Cross-Modal Subspace Clustering via Deep Canonical Correlation Analysis

Paper ID: 2437

Abstract

For cross-modal subspace clustering, the key is to exploit the correlation information between cross-modal data. However, most hierarchical and structural correlation information among cross-modal data cannot be well exploited due to its high-dimensional non-linear property. To tackle this problem, in this paper, we propose an unsupervised framework named Cross-Modal Subspace Clustering via Deep Canonical Correlation Analysis (CMSC-DCCA), which incorporates the correlation constraint with self-expressive layer to make full use of information among the inter-modal data and intra-modal data. More specifically, the proposed model consists of three components: 1) deep canonical correlation analysis (Deep CCA) framework; 2) self-expressive layer; 3) Deep CCA decoders. The Deep CCA model consists of convolutional encoders and correlation constraint, i.e., cross-modal data are sent to its corresponding convolutional encoders and obtain the latent representations, while adding the correlation constraint for the latent representations can make full use of the information of the inter-modal data. Furthermore, the latent representations constrained by the self-expressive layer can perform self-expression properties, which can capture the hierarchical intra-modal correlations of each modal. Then Deep CCA decoders reconstruct data to ensure that the encoded features can preserve the structure of the original data. Experimental results on several real-world datasets demonstrate the proposed method outperforms the state-of-the-art methods.

Introduction

In the era of data explosion, clustering task has attracted considerable attention. Most traditional clustering methods mainly focus on the clustering problem of low-dimensional data. Generally, the traditional clustering methods can be roughly divided into five categories: 1) Non-negative matrix factorization (NMF) clustering (Akata, Thurau, and Bauckhage 2011); 2) Multi-kernel learning strategy (Guo et al. 2014); 3) Subspace based clustering method (Chaudhuri et al. 2009); 4) Self-representation based method (Yin et al. 2015); 5) Graph constraint based method (Xia et al. 2014; Nie, Cai, and Li 2017). Among these above methods, subspace and self-representation based methods have received a lot of attention and achieved remarkable results. However, these traditional methods adopt shallow and linear embedding functions to reveal the intrinsic structure of data, which cannot simulate the high-dimensional nonlinear characteristics of data very well. Especially for subspace based clustering methods, high-dimensional data can easily lead to "curse of dimensionality".

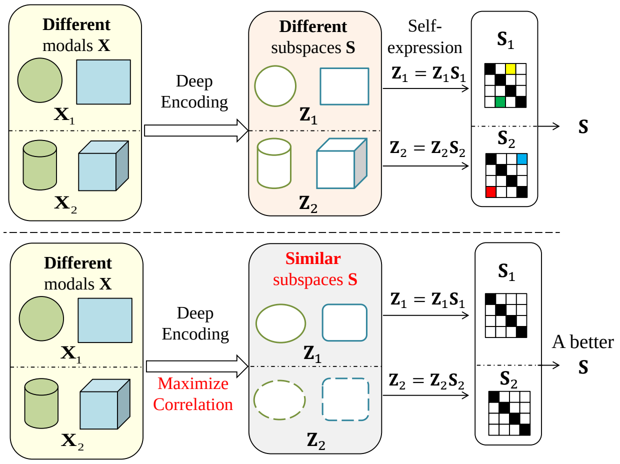


Figure 1: The illustration shows that the learned features of two modal in the subspace are not similar without the correlation constraint, which cannot learn a good shared subspace coefficient matrix $\mathbf{S}$ to replace both $\mathbf{S_1}$ and $\mathbf{S_2}$ to reflect the structure characteristics of the two modal data simultaneously. However, combining the correlation constraint with the self-expressive layer can learn a better shared subspace coefficient matrix $\mathbf{S}$. $\mathbf{X_1}$ and $\mathbf{X_2}$ are the input data. $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are the latent representations in subspace. $\mathbf{S_1}$ and $\mathbf{S_2}$ are the self-expression coefficient matrix.

To solve the "curse of dimensionality" problem, Elhamifar et al. (2013) propose a Sparse Subspace Clustering (SSC) method to reduce dimension. Patel et al. (2015) propose Latent Space Sparse Subspace Clustering (LSSC) (Patel, Van Nguyen, and Vidal 2015) that simultaneously performs dimensionality reduction and sparse coding for SSC. With the development of neural networks, some people use the deep learning to solve subspace clustering problems. For instance, Ji et al. (2017) use stacked auto-encoders as their basic model and adopt self-expression to learn the affinity of data in latent space to cluster. Deep Multi-modal Subspace Clustering (DMSC) (Abavisani and Patel 2018) presents convolutional neural network (CNN) approach for unsupervised multi-modal subspace clustering. However, for cross-modal (Rasiwasia et al. 2010; Jia, Salzmann, and Darrell 2011) subspace clustering, there are some superior methods but not many. For example, Zhang et al. (2007) propose to exploit cross-modal correlation to cluster two modal data. He et al. (2015) use cross-modal learning via the pairwise constraint and aim to find the common hidden structure of cross-modal data. These cross-modal methods reduce the semantic gap of the inter-modal data and improve the clustering accuracy. However, these methods cannot solve the deep data clustering problems very well. Moreover, there are still existing some challenges for cross-modal subspace clustering problem:

- How should we consider the correlation of the inter-modal data and the intra-modal data simultaneously to learn a representative shared subspace representation to improve clustering accuracy?
- How could we guarantee that the features encoded by the deep network can still reflect the structural distribution of the original data?

To address these above challenges, as shown in Figure 2, we propose a novel Cross-Modal Subspace Clustering framework via Deep Canonical Correlation Analysis (CMSC-DCCA) to improve clustering performance. The proposed method consists of three parts: deep canonical correlation analysis (Deep CCA) (Andrew et al. 2013) model, a self-expressive layer and Deep CCA decoder. In the Deep CCA model, we map the high-dimensional nonlinear cross-modal data into latent representations by deep convolution encoders; meanwhile, correlation constraint can maximize the correlation of the inter-modal data to improve clustering accuracy. For the self-expressive layer, a self-expressive loss function is employed for the latent representation, which can reveal intra-modal correlation information of each modal. In addition, we add the Deep CCA decoder method to reconstruct the original data, which ensures the representation processed by the encoder can still reflect the characteristics of the original data very well. Furthermore, Figure 1 shows the purpose of combining the correlation constraint with the self-expression layer. Intuitively, without the correlation constraint, it is difficult to learn a shared subspace representation that reflects the structure of two modal data simultaneously.

The main contributions of our method are summarized as:

- We propose a novel deep cross-modal subspace clustering method (CMSC-DCCA) by incorporating the correlation constraint with the self-expression layer, which can consider both the correlation of the inter-modal data and the distribution of the intra-modal data.
- We add Deep CCA decoder to realize high performance cross-modal clustering by reconstructing the encoded original data, which can ensure that the encoded features can reflect the overall structural distribution of data well.
- We give an efficient algorithm to optimize loss function and train the entire network at once to reduce the calculation of parameters. In addition, we perform spectral clustering on shared coefficient matrix. Experiments show our model achieves the best clustering performance.

Related works

In recent years, with the continuous accumulation of data, data processing plays an increasingly important role in the field of data mining. In real life, a lot of data can be described from different modal. Therefore, cross-modal data processing is getting more and more attention, which focuses on maximizing the correlation of the inter-modal data. In this section, we provide a brief review about the correlation study on cross-modal data.

Among cross-modal data correlation studies, the most famous method is based on Canonical Correlation Analysis (CCA) (Kim, Kittler, and Cipolla 2007). The main idea of the CCA method is to find the mapping vector of each modal to the common space by maximizing the correlation of the inter-modal data. However, the CCA method can only calculate the linear correlation of the inter-modal data. In real applications, the relationship between cross-modal data may be nonlinear. To solve this problem, some nonlinear CCA methods have been proposed, such as, the Kernel Canonical Correlation Analysis (KCCA) (Akaho 2006) and Locality Preserving CCA (LPCCA) (Sun and Chen 2007) methods. However, these methods have high computational complexity. In addition, it is easy over fitting and relatively difficult to choose a suitable kernel function for KCCA method. Deep Canonical Correlation Analysis (Deep CCA) (Andrew et al. 2013) method can learn complex nonlinear transformations of the inter-modal data such that the resulting representations are highly correlated through a multi-layer deep network. The disadvantage of Deep CCA is that cannot reconstruct data and results in poor clustering performance. Wang et al. (2016) further propose Deep Canonically Correlated Auto-encoders (DCCAE) with a auto-encoder to reconstruct data. Deep Generalized Canonical Correlation Analysis (DGCCA) (Benton et al. 2017) method learns a modal-independent representation to cluster, which reduced redundant information for data. However, both DCCAE and DGCCA methods did not consider the correlation of the intra-modal data.

The Proposed Method

In this section, we first give the motivations of Cross-Modal Subspace Clustering via Deep Canonical Correlation Analysis model. Then we introduce the framework of the proposed model. Finally we analyze the loss functions and training process in detail.

Motivations

According to previous related work, there are two motivations for our method as follows:

Motivation 1: There are some cross-modal clustering methods based on canonical correlation analysis, such as (Zhang, Zhuang, and Wu 2007; Jin et al. 2015). However, these methods can only process linear and shallow data. To solve this problem, we propose cross-modal clustering method based on deep canonical correlation analysis, which can process high-dimensional nonlinear data very well. DCCAE (Wang et al. 2016) method ignores the relationship of the intra-modal data, which cannot reflect the discriminant information of data well. In order to make full use of the relationship of the inter-modal data and the information of the intra-modal data, we combine the correlation constraint with the self-expression layer to propose the deep cross-modal subspace clustering method (CMSC-DCCA). Assume that the two modal coded representations are defined as $\mathbf{Z}_1$ and $\mathbf{Z}_2$. The self-expression layer performs self-expression property for each modal: $\mathbf{Z}_1$=$\mathbf{Z}_1\mathbf{S}_1$ and $\mathbf{Z}_2$=$\mathbf{Z}_2\mathbf{S}_2$. $\mathbf{S_1}$ and $\mathbf{S_2}$ are the self-expression coefficient matrix. We use the correlation calculation to maximize the correlation of the inter-modal data $\mathbf{Z}_1$ and $\mathbf{Z}_2$, which purpose is to learn a better shared coefficient matrix $\mathbf{S}$ to replace $\mathbf{S_1}$ and $\mathbf{S_2}$ to reflect the structure of two modal data simultaneously. Our method ensures that different data with high similarity are more likely to be grouped into one class.

Motivation 2: Deep canonical correlation analysis (Deep CCA) (Andrew et al. 2013) method solves nonlinear problems of data using deep neural networks. However, it does not consider whether the data encoded by the neural network can maintain the structure of the original data. Thus, we add decoders based on Deep CCA, which can reconstruct the latent features from deep canonical correlation analysis model. The decoders aim to ensure that the latent features can reflect the structure of the original data to improve clustering performance.

In this paper, our work can not only consider the correlation of the intra-modal data by the self-expression layer, the correlation of the inter-modal data by the correlation constraint, but also preserve the overall data structure by the decoder.

The Framework of CMSC-DCCA

The proposed Cross-Modal Subspace Clustering via Deep Canonical Correlation Analysis (CMSC-DCCA) method consists of three parts: deep canonical correlation analysis (Deep CCA) model, self-expressive layer and Deep CCA decoder. The framework is as shown in Figure 2.

Deep CCA Model: In our Deep CCA model, there are two parts: convolutional encoder and correlation calculation. We assume that two modal data are $\mathbf{X}_1 = \{x_1^i\}_{i=1}^m \in \mathbb{R}^{d_1 \times m}$, $\mathbf{X}_2 = \{x_2^i\}_{i=1}^m \in \mathbb{R}^{d_2 \times m}$, where m is the number of samples. d_1 and d_2 are the corresponding dimensions of the modal 1 and the modal 2. We set two modal convolutional encoders to four layers. $\mathbf{X}_1$ and $\mathbf{X}_2$ are fed as inputs to the convolutional encoders to process and we can obtain the latent representations $(\mathbf{Z}_1|\theta_{e_1}) \in \mathbb{R}^{o \times m}$ and $(\mathbf{Z}_2|\theta_{e_2}) \in \mathbb{R}^{o \times m}$, where $\theta_{e_1}, \theta_{e_2}$ are the parameters of the convolutional encoder 1 and 2, and o is the output dimension of the convolutional encoders. Then we calculate the correlation between $\mathbf{Z}_1$ and $\mathbf{Z}_2$ with the following expression:

$$\arg\max_{\theta_{e_1},\theta_{e_2}} corr(\mathbf{Z}_1, \mathbf{Z}_2) = \arg\max_{\theta_{e_1},\theta_{e_2}} \frac{\mathrm{cov}(\mathbf{Z}_1, \mathbf{Z}_2)}{\sqrt{D(\mathbf{Z}_1)}\sqrt{D(\mathbf{Z}_2)}}, \tag{1}$$

where $corr(\cdot)$ is the correlation between $\mathbf{Z}_1$ and $\mathbf{Z}_2$. $\mathrm{cov}(\mathbf{Z}_1, \mathbf{Z}_2)$ is the covariance of $\mathbf{Z}_1$ and $\mathbf{Z}_2$, and $\sqrt{D(\mathbf{Z}_i)}$ is the variance of $\mathbf{Z}_i$, $i = 1, 2$.

Self-expressiveness Layer: Some self-expressiveness based methods (Rao et al. 2008; Elhamifar and Vidal 2013; Abavisani and Patel 2018) have received a lot of attention, which goal is to express the data point as a linear combination of other points in the same subspace. We obtain the latent representations $\mathbf{Z}_1$ and $\mathbf{Z}_2$ from two modal encoders and send them to the self-expression layer. In the same space, one data point can be represented linearly by other data points. Then we can get the equation:

$$\mathbf{Z}_i = \mathbf{Z}_i\mathbf{S}, \quad s.t., \; diag(\mathbf{S}) = 0\,, \tag{2}$$

where $\mathbf{S}$ is the self-representation coefficient matrix. Two modal share a self-expression coefficient matrix. In order to prevent the trivial solution $\mathbf{S} = \mathbf{I}$, we constraint $diag(\mathbf{S}) = 0$. Then we can leverage the matrix $\mathbf{S}$ to construct the affinity matrix by the following equation:

$$\mathbf{C} = \frac{1}{2}(|\mathbf{S}| + |\mathbf{S}^\top|). \tag{3}$$

Finally, we apply $\mathbf{C}$ for spectral clustering.

Deep CCA Decoders: The Deep CCA decoder of each modal consists of four-layer neural network: one fully connected layer and three deconvolution decoding layers, which aims to reconstruct the latent representations from the self-expression layer and maintain the structural characteristics of the original data. In our model, the outputs $\mathbf{Z_1}\mathbf{S}$ and $\mathbf{Z_2}\mathbf{S}$ from the self-expression layer are used as inputs to the two deep CCA decoders. We can obtain the outputs $(\widehat{\mathbf{X}}_\mathbf{1}|\theta_{d_1})$ and $(\widehat{\mathbf{X}}_\mathbf{1}|\theta_{d_1})$, where θ_{d_1} and θ_{d_2} are network parameters of the decoders.

Loss Function Analysis

The goal of the framework is to combine each part to learn the reliable objective function. We give the loss function analysis for each part and the final objective function. Minimize the loss function to optimize our model.

Deep CCA Loss: In the Deep CCA model, we send the data to the convolutional encoders and obtain the latent representations $\mathbf{Z}_1$ and $\mathbf{Z}_2$. We need to maximize the correlation between $\mathbf{Z}_1$ and $\mathbf{Z}_2$ to improve clustering performance. According to Equation (1), the goal is to jointly learn parameters for both θ_{e_1} and θ_{e_2}. Then we centralize data:

$$\overline{\mathbf{Z}}_1 = \mathbf{Z}_1 - \frac{1}{m}\mathbf{Z}_1\mathbf{1}, \quad \overline{\mathbf{Z}}_1 = \mathbf{Z}_2 - \frac{1}{m}\mathbf{Z}_2\mathbf{1}. \tag{4}$$

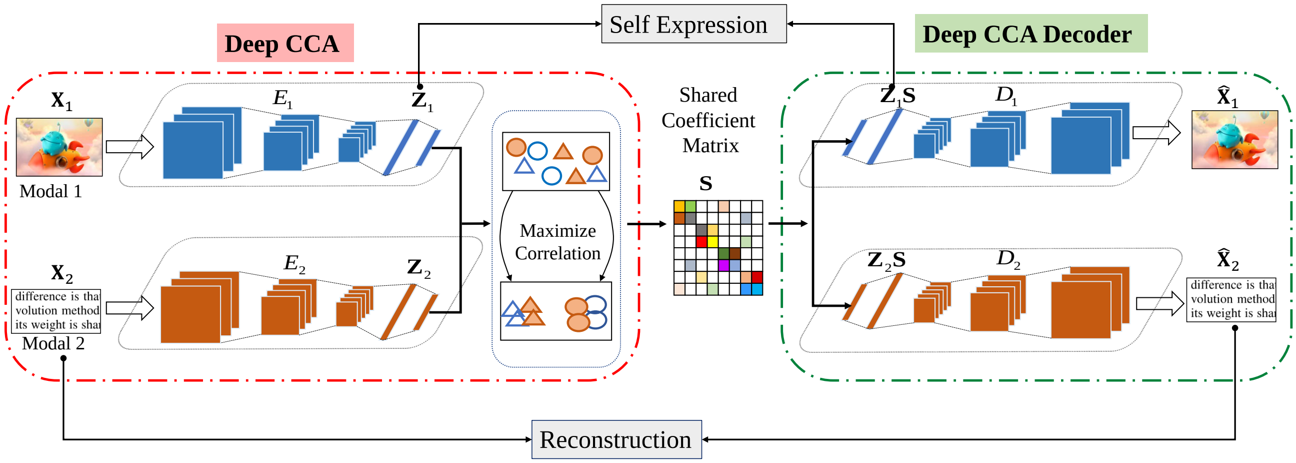


Figure 2: The Framework of Cross-Modal Subspace Clustering via Deep Canonical Correlation Analysis(CMSC-DCCA). $\mathbf{X}_1$ and $\mathbf{X}_2$ are the input data from two modal. $\widehat{\mathbf{X}}_1$ and $\widehat{\mathbf{X}}_2$ are the reconstruct data. E_1 and E_2 are the convolutional encoders. $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are latent representations from the outputs of E_1 and E_2. $\mathbf{S}$ is the shared coefficient matrix. D_1 and D_2 are decoders.

According to the calculation method of (Andrew et al. 2013), the optimization goal is:

$$\begin{aligned}\underset{\theta_{e_1},\theta_{e_2}}{\arg\max}\, corr(\mathbf{Z}_1, \mathbf{Z}_2) &= \underset{\theta_{e_1},\theta_{e_2}}{\arg\max}\, ||\mathbf{T}||_{\mathrm{tr}} \\ &= \underset{\theta_{e_1},\theta_{e_2}}{\arg\max}\, tr(\mathbf{T}^\top\mathbf{T})^{\frac{1}{2}},\end{aligned} \tag{5}$$

where $\mathbf{T} = \widehat{\mathbf{M}}_{11}^{-1/2}\widehat{\mathbf{M}}_{12}\widehat{\mathbf{M}}_{22}^{-1/2}$, $\widehat{\mathbf{M}}_{11} = \frac{1}{m-1}\overline{\mathbf{Z}}_1\overline{\mathbf{Z}}_1^\top + r_1\mathbf{I}$, $\widehat{\mathbf{M}}_{22} = \frac{1}{m-1}\overline{\mathbf{Z}}_2\overline{\mathbf{Z}}_2^\top + r_2\mathbf{I}$, and $\widehat{\mathbf{M}}_{12} = \frac{1}{m-1}\overline{\mathbf{Z}}_1\overline{\mathbf{Z}}_2^\top$. r_1, r_2 are the regularization constants. $tr(\cdot)$ is the trace function of the matrix. Then the final optimization goal for DCCA loss is:

$$loss_{\mathrm{DCCA}} = -\min_{\theta_{e_1},\theta_{e_2}} \quad tr(\mathbf{T}^\top\mathbf{T})^{\frac{1}{2}}. \tag{6}$$

Self-expression Loss: In the self-expression layer, to better perform the self-expressiveness property and acquire a better self-expression coefficient matrix $\mathbf{S}$, we minimize the self-expression loss function:

$$\begin{aligned}loss_{\mathrm{S}} = \min\ ||\mathbf{S}||_{\mathrm{F}} + \sum_{i=1}^{2} ||\mathbf{Z}_i - \mathbf{Z}_i\mathbf{S}||_{\mathrm{F}}^2, \\ s.t.,\ diag(\mathbf{S}) = 0\,.\end{aligned} \tag{7}$$

where $||\cdot||_{\mathrm{F}}$ denotes the matrix Frobenius norm.

Reconstruction Loss: In order to guarantee the effectiveness of the representations processed by the Deep CCA encoders and the self-expression layer, we add the Deep CCA decoders to reconstruct data. The representations $\mathbf{Z}_1\mathbf{S}$ and $\mathbf{Z}_2\mathbf{S}$ from the self-expression layer are fed to the decoders and we can acquire the reconstruct data $(\widehat{\mathbf{X}}_1|\theta_{d_1})$ and $(\widehat{\mathbf{X}}_2|\theta_{d_1})$. Minimize errors between reconstructed and original data to optimize the network. Therefore, the reconstruction loss for the network is:

$$loss_{\mathrm{Re}} = \min_{\theta_{d_1},\theta_{d_2}} \sum_{i=1}^{2} ||\mathbf{X}_i - \widehat{\mathbf{X}}_i||_{\mathrm{F}}^2. \tag{8}$$

```
Algorithm 1: CMSC-DCCA
Input: cross-modal data $\mathbf{X}_1$, $\mathbf{X}_2$; number of categories $K$; initialization parameters $\lambda_1$, $\lambda_2$, $\lambda_3$; learning rate $= 0.001$.
While not converged do:
    Pre-train the networks using equation (8).
    Optimize network parameters $\theta_{e_1}, \theta_{e_2}$ of encoders and $\theta_{d_1}, \theta_{d_2}$ of decoders.
End
While not converged do:
    Train the entire networks using equation (9).
    Update parameters $\theta$ including encoder parameters $\theta_{e_1}, \theta_{e_2}$ and decoder parameters $\theta_{d_1}, \theta_{d_2}$.
End
 Extract the self-representation coefficient matrix $\mathbf{S}$ from the trained networks.
 Compute the affinity matrix $\mathbf{C} = \frac{1}{2}(|\mathbf{S}| + |\mathbf{S}|^\top)$.
 Apply spectral clustering with the affinity matrix $\mathbf{C}$ to complete clustering.
Output: $\theta$, $\mathbf{S}$
```

According to analysis of the loss function for each part, We conclude that the final objective function is as follows:

$$\begin{aligned}\mathrm{Loss} = \min_{\theta}\ ||\mathbf{S}||_{\mathrm{F}} + \lambda_1 \sum_{i=1}^{2} ||\mathbf{Z}_i - \mathbf{Z}_i\mathbf{S}||_{\mathrm{F}}^2 + \\ \lambda_2 \sum_{i=1}^{2} ||\mathbf{X}_i - \widehat{\mathbf{X}}_i||_{\mathrm{F}}^2 - \lambda_3\, tr(\mathbf{T}^\top\mathbf{T})^{\frac{1}{2}}.\end{aligned} \tag{9}$$

where λ_1, λ_2, $\lambda_3 > 0$ are denoted regularization parameters. $||\cdot||_{\mathrm{F}}$ is the matrix Frobenius norm, and $tr(\cdot)$ indicates trace function. θ is all network parameters including $\theta_{e_1}, \theta_{e_2}$ and $\theta_{d_1}, \theta_{d_2}$.

Train Steps

In the proposed model, we use two steps to train the model and optimize the network parameters.

First Step: We pre-train the network using equation (8). We send the cross-modal data $\mathbf{X}_1$ and $\mathbf{X}_2$ to the deep convolutional auto-encoder E_1 and E_2, and obtain the reconstruction data $\widehat{\mathbf{X}}_1$ and $\widehat{\mathbf{X}}_2$ from the decoders D_1 and D_2. In pre-train step, we set the learning-rate to 0.001. Minimize the error between the original data and the reconstruction data to optimize the network and update encoder parameters θ_{e_1}, θ_{e_2} and decoder parameters θ_{d_1}, θ_{d_2}, where we use mean squared error (MSE) (Wang and Bovik 2009) to optimize the objective function. MSE is the expected value of the square of the difference between the parameter estimate and the true value of the parameter. And it can evaluate the degree of change of the data. The smaller the value of the MSE, the better the accuracy of the prediction model describing the experimental data.

Second Step: We train the entire network using equation (9). We use the convolutional encoders E_1, E_2 and the decoders D_1, D_2 from the first step training to train the entire network, i.e., minimizing the total Loss including deep CCA encoder loss $loss_{DCCA}$, the self-expression layer loss $loss_S$ and the reconstruction loss $loss_{Re}$ to update model parameters θ_{e_1} and θ_{e_2}, θ_{d_1} and θ_{d_2}. We obtain the shared coefficient matrix $\mathbf{S}$ from self-expression layer, and calculate the affinity matrix $\mathbf{C} = \frac{1}{2}(|\mathbf{S}| + |\mathbf{S}|^\top)$. Finally, we use the affinity matrix $\mathbf{C}$ and spectral clustering method to complete data clustering. The whole training process is summarized is shown as **Algorithm 1**.

Experiments

In order to evaluate the performance of our proposed CMSC-DCCA model, we conduct the experiments by comparing with nine remarkable baseline approaches on four datasets. Specifically, we first introduce the used four datasets, followed by the experiment results and some analysis.

Experiment Setup

Datasets Settings: The used datasets in our experiments include: 1) FRGC Dataset (Yang, Parikh, and Batra 2016) is an RGB image dataset. In our work, we randomly select 20 objects from the original dataset including 2,462 face images. We set the data size to 32 × 32, use its original RGB picture as the first modal and its corresponding gray picture as the second modal; 2) Fashion-MNIST Dataset (Xiao, Rasul, and Vollgraf 2017) is an image dataset that replaces the MNIST handwritten digit set. The dataset contains images in 10 categories: t-shirt, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag, ankle boot. The Fashion-MNIST dataset has a total of 70,000 different products and the size of each image is 28 × 28. In our work, we randomly select 200 samples per category to make the network easy to handle, and extract their edge features as the second modal; 3) YTF Dataset (Wolf, Hassner, and Maoz 2011) which is a face videos dataset is designed for studying the problem of unconstrained face recognition in videos. This dataset contains 3,425 videos of 1,595 different people. In our work, we select 41 subjects from YTF dataset, and set the size as 55 × 55. We use its original RGB picture as the first modal, and the gray picture converted from the original RGB picture as the second modal; 4) COIL-20 Dataset contains 1440 images of 20 objects, and each object is captured from varying angles with a 5 degrees interval, one image every 5 degrees and 72 images per object. In this paper, we use 1440 images and extract their edge features. The 1440 original images and the edge feature images are used as two modal data.

Table 1: The parameters of convolution encoders.

Encoders	Convolution kernel size	Stride	Padding
Encoder1	4×4	2	1
Encoder2	3×3	1	1
Encoder3	4×4	2	1

Implementation details: In our model, we use the four-layer encoders including three convolution encoding layers and a fully connected layer, and the corresponding decoders consists of a fully connected layer and three deconvolution decoding layers. More specific settings are given in Table 1.

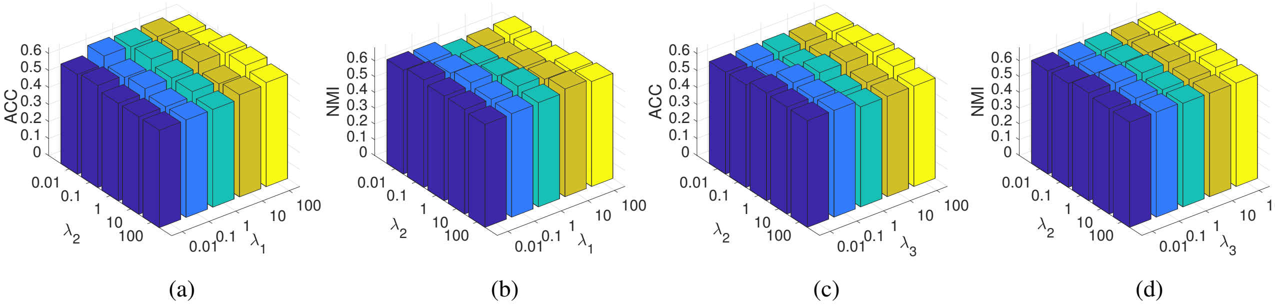
We implement our method and other non-linear methods with the public toolbox of PyTorch. We run all the experiments on the platform of Ubuntu Linux 16.04 with NVIDIA Titan Xp Graphics Processing Units (GPUs) and 64 GB memory size. We use Adam (Kingma and Ba 2015) optimizer with default parameter setting to train our model and fix the learning rate as 0.001.

Experimental Results

Comparison with Existing Approaches: To evaluate the effectiveness of the proposed method, we compare with nine algorithms, including two classic single-modal clustering methods and seven state-of-the-art multi-modal clustering methods. K-means clustering (Hartigan and Wong 1979) and Deep Embedding Clustering (DEC) (Xie, Girshick, and Farhadi 2016) are single-modal clustering methods and are regarded as the baseline algorithms for comparison. Robust Multi-View K-Means Clustering (RMKMC) (Xiao, Nie, and Huang 2013) integrates heterogeneous representations of large scale data; Binary Multi-View Clustering (BMVC) (Zhang et al. 2018) dexterously manipulates multi-view image data and easily scaled to large data; Joint Framework For Deep Multi-view Clustering (DMJC) (Lin et al. 2018) designs two ingenious variants of deep multi-view joint clustering models; Deep Multi-modal Subspace Clustering (DMSC) (Abavisani and Patel 2018) presents convolutional neural network (CNN) based approaches for unsupervised multi-modal subspace clustering; Deep Canonical Correlation Analysis (DCCA) (Andrew et al. 2013) computes representations of the two modal by passing them through multiple stacked layers of nonlinear transformation; Deep Canonically Correlated Auto-Encoders (DC-CAE) (Wang et al. 2016) optimizes the combination of canonical correlation between the learned bottle neck rep-

Table 2: The clustering accuracy rate(ACC)(%) and the normalized mutual information(NMI)(%) on four datasets.

Methods	Fashion-MNIST		COIL-20		FRGC		YTF	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
kmeans	51.27	49.99	57.49	73.22	23.62	27.12	56.01	75.23
DEC	51.80	54.60	68.00	80.25	37.80	50.50	37.10	44.60
RMKMC	53.32	52.87	60.97	74.93	23.52	25.85	57.21	74.56
BMVC	45.36	38.05	34.31	40.33	41.51	45.92	28.13	38.28
DMJC	61.41	63.41	72.99	81.58	44.07	59.79	61.15	77.40
DMSC	59.55	65.07	74.10	86.82	60.28	75.51	62.80	80.16
DCCA	52.74	53.82	55.76	64.91	22.91	24.75	45.19	60.35
DCCAE	55.95	52.93	61.60	71.56	32.33	31.22	45.57	60.15
DGCCA	56.28	57.04	54.01	62.40	23.76	24.53	47.26	61.38
CMSC-DCCA	**62.95**	**68.33**	**82.64**	**91.45**	**70.80**	**78.55**	**66.15**	**82.67**



(a) (b) (c) (d)

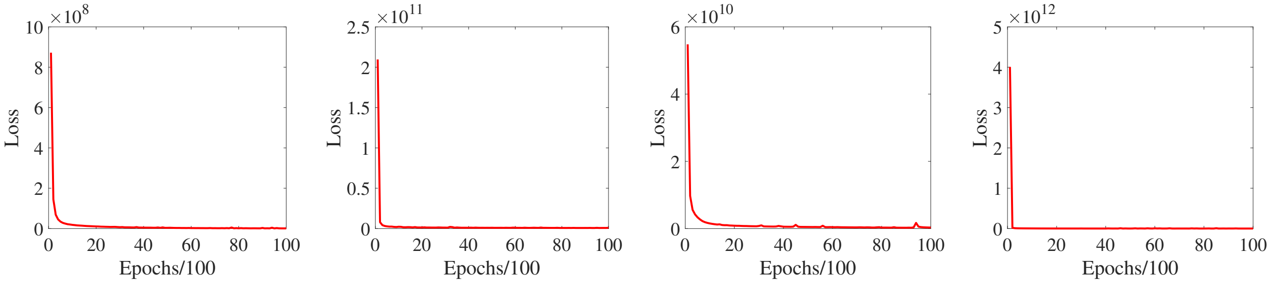
Figure 3: The effect of parameters λ_1, λ_2 and λ_3 on Fashion-MNIST dataset, where λ_1 is the regularization parameter of the self-expression error, λ_2 is the regularization parameter of the reconstruction error and λ_3 is the regularization parameter of the correlation calculation. (a) and (b) are the clustering results in terms of ACC and NMI, when fixing λ_3 and varying λ_1 and λ_2. (c) and (d) are clustering results in terms of ACC and NMI, when fixing λ_1, and varying λ_2 and λ_3.

resentations and the reconstruction errors of the autoencoders; Deep Generalized Canonical Correlation Analysis (DGCCA) (Benton et al. 2017) is a method for learning nonlinear transformations of arbitrarily multi-modal data. For DCCA and DCCAE methods, only two modal data as inputs can be sent to these models. For RMKMC, BMVC, DMJC and DMSC methods, there are mutli-modal data as inputs to these four models. For comparison, we select two modal data to do contrast experiments on four datasets.

Performance Evaluation: In order to evaluate the performance of clustering algorithms, we adopt two metrics (i.e., clustering accuracy (ACC) (Kuhn 1955), normalized mutual information (NMI) (Xu, Liu, and Gong 2003)) to measure the performance by comparing with nine baseline methods on four datasets. The correct clustering should assign the high similarity data to the same class, and different data to different classes. Therefore, the bigger the value of ACC and NMI is, the better the clustering performance of the corresponding method will be.

The clustering performance of the existing methods on four datasets are reported in Table 2. From the presented results, we can have the following observations: 1) our proposed CMSC-DCCA model can achieve the best performance on all the four datasets in terms of both ACC and NMI, which verifies the impact of improved clustering performance via the correlations among both inter-modal and intra-modal data. 2) Our proposed model significantly outperforms both K-means and DEC among most cases, e.g., for the FRGC dataset, K-means and DEC are only 23.62% and 37.80% in terms of ACC, and 37.80% and 50.50% in terms of NMI. It is because that they are single modal clustering methods, which does not consider information of other modal, and fully reflect the data characteristics. 3) The reason why DCCA obtain poor clustering performance (e.g., ACC and NMI are only 22.91% and 24.75% on the FRGC dataset) is that it cannot reconstruct data to ensure that the representations after the encoded network can still reflect the structure of the original data; since DCCAE cannot consider the relationships among intra-modal data, which cause ACC and NMI to be 32.33% and 31.22% on the FRGC dataset. Additionally, even though DMJC and DMSC which are designed for cross-modal data clustering perform better than other methods on most cases, our proposed model also has a little improvement. It is because that these two methods cannot make full use of the correlations among the inter-modal data.

Parameters Analysis: In our model, there are three regularization parameters λ_1, λ_2 and λ_3. We use the method



(a) Fashion-MNIST (b) COIL-20 (c) FRGC (d) YTF

Figure 4: The loss curve of our method on four datasets. (a), (b), (c) and (d) are the loss iteration curve on Fashion-MNIST, COIL-20, FRGC and YTF datasets. We set 10000 epochs to train the entire network and obtain a loss value for each 100 epochs.

of controlling variables to analyze the parameters. Firstly, we fix the regularization parameters of the correlation calculation λ_3, and vary the regularization parameters of the self-expression error and the reconstruction error λ_1 and λ_2 in a range $\{0.01, 0.1, 1, 10, 100\}$. Then we fix λ_1, and also vary λ_2 and λ_3 in range $\{0.01, 0.1, 1, 10, 100\}$. Since the strategies of setting parameters are the same on all the four datasets, we only show the effect of parameters on Fashion-MNIST dataset for simplicity. From the presented in Figure 3, notice that 1) our method can achieve the best ACC and NMI values on Fashion-MNIST dataset when $\lambda_1 = 0.1$, $\lambda_2 = 1$ and $\lambda_3 = 1$; 2) our method is stable since varying parameters has little influence on the clustering performance. In addition, the values of best parameters are $\lambda_1 = 1$, $\lambda_2 = 1$ and $\lambda_3 = 1$ for YTF dataset, and are $\lambda_1 = 0.1$, $\lambda_2 = 1$ and $\lambda_3 = 1$ for both FRGC and COIL-20 datasets.

Convergence Analysis: In order to investigate the convergence of our proposed model, in our subsection, we plot the corresponding loss in the Eq. (9) on four datasets. As depicted in Figure 4, the values of objective function loss decrease with respect to iterations on all the four datasts, and the values approach to be a fixed value after a few iterations (less than 20 iterations), where each iteration includes 100 epochs. Therefore, our proposed optimization algorithm is reliable and converges quickly.

Through the analysis, we can find that our cross-modal clustering performance is highly related to the following aspects: 1) The regularization parameters play a key role in clustering performance. We thus fix the best regularization parameters to training network which aims to learn better network parameters; 2) The clustering performance is closely associated with the number of the pre-training epochs. An appropriate pre-training epochs can improve overall clustering performance; 3) Data preprocessing and initialization methods may also have an impact on the clustering performance. Clustering methods can affect the clustering speed in the final stage of the network training.

Ablation Study: In this subsection, we perform ablation study on our model, even the network is hard to analyze in the background of deep neural networks. For our model which consists of three components, we conduct experiment on Fashion-MNIST dataset with different components ablation, i.e., correlation constraint in deep CCA model, a self-expressive layer and deep CCA decoders. As shown in Table 3, we can find that 1) the correlation constraint has a certain impact on clustering performance, which maximizes the correlation of the inter-modal data and obtains a better common subspace representation; 2) self-expressive layer has significant effect on the proposed model, i.e., the correlation among intra-modal data play an important role in clustering performance; 3) deep CCA decoder has the biggest impact on the proposed method, whose role is to ensure the overall structure of the data and make the encoded data reliable. These above observations indicate that all the three components in our proposed CMSC-DCCA model are designed reasonably.

Table 3: Ablation Study on Fashion-MNIST dataset in terms of ACC (%) and NMI (%).

Methods	ACC	NMI
Without correlation constraint	60.25	61.02
Without self-expression	55.75	56.01
Without decoder	50.50	52.77
CMSC-DCCA	**62.95**	**68.33**

Conclusions

We propose a novel clustering method named Cross-Modal Subspace Clustering via Deep Canonical Correlation Analysis (CSMC-DCCA). We maximize the correlation of the inter-modal data by correlation constraint to make the data with high similarity be better grouped into one class and make full use of the information of the intra-modal data by the self-expressive layer. We construct the shared subspace coefficient matrix based on the self-expression layer and the correlation constraint. At the same time, we reconstruct the data by the decoder to ensure the overall structure of the data. Then we optimize the objective function by training the entire network and apply the spectral clustering method to implement clustering. Our experiments demonstrate that the proposed method provides significant improvement over the several state-of-the-art clustering methods.

References

Abavisani, M., and Patel, V. M. 2018. Deep multimodal subspace clustering networks. *IEEE Journal of Selected Topics in Signal Processing* 12(6):1601–1614.

Akaho, S. 2006. A kernel method for canonical correlation analysis. *arXiv preprint arXiv:cs/0609071*.

Akata, Z.; Thurau, C.; and Bauckhage, C. 2011. Non-negative matrix factorization in multimodality data for segmentation and label prediction. *CVWW*.

Andrew, G.; Arora, R.; Bilmes, J. A.; and Livescu, K. 2013. Deep canonical correlation analysis. In *ICML*, 1247–1255.

Benton, A.; Khayrallah, H.; Gujral, B.; Reisinger, D. A.; Zhang, S.; and Arora, R. 2017. Deep generalized canonical correlation analysis. *arXiv preprint arXiv:1702.02519*.

Chaudhuri, K.; Kakade, S. M.; Livescu, K.; and Sridharan, K. 2009. Multi-view clustering via canonical correlation analysis. In *ICML*, 129–136.

Elhamifar, E., and Vidal, R. 2013. Sparse subspace clustering: Algorithm, theory, and applications. *TPAMI* 35(11):2765–2781.

Guo, D.; Zhang, J.; Liu, X.; Cui, Y.; and Zhao, C. 2014. Multiple kernel learning based multi-view spectral clustering. In *ICPR*, 3774–3779.

Hartigan, J. A., and Wong, M. A. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society* 28(1):100–108.

He, R.; Zhang, M.; Wang, L.; Ji, Y.; and Yin, Q. 2015. Cross-modal subspace learning via pairwise constraints. *TIP* 24(12):5543–5556.

Ji, P.; Zhang, T.; Li, H.; Salzmann, M.; and Reid, I. D. 2017. Deep subspace clustering networks. In *NeurIPS*, 24–33.

Jia, Y.; Salzmann, M.; and Darrell, T. 2011. Learning cross-modality similarity for multinomial data. In *ICCV*, 2407–2414.

Jin, C.; Mao, W.; Zhang, R.; Zhang, Y.; and Xue, X. 2015. Cross-modal image clustering via canonical correlation analysis. In *AAAI*, 151–159.

Kim, T.; Kittler, J.; and Cipolla, R. 2007. Discriminative learning and recognition of image set classes using canonical correlations. *TPAMI* 29(6):1005–1018.

Kingma, D. P., and Ba, J. L. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kuhn, H. W. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2(1):83–97.

Lin, B.; Xie, Y.; Qu, Y.; and Li, C. 2018. Deep multi-view clustering via multiple embedding.

Nie, F.; Cai, G.; and Li, X. 2017. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI*, 2408–2414.

Patel, V. M.; Van Nguyen, H.; and Vidal, R. 2015. Latent space sparse and low-rank subspace clustering. *IEEE J-STSP* 9(4):691–701.

Rao, S. R.; Tron, R.; Vidal, R.; and Ma, Y. 2008. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *CVPR*, 1–8.

Rasiwasia, N.; Pereira, J. C.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 251–260.

Sun, T., and Chen, S. 2007. Locality preserving cca with applications to data visualization and pose estimation. *Image and Vision Computing* 25(5):531–543.

Wang, Z., and Bovik, A. C. 2009. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE* 26(1):98–117.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. A. 2016. On deep multi-view representation learning: Objectives and optimization. *arXiv: Learning*.

Wolf, L.; Hassner, T.; and Maoz, I. 2011. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 529–534.

Xia, R.; Pan, Y.; Du, L.; and Yin, J. 2014. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *AAAI*, 2149–2155.

Xiao, C.; Nie, F.; and Huang, H. 2013. Multi-view k-means clustering on big data. In *IJCAI*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv: Learning*.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*, 478–487.

Xu, W.; Liu, X.; and Gong, Y. 2003. Document clustering based on non-negative matrix factorization. In *ACM SIGIR*, 267–273.

Yang, J.; Parikh, D.; and Batra, D. 2016. Joint unsupervised learning of deep representations and image clusters. In *CVPR*, 5147–5156.

Yin, Q.; Wu, S.; He, R.; and Wang, L. 2015. Multi-view clustering via pairwise sparse subspace representation. *Neurocomputing* 156:12–21.

Zhang, Z.; Liu, L.; Shen, F.; Shen, H. T.; and Shao, L. 2018. Binary multi-view clustering. *TPAMI* PP(99):1–1.

Zhang, H.; Zhuang, Y.; and Wu, F. 2007. Cross-modal correlation learning for clustering on image-audio dataset. In *ACM MM*, 273–276.