{\widetilde{t} \times s}$, $\Sigma_b \in \mathbf{R}^{s \times s}$ is diagonal and $s = \mathrm{rank}(B)$ is the rank of matrix B. It follows that $BB^T = U_b\Sigma_q^2 U_b^T$. Hence we have

$$\begin{aligned}
&\widetilde{C}_{xx}^{-1}\widetilde{C}_{xy}\widetilde{C}_{yy}^{-1}\widetilde{C}_{yx} \\
&= U_1(\Sigma_t^2 + \lambda_x I)^{-1}\Sigma_t V_1^T D_x^{-\frac{1}{2}}\widetilde{P}^T BB^T\widetilde{P}D_x^{-\frac{1}{2}}V_1\Sigma_t U_1^T \\
&= U_1(\Sigma_t^2 + \lambda_x I)^{-1}\Sigma_t V_1^T D_x^{-\frac{1}{2}}\widetilde{P}^T U_b\Sigma_b^2 U_b^T\widetilde{P}D_x^{-\frac{1}{2}}V_1\Sigma_t U_1^T \\
&= U_1(\Sigma_t^2 + \lambda_x I)^{-1}\Sigma_t QQ^T\Sigma_t U_1^T
\end{aligned} \tag{20}$$

where $Q = V_1^T D_x^{-\frac{1}{2}} \widetilde{P}^T U_b \Sigma_b$. Define two diagonal matrices as follows

$$\begin{aligned} \Lambda_1 &= (\Sigma_t^2 + \lambda_x I)^{-1} \Sigma_t \\ \Lambda &= \Lambda_1^{\frac{1}{2}} \Sigma_t^{-\frac{1}{2}} \end{aligned} \tag{21}$$

It is easy to show that $\Lambda^{-1}\Lambda_1 = \Sigma_t \Lambda$. Hence, we have

$$\begin{aligned} &\widetilde{C}_{xx}^{-1} \widetilde{C}_{xy} \widetilde{C}_{yy}^{-1} \widetilde{C}_{yx} \\ &= U_1 (\Sigma_t^2 + \lambda_x I)^{-1} \Sigma_t Q Q^T \Sigma_t U_1^T \\ &= U_1 \Lambda (\Lambda^{-1} \Lambda_1) Q Q^T \Sigma_t \Lambda \Lambda^{-1} U_1^T \\ &= U_1 \Lambda \widetilde{\Lambda} Q Q^T \widetilde{\Lambda} \Lambda^{-1} U_1^T \end{aligned} \tag{22}$$

where $\widetilde{\Lambda} = \Lambda^{-1}\Lambda_1 = \Sigma_t \Lambda$. Denote $\Phi = \widetilde{\Lambda} Q \in \mathbf{R}^{t \times r_b}$ and $\Phi = MSN^T$ be the SVD of Φ, where $M \in \mathbf{R}^{t \times r_b}$, $N \in \mathbf{R}^{r_b \times r_b}$ and $S \in \mathbf{R}^{r_b \times r_b}$ is diagonal. It follows that

$$\begin{aligned} &\widetilde{C}_{xx}^{-1} \widetilde{C}_{xy} \widetilde{C}_{yy}^{-1} \widetilde{C}_{yx} \\ &= U_1 \Lambda \widetilde{\Lambda} Q Q^T \widetilde{\Lambda} \Lambda^{-1} U_1^T \\ &= U_1 \Lambda \Phi \Phi^T \Lambda^{-1} U_1^T \\ &= U_1 \Lambda M S N^T N S^T M^T \Lambda^{-1} U_1^T \\ &= U_1 \Lambda M S S^T M^T \Lambda^{-1} U_1^T \end{aligned} \tag{23}$$

Therefore, the eigen-decomposition of the matrix $\widetilde{C}_{xx}^{-1} \widetilde{C}_{xy} \widetilde{C}_{yy}^{-1} \widetilde{C}_{yx}$ is summarized in the following theorem:

Theorem 1. *There are ℓ nonzeros eigenvalues for the matrix $\widetilde{C}_{xx}^{-1} \widetilde{C}_{xy} \widetilde{C}_{yy}^{-1} \widetilde{C}_{yx}$. Specifically, the solution to problem (14), which consists of the eigenvectors corresponding to the top eigenvalues of $\widetilde{C}_{xx}^{-1} \widetilde{C}_{xy} \widetilde{C}_{yy}^{-1} \widetilde{C}_{yx}$ is given by*

$$W = U_1 \Lambda M_\ell \tag{24}$$

where M_ℓ represents the first ℓ columns of M.

Table 1 summarizes the overall algorithm for solving (14).

Table 1: Algorithm:sml-CCA

Input: Cross-modal data matrices X, Y, label matrix Z, η
Output: Projection matrices W, V
Use the semantic projection (10) to compute $P(Z)$
Compute the SVD of $\widetilde{X}$ and $\widetilde{Y}$
Compute B according to Equation (18)
Compute the SVD of $B = U_b \Sigma_b V_b^T$
Compute Λ_1, Λ and $\widetilde{\Lambda}$
Compute the SVD of $\Phi = MSN^T$
Compute $W = U_1 \Lambda M_\ell$
Compute V according to (6)

4.4. Complexity analysis

In this section, we analyze the computational complexity of our proposed method. From above optimization process, we can avoid explicitly compute and store the weight covariance matrix. However, we still need to compute D_x and D_y. If we write D_x as $D_x = \mathrm{diag}(P(Z)^T(P(Z)e))$, we can compute D_x with $\mathcal{O}(nc)$ time complexity, where c is the number of labels and n is the number of samples. Similar result can be observed for D_y. Therefore, the main computational cost of our algorithm is dominated by the SVD of $\widetilde{X}$, $\widetilde{Y}$, B and Φ. Specifically, the SVD for $\widetilde{X}$ and $\widetilde{Y}$ take $\mathcal{O}(nd^2)$ time complexity assuming $n > d$, where $d = \max\{d_x, d_y\}$. The SVD of B takes $\mathcal{O}((n + c')\widetilde{t}^2)$ time complexity. Similarly, the SVD of Φ takes $\mathcal{O}(ts^2)$ time complexity.

Table 2 summarizes the relevant matrices and their associated computational complexities. Typically, for cross-modal retrieval tasks, we usually

have $n \gg d$, where $d = \max(d_x, d_y)$. Therefore, t, $\widetilde{t}$ and s are relatively small. Thus, the cost of the proposed algorithm for computing the eigenvectors is dominated by the cost for computing the SVD of $\widetilde{X}$ and $\widetilde{Y}$. It can be concluded from Table 2 that our proposed method has linear time complexity in terms of the number of samples.

Table 2: Summary of relevant matrices and the associated complexity of each relevant matrix

Matrix	Size	Computation	Complexity
$\widetilde{X}$	$d_x \times n$	SVD	$\mathcal{O}(nd_x^2)$
$\widetilde{Y}$	$d_y \times n$	SVD	$\mathcal{O}(nd_y^2)$
B	$(n+c) \times \widetilde{t}$	SVD	$\mathcal{O}((n+c)\widetilde{t}^2)$
Q	$s \times t$	SVD	$\mathcal{O}(ts^2)$

5. Experiments

In this section, three multimedia datasets namely Pascal VOC dataset [30], NUS-WIDE [31] and MIRFlickr [9] are used to verify the effectiveness of our proposed scalable cross-modal retrieval method. In addition, Two cross-modal tasks: use an image query to search the relevant texts from the text view (denoted as I2T) and use text query to search the relevant image in the image view (denoted as T2I) are carried out to evaluate the retrieval performance of the compared algorithms. We describe the details of experimental settings and results in the following.

5.1. Datasets and features

Details of the above-mentioned three datasets are as follows:

- The Pascal VOC dataset [30] consists of 9963 images with their tags pairs categorized into 20 classes. Each image is described by a 576-dimension feature vector which consists of a 512-dimensional GIST feature vector and a 64-dimensional HSV color histograms, while the text is represented by the relative and absolute tag ranks. We use 4952 pairs for the query samples and the rest are used as training set.

- NUS-WIDE [31] contains 269,648 images and the associated tags, with a total number of 5,018 unique tags. In our experiment, we randomly select 1000 samples for query and 10000 samples for the training set. We use the AlexNet [32] to extract deep *fc7* features for each image by a 4096-dimensional vectors, and use bag-of- word vector to represent each text. The corresponding annotated tags of each image is represented by a 1000-D vector, where each dimension is a binary value to indicate whether a tag appears or not.

- The MIRFlickr [9] dataset consists of 25,000 images collected from Flickr website. Each image is associated with several textual tags. In our experiment, we select those pairs which have at least 20 textual tags. We also use the AlexNet [32] to extract deep *fc7* features for each image by a 4096-dimensional vectors, and use bag-of- word vector to represent each text. We randomly select 2000 samples to form the query set and the rest are used as the training set.

5.2. Compared methods and evaluation metric

Our proposed algorithm is compared against the following seven state-of-the-art methods:

- PCA. We apply principal component analysis (PCA) to each modal separately.

- CCA. Canonical correlation analysis [4] learns a common subspace by maximizing the correlation of two modalities. In our implementation, we use the matlab function *canoncorr* which is highly optimized and fast.

- CCA3. Multiview canonical correlation analysis proposed in [11]. CCA3 learns the projection matrices by incorporating the semantic matrix as the third view.

- ml-CCA. multi-label correlation analysis proposed in [10] for cross-modal retrieval.

- DCCA. Deep canonical correlation analysis which is a nonlinear extension of traditional CCA with deep learning technique [23].

- DCCAE. Deep canonically correlated autoencoders which consists of two autoencoders and optimizes the combination of canonical correlation and the reconstruction errors of the autoencoders [24].

- MMSAE. MMSAE is a two stage cross-modal retrieval method which employs autoencoder to preserve feature and semantic information [27].

For all the compared methods , we empirically set the dimension of the common subspace to 10 (i.e. $\ell = 10$). An in-depth experimental verification is presented in Section 5.4.

Two popular criteria: mean average precision (mAP) and precision-recall curve are utilized to evaluate the performance of different methods. These two criteria are defined as follows.

- mAP: The average precision (AP) is defined as $\mathrm{AP}(q) = \frac{1}{R'} \sum_{r=1}^{R} P_q(r)\delta_q(r)$, where R' is the total number of relevant samples in the retrieved set, R is the number of retrieved samples. $P_q(r)$ denotes the top-r precision of the q-th query, and $\delta_q(r) = 1$ if the r-th data item is relevant to the q-th query;otherwise $\delta_q(r) = 0$. Clearly, the larger the mAP value, the better the performance. The mAP over Q queries can be computed by

$$\mathrm{mAP} = \frac{1}{Q} \sum_{q=1}^{Q} \mathrm{AP}(q)$$

- Precision-recall: The precision and recall are defined as

$$\text{precision} = \frac{\text{the number of retrieved relevant items}}{\text{the number of all retrieved items}}$$

$$\text{recall} = \frac{\text{the number of retrieved relevant items}}{\text{the number of all relevant items}}$$

  Precision-recall reflects the precision values at different recall levels. Typically, the area under the precision-recall curve is computed and a larger value indicates better performance.

In addition to these two metrics, we also report the score of precision@K (P@K) to evaluate the performance of the compared methods. Specifically, we include precision values at top 5 results (P@5), top 10 results (P@5) and top 15 results (P@15).

5.3. Experimental results

The mAP values on three datasets are presented in Tables 3 to 5 for I2T and T2I tasks. Note that DCCA, DCCAE and MMSAE are deep learning based methods. Several observations can be made from these tables. First, ml-CCA, MMSAE and sml-CCA which utilize the multi-label information to learn the common subspace outperform the other compared methods. This implies the importance of integrating semantic multiple labels for cross-modal retrieval tasks. Second, our sml-CCA achieves better results than traditional common subspace learning algorithms PCA, CCA, CCA3 and ml-CCA. Specifically, compared with ml-CCA, our sml-CCA improves average mAP values of two tasks 2.4%, 14.3%, 4.0% on Pascal Voc, NUS-WIDE and MIRFlickr data sets respectively. In addition, our sml-CCA even outperform deep learning based methods DCCA and DCCAE because of taking semantic information into account.

Tables 6-8 depict precisions at different K (P@K). It is clear from Tables 6 and 8 that MMSAE and sml-CCA outperform the other compared methods. This again shows that integrating semantic information can significantly improve retrieval performance. It is interesting to note that MMSAE achieves the best performance. This can be attribute to the powerful representation learning ability of autoencoder introduced in MMSAE. Moreover, unlike D-CCA and DCCAE, MMSAE can integrate the semantic information to learn discriminant representation.

In addition to mAP and P@K, the precision-recall curves are plotted in Figures 2 to 4. The results keep consistent with the mAP values and P@K values reported in the previous tables, which further demonstrates the

effectiveness of our proposed sml-CCA.

Table 3: mAP values on Pascal VOC dataset

Task	PCA	CCA	CCA3	ml-CCA	DCCA	DCCAE	MMSAE	sml-CCA
I2T	0.2737	0.3686	0.2460	0.4686	0.4052	0.4087	0.4871	0.4510
T2I	0.2286	0.3759	0.2583	0.4846	0.4067	0.4104	0.5258	0.5661
Avarage	0.2511	0.3722	0.2521	0.4766	0.4060	0.4479	0.4485	**0.4884**

Table 4: mAP values on NUS-WIDE dataset

Task	PCA	CCA	CCA3	ml-CCA	DCCA	DCCAE	MMSAE	sml-CCA
I2T	0.7339	0.7099	0.7053	0.7325	0.7185	0.7165	0.8338	0.8328
T2I	0.7046	0.7107	0.7886	0.7205	0.7147	0.7166	0.8331	0.8290
Average	0.7192	0.7103	0.7470	0.7265	0.7166	0.7166	**0.8335**	0.8309

Table 5: mAP values on MIRFlickr dataset

Task	PCA	CCA	CCA3	ml-CCA	DCCA	DCCAE	MMSAE	sml-CCA
I2T	0.5983	0.6342	0.5536	0.6813	0.6558	0.6444	0.7387	0.6985
T2I	0.5698	0.6426	0.6450	0.6826	0.6564	0.6450	0.7655	0.7193
Average	0.5796	0.6384	0.5993	0.6820	0.6561	0.6447	**0.7521**	0.7089

5.4. Dimension of common space analysis

In our previous experiments, we empirically set $\ell = 10$. In this subsection, we analysis the effect of dimension ℓ of common subspace for cross-modal retrieval task on three datasets. We change ℓ in the range of [2, 20] with other parameters fixed.

Table 6: Top K precision on Pascal VOC dataset

Task	Method	P@5	P@10	P@15
I2T	PCA	0.2573	0.2523	0.2479
	CCA	0.4225	0.4198	0.4177
	CCA3	0.2504	0.2406	0.2381
	ml-CCA	0.5200	0.5198	0.5193
	DCCA	0.4834	0.4811	0.4785
	DCCAE	0.4788	0.4757	0.4736
	MMSAE	**0.5483**	**0.5474**	**0.5455**
	sml-CCA	0.4864	0.4837	0.4801
T2I	PCA	0.2141	0.2015	0.1979
	CCA	0.5395	0.5307	0.5257
	CCA3	0.2488	0.2465	0.2426
	ml-CCA	0.7485	0.7401	0.7346
	DCCA	0.6372	0.6354	0.6301
	DCCAE	0.6409	0.6284	0.6286
	MMSAE	0.7583	0.7387	0.7336
	sml-CCA	**0.8112**	**0.7986**	**0.7928**

Figures 5 to 7 depict the mAP values versus dimension. Figure 5 reveals that our method sml-CCA can get better performance than the other compared methods when the dimension is larger than 6. Similar observations can be drawn from Figures 6 and 7.

In summary, these results show that our method perform well in a wide range of dimension. Hence our choice of ℓ in the experiments is reasonable.

Table 7: Top K precision on NUS-WIDE dataset

Task	Method	P@5	P@10	P@15
I2T	PCA	0.7888	0.7827	0.7797
	CCA	0.8128	0.7803	0.7658
	CCA3	0.6334	0.6583	0.6724
	ml-CCA	0.7834	0.7829	0.7813
	DCCA	0.7530	0.7514	0.7508
	DCCAE	0.8351	0.8177	0.8096
	MMSAE	**0.9177**	**0.9183**	**0.9178**
	sml-CCA	0.9110	0.9134	0.9142
T2I	PCA	0.6570	0.6810	0.6828
	CCA	0.8205	0.7890	0.7614
	CCA3	0.7417	0.7468	0.7496
	ml-CCA	0.7568	0.7569	0.7533
	DCCA	0.7685	0.7638	0.7600
	DCCAE	0.8519	0.8356	0.8232
	MMSAE	**0.9486**	**0.9456**	**0.9445**
	sml-CCA	0.9380	0.9366	0.9383

5.5. Parameter sensitivity analysis

In this section, we conduct parameter analysis to empirically show how to choose the value of η.

Figure 8 shows the influence of the mAP values with respect to η on three datasets. We can observe that mAP values on Pascal VOC dataset go up as the value η increases from 1 to 10 while the mAPs values on NUS-WIDE

Table 8: Top K precision on MIRFlickr dataset

Task	Method	P@5	P@10	P@15
I2T	PCA	0.6345	0.6328	0.6322
	CCA	0.7473	0.7464	0.7460
	CCA3	0.5581	0.5559	0.5569
	ml-CCA	0.8368	0.8350	0.8338
	DCCA	0.8283	0.8293	0.8298
	DCCAE	0.7786	0.7810	0.7831
	MMSAE	**0.8559**	**0.8545**	**0.8545**
	sml-CCA	0.8475	0.8468	0.8480
T2I	PCA	0.5781	0.5775	0.5723
	CCA	0.7832	0.7837	0.7844
	CCA3	0.6574	0.6579	0.6578
	ml-CCA	0.7917	0.7900	0.7907
	DCCA	0.8237	0.8214	0.8210
	DCCAE	0.8003	0.7966	0.7951
	MMSAE	**0.8806**	**0.8814**	**0.8823**
	sml-CCA	0.8698	0.8687	0.8679

and MIRFlickr datasets reach relative stable between 1 and 10. Therefore, we can choose the optimal value within the range of $[1, 10]$.

From the above analysis, we can conclude that sml-CCA can achieve stable performance under a wide range of parameter values.

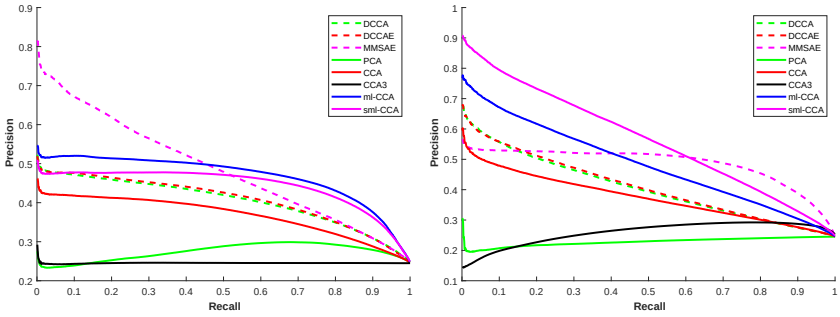

Figure 2: Precision-recall curves of Pascal VOC dataset in I2T (left) and T2I (right) tasks

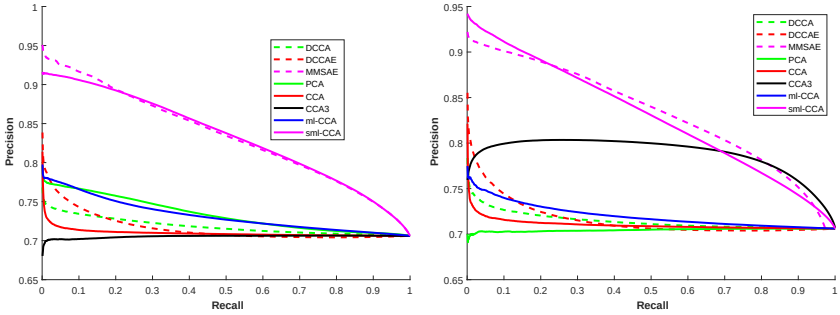

Figure 3: Precision-recall curves of NUS-WIDE dataset in I2T (left) and T2I (right) tasks

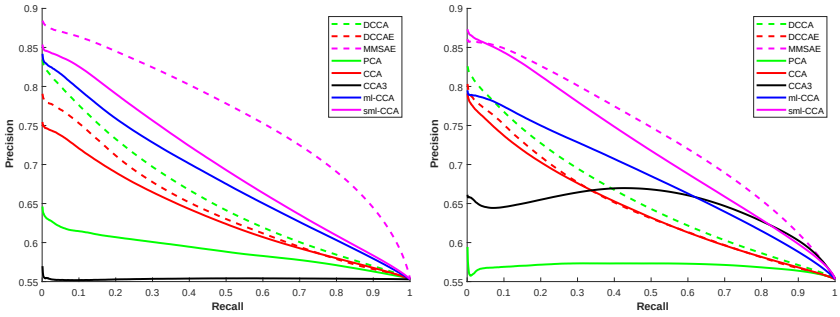

Figure 4: Precision-recall curves of MIRFlickr dataset in I2T (left) and T2I (right) tasks

5.6. Scalability evaluation

In this experiment, we use datasets Pascal VOC and MIRFlickr to compare the scalability of sml-CCA, MMSAE, ml-CCA and CCA. More precisely,

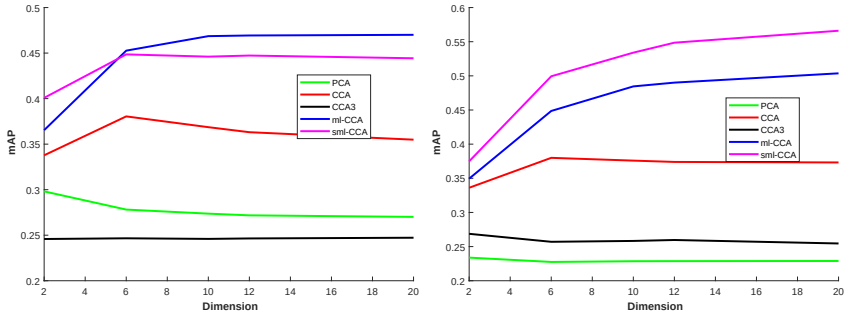

Figure 5: Effect of the dimension of the common subspace in I2T (left) and T2I (right) tasks on Pascal Voc dataset

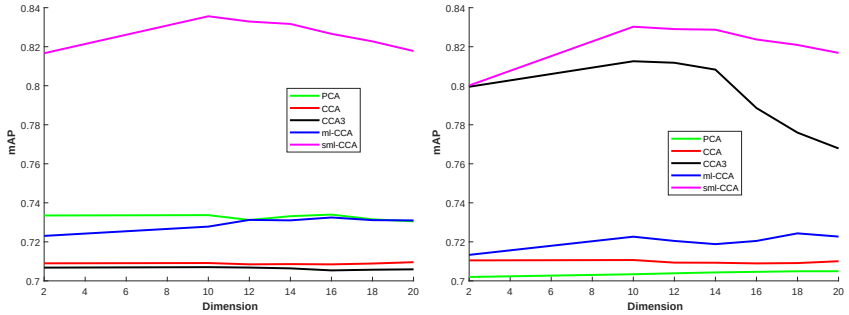

Figure 6: Effect of the dimension of the common subspace in I2T (left) and T2I (right) tasks on Nuswide dataset

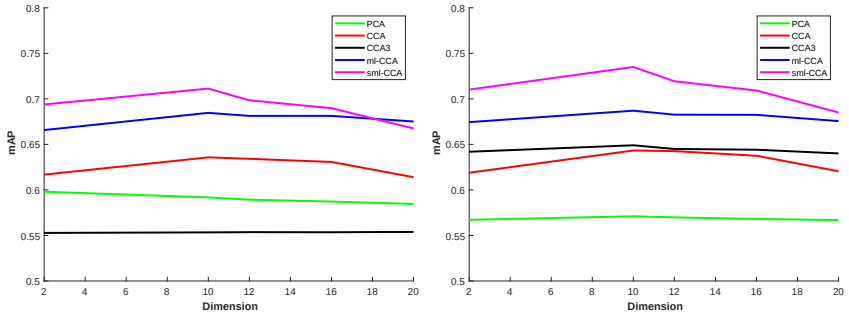

Figure 7: Effect of the dimension of the common subspace in I2T (left) and T2I (right) tasks on MIRFlickr dataset

we evaluate the computational time for the steadily increasing data dimension under the fixed number of data samples and vice versa.

Figure 9 displays the computational time of the compared methods with growing size of training sample under fixed dimension. We can see that sml-CCA is much faster than ml-CCA. This is due to the fact that sml-CCA has the linear time complexity in terms of the number of samples, while ml-CCA has the quadratic complexity. On top of that, Figure 10 displays the computation time of the compared methods as data dimension increases when the number of training samples fixed . Similar trends can be observed as in Figure 9. Note that sml-CCA requires only $\mathcal{O}(d^2)$ computation time complexity, which is smaller than that of ml-CCA with $\mathcal{O}(d^3)$ time complexity. From Figure 9 and 10, we arrive at the conclusion that the proposed sml-CCA is potentially much more scalable than ml-CCA.

It is interesting to note that CCA is much faster than sml-CCA. We conjecture that CCA does not involve the computation of semantic similarity matrix which is a huge save of computation cost. Moreover, in our implementation, we use the highly optimized matlab function *canoncorr* for CCA.

6. Conclusion

In this paper, we have proposed a scalable multi-label canonical correlation analysis (sml-CCA) for cross-modal retrieval. sml-CCA combines the merits of feature correlation and semantic correlation to boost the performance. A novel semantic transformation is further introduced to elegantly avoid the expensive computation of the semantic similarity matrix. We then design an efficient learning algorithm to learn the projection vectors.

Experimental results on three multi-label datasets have demonstrated the effectiveness and efficiency of the proposed approach.

In the future, we will extend our current work to learn nonlinear projections by using the kernel technique. In addition, deep learning can also be employed in our framework to learn more powerful projections.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant 61806097, 61602248), the Academy of Finland for project MiGA (grant 316765), ICT 2023 project (grant 328115), Infotech Oulu and the China Scholarship Council.

As well, the authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

Reference

[1] Yang, Y., Wu, F., Xu, D., Zhuang, Y., Chia, L.T.. Cross-media retrieval using query dependent search methods. Pattern Recognition 2010;43(8):2927 – 2936.

[2] Wu, F., Jing, X.Y., Wu, Z., Ji, Y., Dong, X., Luo, X., et al. Modality-specific and shared generative adversarial network for cross-modal retrieval. Pattern Recognition 2020;104:107335.

[3] Peng, Y., Huang, X., Zhao, Y.. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. IEEE Transactions on circuits and systems for video technology 2017;28(9):2372–2385.

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

[5] Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., et al. A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM international conference on Multimedia. 2010, p. 251–260.

[6] Pereira, J.C., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G.R., Levy, R., et al. On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE transactions on pattern analysis and machine intelligence 2013;36(3):521–535.

[7] Bhatt, G., Jha, P., Raman, B.. Representation learning using step-based deep multi-modal autoencoders. Pattern Recognition 2019;95:12 – 23.

[8] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009, p. 248–255.

[9] Huiskes, M.J., Lew, M.S.. The mir flickr retrieval evaluation. In: Proceedings of the 1st ACM international conference on Multimedia information retrieval. ACM; 2008, p. 39–43.

[10] Ranjan, V., Rasiwasia, N., Jawahar, C.. Multi-label cross-modal retrieval. In: Proceedings of the IEEE International Conference on Computer Vision. 2015, p. 4094–4102.

[11] Gong, Y., Ke, Q., Isard, M., Lazebnik, S.. A multi-view embedding space for modeling internet images, tags, and their semantics. International journal of computer vision 2014;106(2):210–233.

[12] Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.. Generalized multiview analysis: A discriminative latent space. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2012, p. 2160–2167.

[13] Cao, G., Iosifidis, A., Chen, K., Gabbouj, M.. Generalized multi-view embedding for visual recognition and cross-modal retrieval. IEEE transactions on cybernetics 2017;48(9):2542–2555.

[14] Yuan, Y.H., Sun, Q.S., Ge, H.W.. Fractional-order embedding canonical correlation analysis and its applications to multi-view dimensionality reduction and recognition. Pattern Recognition 2014;47(3):1411 – 1424.

[15] Li, D., Dimitrova, N., Li, M., Sethi, I.K.. Multimedia content processing through cross-modal association. In: Proceedings of the eleventh ACM international conference on Multimedia. 2003, p. 604–611.

[16] Zhai, X., Peng, Y., Xiao, J.. Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In: Twenty-seventh AAAI conference on artificial intelligence. 2013, p. 1198–1204.

[17] Zhai, X., Peng, Y., Xiao, J.. Learning cross-media joint representation with sparse and semisupervised regularization. IEEE Transactions on Circuits and Systems for Video Technology 2013;24(6):965–978.

[18] Peng, Y., Zhai, X., Zhao, Y., Huang, X.. Semi-supervised cross-media feature learning with unified patch graph regularization. IEEE transactions on circuits and systems for video technology 2015;26(3):583–596.

[19] Zhang, L., Ma, B., Li, G., Huang, Q., Tian, Q.. Generalized semi-supervised and structured subspace learning for cross-modal retrieval. IEEE Transactions on Multimedia 2017;20(1):128–141.

[20] Zhuang, Y.T., Wang, Y.F., Wu, F., Zhang, Y., Lu, W.M.. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In: Twenty-Seventh AAAI Conference on Artificial Intelligence. 2013, p. 1070–1076.

[21] Zhu, F., Shao, L., Yu, M.. Cross-modality submodular dictionary learning for information retrieval. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. 2014, p. 1479–1488.

[22] Shang, F., Zhang, H., Zhu, L., Sun, J.. Adversarial cross-modal retrieval based on dictionary learning. Neurocomputing 2019;355:93–104.

[23] Andrew, G., Arora, R., Bilmes, J., Livescu, K.. Deep canonical correlation analysis. In: International conference on machine learning. 2013, p. 1247–1255.

[24] Wang, W., Arora, R., Livescu, K., Bilmes, J.. On deep multi-view representation learning. In: International Conference on Machine Learning. 2015, p. 1083–1092.

[25] Zhen, L., Hu, P., Wang, X., Peng, D.. Deep supervised cross-modal retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019, p. 10394–10403.

[26] Peng, Y., Qi, J., Huang, X., Yuan, Y.. Ccl: Cross-modal correlation learning with multigrained fusion by hierarchical network. IEEE Transactions on Multimedia 2017;20(2):405–420.

[27] Wu, Y., Wang, S., Huang, Q.. Multi-modal semantic autoencoder for cross-modal retrieval. Neurocomputing 2019;331:165 – 175.

[28] Golub, G.H., Van Loan, C.F.. Matrix computations; vol. 3. JHU press; 2012.

[29] Jiang, Q.Y., Li, W.J.. Scalable graph hashing with feature transformation. In: Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015, p. 2248–2254.

[30] Hwang, S.J., Grauman, K.. Reading between the lines: Object localization using implicit cues from image tags. IEEE transactions on pattern analysis and machine intelligence 2011;34(6):1145–1158.

[31] Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.. Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. 2009, p. 1–9.

[32] Krizhevsky, A., Sutskever, I., Hinton, G.E.. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012, p. 1097–1105.

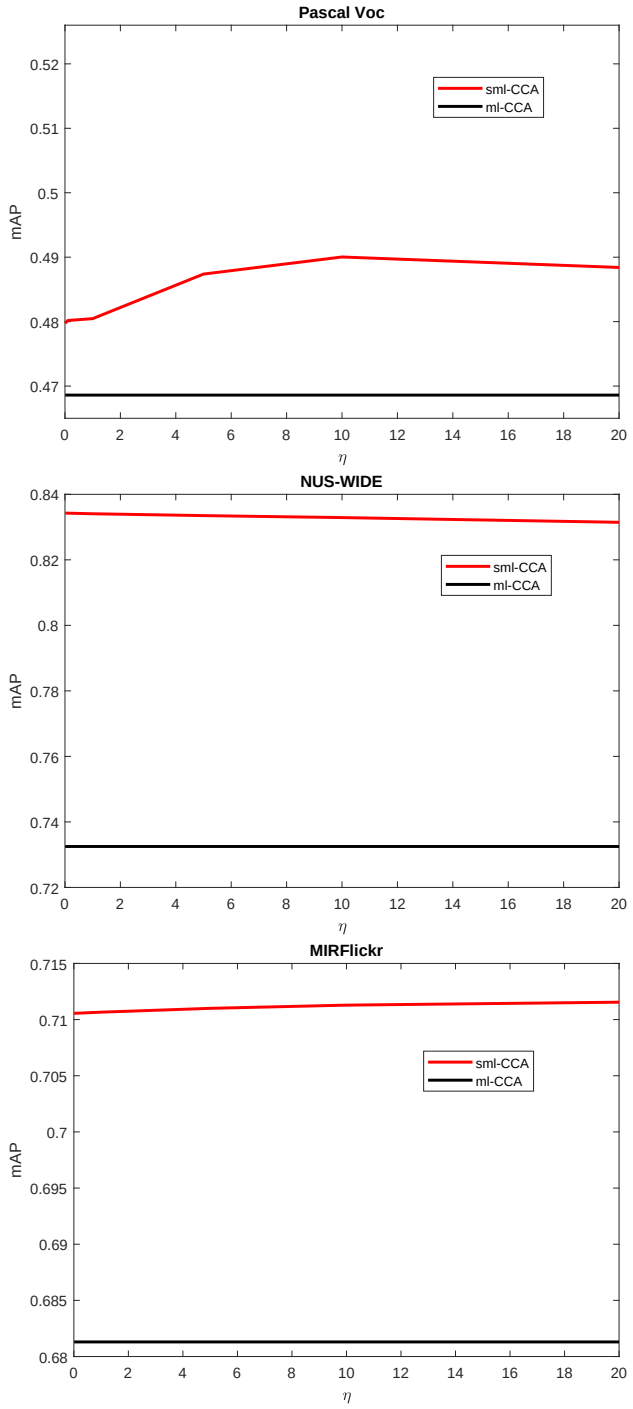

Figure 8: Parameter sensitivity analysis of η on Pascal Voc (top), NUS-WIDE (middle) and MIRFlickr (bottom) datasets

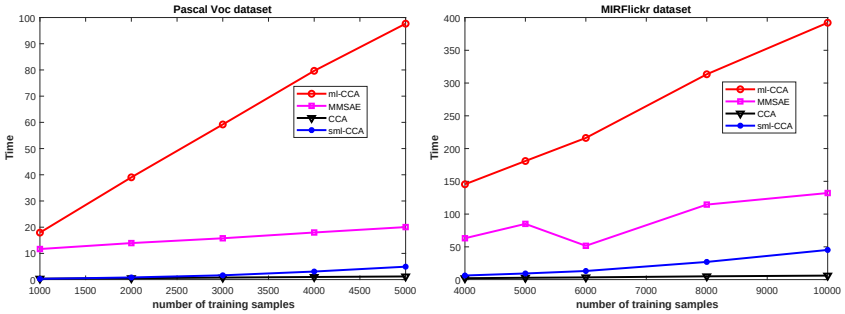

Figure 9: Comparison of computation time for sml-CCA, MMSAE, ml-CCA and CCA on Pascal Voc (left) and MIRFlickr (right) datasets as the sample size increases.

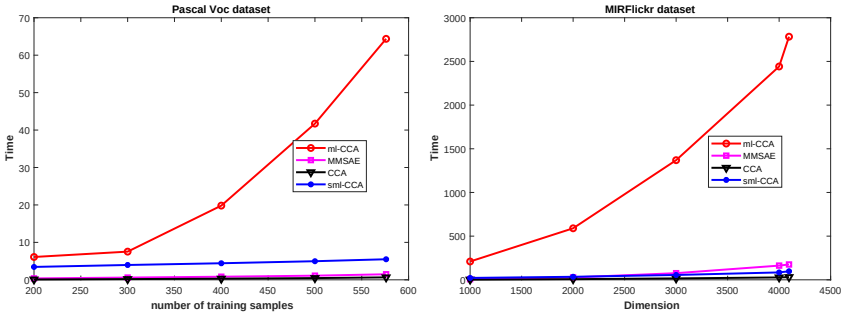

Figure 10: Comparison of computation time for sml-CCA, MMSAE, ml-CCA and CCA on Pascal Voc (left) and MIRFlickr (right) datasets as the dimension increases.