Deep Multimodal Clustering with Cross Reconstruction

Xianchao Zhang^{1,2}, Xiaorui Tang^{1,2}, Linlin Zong^{1,2}(✉), Xinyue Liu^{1,2},
and Jie Mu^{1,2}

¹ School of Software, Dalian University of Technology, Dalian 116620, China
llzong@dlut.edu.cn

² Key Laboratory for Ubiquitous Network and Service Software
of Liaoning Province, Dalian 116024, China

Abstract. Recently, there has been surging interests in multimodal clustering. And extracting common features plays a critical role in these methods. However, since the ignorance of the fact that data in different modalities shares similar distributions in feature space, most works did not mining the inter-modal distribution relationships completely, which eventually leads to unacceptable common features. To address this issue, we propose the deep multimodal clustering with cross reconstruction method, which firstly focuses on multimodal feature extraction in an unsupervised way and then clusters these extracted features. The proposed cross reconstruction aims to build latent connections among different modalities, which effectively reduces the distribution differences in feature space. The theoretical analysis shows that the cross reconstruction reduces the Wasserstein distance of multimodal feature distributions. Experimental results on six benchmark datasets demonstrate that our method achieves obviously improvement over several state-of-arts.

Keywords: Deep learning · Unsupervised learning · Multimodal clustering

1 Introduction

Clustering is a vital research topic in data science and machine learning. Multimodal clustering is an important field in clustering and has made great progress. Multimodal clustering aims to divide the multimodal data information into different clusters in an unsupervised manner. The existing works usually are based on spectral clustering [14], subspace clustering [1], deep clustering [23] etc. On account of the amazing performance for feature extraction and dimensionality reduction tasks, deep clustering receives much attention in recent years [12, 17, 20, 23]. For deep multimodal clustering, the most common method extracts common features from different modalities [19] by using multiple deep neural networks (DNN) and clusters on the common features.

This work was supported by National Science Foundation of China (No. 61876028; No. 61632019; No. 61972065; No. 61806034).

© Springer Nature Switzerland AG 2020

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12084, pp. 305–317, 2020.

https://doi.org/10.1007/978-3-030-47426-3_24

Fig. 1. The two stages of DMCR framework: The left part illustrates the first stage. Two autoencoders extract features from two modalities, VIB is used both in E_1 and E_2 . $\hat{x}^1$ are reconstructed from z^1 , $\hat{x}^2$ are reconstructed from z^2 . The dotted lines illustrate the cross reconstruction method, $\hat{x}^1$ are reconstructed from z^2 , $\hat{x}^2$ are reconstructed from z^1 . The right part illustrates the second stage. The fusion layers fuse the features z^1 and z^2 to common features z^* , which are used for clustering.

So far, the deep neural networks (DNN) based multimodal clustering methods can be divided into three categories: autoencoder based methods [16], Deep Boltzmann Machine (DBM) based methods [19] and deep canonical correlation analysis (DCCA) [3] based methods [21]. The autoencoder based methods use autoencoders to extract common features of different modalities and choose common features that can best reconstruct the input data [16, 21]. However, autoencoders do not really discover the similarity of the common feature distributions. The DBM based methods [19] learn a joint representation of different modalities by DBM. But due to the high computational costs in high-dimensional data space [11], the DBM based methods have not been widely studied in recent years. The DCCA based methods [21] learn features that are most correlated from different modalities by canonical correlation analysis (CCA). Like autoencoder based methods, DCCA based methods use autoencoders to extract features from different modalities, called deep canonically correlated autoencoder (DCCAE) [21]. The difference is that DCCAE further optimizes the canonical correlation among features of different modalities. But DCCA based methods lack the analysis of probability theory, which makes it difficult to measure the distribution differences of different modalities. Moreover, data of different modalities contain different numerical characteristics, and may not show obvious correlation. In this case, the depth typical correlation analysis may not be effective.

In this paper, we focus on multimodal clustering by extracting the common features of multimodal data in an unsupervised way, and reduce the distribution differences of different modalities in feature space. Firstly, we apply Variational Information Bottleneck (VIB) [2] to extract features from different modalities. By minimizing mutual information between raw data and the extracted features, VIB can control the amount of information owing through the network when extracting features. And due to the mutual information, VIB also provides explicit probabilistic analysis on feature space. Secondly, we apply cross recon-

struction method during extracting features from different modalities, which can effectively reduce the distribution differences of different modalities in feature space. We also provide theoretical analysis to prove the similarity of multimodal distributions. Thirdly, we fuse the extracted features to common features using fusion layers. Finally, we cluster the common features using clustering algorithm. The entire process above constitute our deep multimodal clustering with cross reconstruction (DMCR).

The contributions of this work are summarized as below: (1) We propose a novel deep multimodal clustering algorithm, which can effectively reduce the distribution differences among different modalities in feature space. (2) We provide a theoretical analysis to prove that the proposed cross reconstruction method effectively reduce the distribution difference of different modalities in feature space. (3) Experiments show obviously improvement over state-of-the-art multimodal clustering methods on six benchmark multimodal datasets.

2 Related Work

2.1 Deep Clustering

The existing deep clustering methods are roughly divided into two categories: two-stage methods and end-to-end methods [21].

The two-stage methods first extract features of the data by deep learning method, and finally apply the clustering methods to the features. Tian et al. [20] uses autoencoder to extract the features of graph and finally uses k-means to cluster. Chen [7] applies Deep Belief Network (DBN) to extract features and finally uses non-parametric maximum-margin clustering to cluster the features.

The end-to-end methods jointly optimize the feature extraction and clustering. The joint unsupervised learning (JULE) algorithm [24] uses a recurrent framework for joint unsupervised learning of deep representations and image clustering, which are optimized jointly in training process. The deep embedding clustering (DEC) algorithm [23] clusters a set of data points in a jointly optimized feature space. Based on DEC, the improved deep embedding clustering (IDEC) algorithm [12] jointly optimization and preserve local structure of data generating distribution.

2.2 Multimodal Clustering

Based on the basic algorithms used in the multimodal methods, the existing multimodal clustering are roughly divided into two categories: the traditional clustering based methods [5, 6, 10, 22, 25] and deep clustering based methods [3, 16, 19, 21].

The traditional clustering based methods learn a consensus matrix or minimize the divergence of multiple views simultaneously. For example, the multi-view spectral clustering (MMSC) algorithm [5] learns a commonly shared graph Laplacian matrix by unifying different modalities. Gao et al. [10] proposes a

novel NMF-based multi-view clustering algorithm by searching for a factorization that gives compatible clustering solutions across multimodal. The diversity-induced multi-view subspace clustering (DIMSC) algorithm [6] extend the existing subspace clustering into the multimodal domain. The low-rank tensor constrained multi-view subspace clustering (LT-MSC) algorithm [25] introduces a low-rank tensor constraint to explore the complementary information from multimodal data. The exclusivity-consistency regularized multi-view subspace clustering (ECMSC) algorithm [22] attempts to harness the complementary information between different representations by introducing a novel position-aware exclusivity term.

Deep clustering based methods firstly jointly learn low-dimensional features from multimodal data, and then cluster the features. Ngiam et al. [16] proposes a series of frameworks for deep multimodal learning based on autoencoders. Srivastava and Salakhutdinov [19] proposes a deep multimodal representation learning framework, which learns a joint representation of different modalities by DBM. The DCCA [3] learns complex nonlinear transformations of two modalities of data such that the resulting representations are highly linearly correlated. The DCCAE [21] add an autoencoder regularization term to DCCA.

3 The Proposed Algorithm

In this section, we introduce our DMCR algorithm in detail. Consider the problem of clustering a set of n points $X = \{x_1, x_2, \dots, x_n\}$ into k clusters $\{c_1, c_2, \dots, c_k\}$, each data point x_i contains m modalities $x_i = \{x_i^1, x_i^2, \dots, x_i^m\}$. These modalities have different dimensions, i.e. $x_i^1 \in R^{d_1}$, $x_i^2 \in R^{d_2}, \dots, x_i^m \in R^{d_m}$. Data in the same modality have the same dimensions. Multimodal clustering partitions data in m modalities into k clusters.

Figure 1 illustrates the framework of DMCR for two modalities. The DMCR has two stages. In the first stage, we extract features from different modalities using VIB and cross reconstruction to constrain encoders, which ensures the extracted features of different modalities sharing similar distributions. In the first stage of Fig. 1, E_1 and D_1 are the encoder and decoder used for the first modality, E_2 and D_2 are the encoder and decoder used for the second modality; z^1 and z^2 are features extracted from x^1 and x^2 ; VIB in E_1 and E_2 denote the VIB regularization terms; $\hat{x}^1$ and $\hat{x}^2$ are data reconstructed from z^1 and z^2 ; $\hat{x}^1$ and $\hat{x}^2$ are data reconstructed from cross reconstruction; In the second stage, we fuse the features to common features by fusion layers, then we cluster these fused common features. In the second stage of Fig. 1, the fusion layers fuse features from different modalities to common features, z^* is the common features. We describe our algorithm next.

3.1 Multimodal Feature Extraction

Multimodal data contain modality-unique features and modality-common features. It is difficult to extract the modality-common features directly from

different modalities with traditional encoders. We first use the Variational Information Bottleneck (VIB) [2] to extract the modality-common features among different modalities. And then apply Cross Reconstruction method to ensure the modality-common features of different modalities satisfy similar distribution.

Multimodal Feature Extraction with VIB. We adopt deep autoencoders as a features extractor. For a given input x^i in the i -th modality, the encoder aims to get a feature z^i in a low-dimensional space, and the decoder aims to reconstruct the input from the latent representation. So, the goal of the deep autoencoder is to get a good reconstruction $\hat{x}^i$ for input x^i .

But autoencoders cannot control the amount of information contained in extracted features, which makes it difficult to extract modality-common features from each modality.

Given that VIB is able to control the scale of feature information, we use the VIB [2] regularization term in the encoders to eliminate unique features and extract the common features. The loss function of extracting features of the i -th modality is:

$$\min_{\theta^i, \varphi^i} E_{x^i \sim p(x^i)} [-E_{z^i \sim p(z^i|x^i; \theta^i)} [\log q(\hat{x}^i|z^i; \varphi^i)] + \beta KKL(p(z^i|x^i; \theta^i) || q(z^i))], \quad (1)$$

where $p(z^i|x^i; \theta^i)$ denotes the encoder for the i -th modality; θ^i is the parameter of the encoder; $q(x^i|z^i; \varphi^i)$ denotes the decoder for the i -th modality; z^i are generated with reparameterization trick [2, 13]; φ^i is the parameter of the decoder; β controls the weight of the VIB regularization term. The first term $-E_{z^i \sim p(z^i|x^i; \theta^i)} [\log q(\hat{x}^i|z^i; \varphi^i)]$ in Eq. 1 is the reconstruction loss of autoencoder for the i -th modality. The second term in Eq. 1 is the VIB regularization loss for the i -th modality.

Cross Reconstruction. In the last section, we only extract modality-common features from each modality using VIB. In multimodal learning tasks, the basic task is mining the relationships among different modalities. Previous work, such as DCCAE [21] learn complex nonlinear transformations of two modalities of data such that the resulting representations are highly linearly correlated. But data of different modalities have different numerical characteristics, correlation constraints on different modalities may failed to capture the statistical properties of implicit features. Considering the fact that modality-common features of different modalities share similar distributions, we propose a cross reconstruction method to promote the features extraction process.

The dotted lines in Fig. 1 represent the our cross reconstruction process. The decoders D_1 and D_2 both take z^1 and z^2 as inputs simultaneously. Specifically, the green dotted lines in Fig. 1 represent reconstructing x^1 with z^2 using D_1 and the red dotted lines in Fig. 1 represent reconstructing x^2 with z^1 using D_2 . The loss function of reconstructing x^j with z^i ($j \neq i$) is:

$$\min_{\theta^i, \varphi^j} E_{x^i \sim p(x^i)} [-E_{z^i \sim p(z^i|x^i; \theta^i)} [\log q(\hat{x}^j|z^i; \varphi^j)]]. \quad (2)$$

where $q(\hat{x}^j|z^i; \varphi^j)$ denotes the decoder used for reconstructing x^j with z^i . Note that, it is also the same decoder for reconstructing x^j with z^j .

As a regularization term, cross reconstruction restrains the similarity of different modalities in distribution, which will analysed later.

The Overall Loss Function. The complete loss function of extracting feature of the i -th modality is:

$$\begin{aligned} \min E_{x^i \sim p(x^i)} [& [-E_{z^i \sim p(z^i|x^i; \theta^i)} [\log q(\tilde{x}^i|z^i; \varphi^i)] + \beta KL(p(z^i|x^i; \theta^i) || q(z^i)) \\ & + \gamma [\sum_{\substack{M, j \neq i \\ j=1}} -E_{z^i \sim p(z^i|x^i; \theta^i)} [\log q(\hat{x}^j|z^i; \varphi^j)]]], \end{aligned} \quad (3)$$

where j represents a modality except the i -th modality; the last term of Eq. 3 is cross reconstruction regularization loss of reconstructing x^j with z^i ; γ controls the weight of the cross reconstruction regularization term.

Algorithm 1. Training Process of Deep Multimodal Clustering with Cross Reconstruction.

Input:

```

    The number of modality  $m$ ;
    The dataset of  $n$  data points for each modality  $i$ :  $x^i = \{x_1^i, x_2^i, \dots, x_n^i\}$ ,  $i \in [1, m]$ .
1: for number of training iterations do
2:   Sampling a minibatch of samples from each modality;
3:   for each  $i \in [1, m]$  do
4:      $\mu(x^i), \sigma(x^i) = E_i(x^i)$ ;
5:      $z^i = \text{Reparameterization}(\mu(x^i), \sigma(x^i))$ ;
6:      $\tilde{x}^i = D_i(z^i)$ ;
7:     for each  $j \in [1, m] \setminus i$  do
8:        $\hat{x}^j = D_j(z^i)$ ;
9:     end for
10:    Training  $E_i$  with Eq.3;
11:   end for
12: end for
13: for number of training iterations do
14:   Sampling a minibatch of samples from each modality;
15:   for each  $i \in [1, m]$  do
16:      $\mu(x^i), \sigma(x^i) = E_i(x^i)$ ;
17:      $z^i = \text{Reparameterization}(\mu(x^i), \sigma(x^i))$ ;
18:   end for
19:    $z^* = \text{Fusion}(z^1, z^2, \dots, z^m; \eta)$ ;
20:   Training the fusion layer with Eq.5;
21: end for
```

3.2 Feature Fusion

After extracting features from different modalities, we fuse these features to common features, then cluster on the common features. As shown in the second stage of Fig. 1, we use fusion layers to fuse extracted features. The fusion layers consist of fully connected layers, and z^* is the fused common features:

$$z^* = \text{Fusion}(z^1, \dots, z^m; \eta), \quad (4)$$

where η is the parameter of the fusion layers. We use L2 loss between z^* and extracted features of different modalities to train the fusion layers:

$$\min_{\eta} \sum_{i=1}^m \|z^* - z^i\|_2. \quad (5)$$

3.3 The DMCR Algorithm

In this section, we describe the training process and clustering process of DMCR. The entire training process has two steps: (1) Training the autoencoders with Eq. 3. (2) Training the fusion layers according to Eq. 5. The training process of DMCR is summarized in Algorithm 1.

After training DMCR, we get the common features associated with multimodal data. Then we cluster the common features. Here, we choose k -means as the final clustering algorithm. The clustering process of DMCR is summarized in Algorithm 2.

Algorithm 2. Clustering Process of Deep Multimodal Clustering with Cross Reconstruction.

Input:

- The number of modality m ;
- The dataset of n data points for each modality i : $x^i = \{x_1^i, x_2^i, \dots, x_n^i\}$, $i \in [1, m]$;
- The number of clusters k .

Output:

- Cluster assignments for n multimodal data points: $C = \{c_1, c_2, \dots, c_n\}$, $c_i \in [1, k]$, $i \in [1, n]$.
 - 1: **for** each $i \in [1, m]$ **do**
 - 2: $\mu(x^i), \sigma(x^i) = E_i(x^i)$;
 - 3: $z^i = \text{Reparameterization}(\mu(x^i), \sigma(x^i))$;
 - 4: **end for**
 - 5: $z^* = \text{Fusion}(z^1, z^2, \dots, z^m; \eta)$;
 - 6: $C = k\text{-means}(z^*, k)$.
-

4 Theoretical Analysis

As mentioned previously, cross reconstruction can build implicit connection among different modalities. In this section, we explore the probability theory of connection among different modalities built by cross reconstruction method. Here we assume $q(\hat{x}^i|z^i; \varphi^i)$ to be a Gaussian distribution in the i -th modality with mean $\mu_i(z^i)$ and variance $\sigma_i^2(z^i)$, $q(\hat{x}^i|z^j; \varphi^i)$ to be a Gaussian distribution in the i -th modality with mean $\mu_j(z^j)$ and variance $\sigma_j^2(z^j)$. And then we can further derive the loss function of cross reconstruction:

$$\begin{aligned} L_i &= E_{x^i \sim p(x^i)} [-E_{z^i \sim p(z^i|x^i; \theta^i)} [\log q(\hat{x}^i|z^i; \varphi^i)]] \\ &= E_{x^i \sim p(x^i)} [E_{z^i \sim p(z^i|x^i; \theta^i)} [\frac{1}{2} \log 2\pi + \frac{1}{2} \log \sigma_i^2(z^i) + \frac{(x^i - \mu_i(z^i))^2}{2\sigma_i^2(z^i)}]], \\ L_j &= E_{x^j \sim p(x^j)} [-E_{z^j \sim p(z^j|x^j; \theta^j)} [\log q(\hat{x}^i|z^j; \varphi^i)]] \\ &= E_{x^j \sim p(x^j)} [E_{z^j \sim p(z^j|x^j; \theta^j)} [\frac{1}{2} \log 2\pi + \frac{1}{2} \log \sigma_j^2(z^j) + \frac{(x^i - \mu_j(z^j))^2}{2\sigma_j^2(z^j)}]], \end{aligned}$$

where L_i represents the loss of reconstructing x^i with z^i ; L_j represents the loss of reconstructing x^j with z^j .

Combining L_i and L_j , both $\mu_i(z^i)$ and $\mu_j(z^j)$ reduce the difference from x^i , which means that the difference between $\mu_i(z^i)$ and $\mu_j(z^j)$ also decrease. Note that both $\mu_i(z^i)$ and $\mu_j(z^j)$ are outputs of decoder D_i , so the input of D_i , i.e. z^i and z^j are similar. z^i and z^j are generated with reparameterization trick:

$$\begin{aligned} z^i &= \text{Reparameterization}(\mu(x^i), \sigma(x^i)), \\ z^j &= \text{Reparameterization}(\mu(x^j), \sigma(x^j)), \end{aligned}$$

where $\mu(x^i)$ and $\sigma(x^i)$ are the outputs of E_i , $\mu(x^j)$ and $\sigma(x^j)$ are the outputs of E_j ; z^i and z^j are randomly sampled from Gaussian distribution $N_i(\mu(x^i), \sigma(x^i))$ and $N_j(\mu(x^j), \sigma(x^j))$ respectively. So the Wasserstein distance [4] between N_i and N_j also decrease:

$$W(N_i, N_j) = \inf_{\epsilon \in \Pi(N_i, N_j)} E_{(z_i, z_j) \sim \epsilon} [\|z_i - z_j\|], \quad (6)$$

where $\Pi(N_i, N_j)$ denotes the set of all joint distributions where the marginals of $\epsilon(z_i, z_j)$ are N_i and N_j respectively. Therefore, under the constraints of cross reconstruction, these encoders will reduce the distribution differences of multi-modal features. So we prove that the cross reconstruction constrain extracted features to share similar distributions in feature spaces of different modalities.

5 Experiments

5.1 Description of Datasets

We test our model on six multimodal datasets: Digits¹, CNN², AwA³, Cal101⁴, LUse-21⁵ and Scene-15 [9]: Digits contains three modalities of 2000 samples belonging to 10 clusters. The three modalities respectively have 76, 216 and 240 dimensions. CNN is a news dataset that contains two modalities of 2107 samples belonging to 7 clusters. The first modality consists of text contents, the second modality contains the images of articles. AwA contains three modalities of 5814 samples belonging to 10 clusters. These three modalities are local self-similarity features, SIFT features and SURF features. Cal101, LUse-21 and Scene-15 contain three modalities: we extract LBP, GIST and CENTRIST descriptors from these datasets as three modalities. Cal101 contains 712 samples and these samples belong to 10 clusters. LUse-21 contains 2100 samples belonging to 21 clusters. Scene-15 contains 3000 samples assigned to 15 clusters.

¹ <https://archive.ics.uci.edu/ml/datasets/Multiple+Features>.

² <https://sites.google.com/site/qianmingjie/home/datasets/cnn-and-fox-news>.

³ <https://cvml.ist.ac.at/AwA/>.

⁴ http://www.vision.caltech.edu/Image_Datasets/Caltech101/.

⁵ <http://weegee.vision.ucmerced.edu/datasets/landuse.html>.

Table 1. Clustering ACC (%)

	DEC	IDEC	JULE	DCCAE	MMSC	RMKMC	LT-MSC	DIMSC	ECMSC	DMC	DMCR
Digits	44.07	46.70	76.36	66.19	54.00	73.30	73.35	52.10	77.40	85.55	90.75
CNN	36.96	31.69	40.00	34.27	32.02	34.21	21.26	21.64	25.34	55.52	66.68
AwA	16.10	20.08	23.05	24.31	20.28	23.17	24.17	20.79	22.19	26.23	28.67
Cal101	35.74	43.54	63.20	62.85	49.02	51.97	61.35	39.47	53.90	62.50	66.57
LUse-21	12.34	25.14	27.14	23.72	20.38	27.24	30.86	22.81	25.29	30.14	31.62
Scene-15	16.97	25.03	38.10	34.95	18.20	40.03	43.93	23.87	42.70	44.60	54.00

5.2 Comparing Methods

We compare the proposed DMCR algorithm with the following baselines: (1) Single modal clustering: DEC [23], IDEC [12] and JULE [24]. We test these methods on each modality data and take the best result as their final result. Multimodal clustering: MMSC [5], RMKMC [10], DIMSC [6], LT-MSC [25], ECMSC [22] and DCCAE [21]. Among these multimodal clustering methods, DCCAE is a two-modal method. In order to extend DCCAE to multimodal clustering task, we combine every two of the multiple modalities, and take the average results as the final result. (3) The simplified DMCR: the DMCR without cross reconstruction regularization term, called DMC.

Table 2. Clustering NMI (%)

	DEC	IDEC	JULE	DCCAE	MMSC	RMKMC	LT-MSC	DIMSC	ECMSC	DMC	DMCR
Digits	35.11	44.46	70.25	67.59	57.21	70.74	70.40	39.25	71.21	74.98	83.17
CNN	20.79	13.28	18.44	14.27	11.31	13.07	13.84	12.67	18.87	37.74	50.93
AwA	4.10	6.73	8.33	10.13	1.81	9.23	9.24	7.61	7.33	10.20	12.72
Cal101	22.63	31.00	54.12	57.96	45.55	50.49	53.41	33.74	51.53	57.51	64.31
LUse-21	15.18	28.02	34.99	32.05	26.56	30.55	35.62	21.69	25.93	39.17	40.39
Scene-15	16.10	19.92	35.84	37.82	14.44	39.03	41.36	19.39	41.16	40.58	55.55

5.3 Model and Parameter Settings

The model and parameter settings of our experiments are follows: (1) We keep these parameter settings of comparing methods as the original papers. During training, we fine-tune the parameters of these methods to get the best performance as the final result. (2) We use three autoencoders to handle three modalities and two autoencoders for two modalities. The encoders and decoders are composed of fully connected layers. We use sigmoid as the activation function in the last layer of decoders, and use ReLU activation in the other layers of encoders and decoders. The parameters of our method are randomly initialized and we set the learning rate of Adam to 0.001. (3) We note that the k -means, which is the final step of DMCR, can be replaced by other clustering algorithms. But considering the interpretation of the Euclidean distance in the feature space as diffusion distance in the input space [8, 15, 18], we choose k -means as the final clustering algorithm.

5.4 Experiment Results

Clustering Results. We evaluate our approach on three metrics of Accuracy (ACC), Normalized Mutual Information (NMI), and Purity. The experiment results of clustering ACC, NMI and Purity are summarized in Table 1, Table 2 and Table 3. The best results are marked in bold.

Firstly, we compare DMCR with the single modal algorithms DEC, IDEC and JULE. DMCR performs better than DEC and IDEC on each dataset in terms of ACC, NMI and Purity. DMCR outperforms JULE in most cases in terms of ACC, NMI and Purity. As exceptional case, DMCR performs worse than JULE on Cal101 dataset in terms of Purity. The Purity of JULE only increased by 4%, however, the NMI of JULE is decreased by 10%. Generally, it can be seen that DMCR outperforms the single modal algorithms when clustering multimodal data. The results indicate that it is reasonable to ensemble multimodal.

Secondly, we compare DMCR with multi-modal methods MMSC, RMKMC, DIMSC, LT-MSC, ECMSC and DCCAE. DMCR outperforms MMSC, RMKMC, DIMSC and ECMSC on each dataset in terms of ACC, NMI and Purity. DMCR outperforms LT-MSC and DCCAE in most cases. Taking NMI for example, the NMI of DMCR raises 12% on the Digits dataset, 3% on the AWA dataset, 13% on the Cal101 dataset, 10% on the LUse-21 dataset, 14% on the LUse-21 dataset. As exceptional cases, LT-MSC achieves a better Purity on Cal101 which is about 0.1% higher than DMCR, and DCCAE achieves a better Purity on Cal101 which is about 2% higher than DMCR. However, the NMI of LT-MSC is decreased by 10%, and the NMI of DCCAE is decreased by 7%. Generally, DMCR outperforms the other multimodal methods.

Finally, we compare DMCR with DMC, we can find that DMCR outperforms DMC, especially on Digits, CNN and Scene-15, which is more than 5% higher than DMC. That proves that the cross reconstruction regularization is effective for extracting common features of different modalities. From these tables, we can also observe that DMC also demonstrates strong competitiveness compared to other models. That proves that VIB is a effective feature extraction method that is universally applicable to different datasets. Note that DEC and IDEC just use an autoencoder without VIB, and do not perform well on multimodal datasets.

In summary, it can be concluded that our method performs the best on the multimodal datasets. The proposed cross reconstruction regularization improves the results of multimodal clustering, which further proves that it is beneficial to establish a connection among different modalities.

Parameter Setting Results. In Table 4, we explore the effect of the parameters β and γ to clustering performance of DMCR on each dataset. Due to space limitation, we only present the experimental results on the Digits dataset and Cal101 dataset in this paper. Both β and γ vary in the set $[0, 0.5, 1]$ and the best results are marked in bold.

Table 3. Clustering Purity (%)

	DEC	IDEC	JULE	DCCAE	MMSC	RMKMC	LT-MSC	DIMSC	ECMSC	DMC	DMCR
Digits	47.97	49.00	77.44	69.92	58.25	76.40	77.35	53.50	78.13	85.55	90.75
CNN	40.84	37.66	40.80	35.19	32.21	34.87	25.63	24.54	25.49	56.95	61.70
AwA	17.96	21.87	23.58	25.68	20.64	24.70	27.35	23.89	23.24	27.95	31.84
Cal101	37.88	47.19	68.68	66.29	54.21	58.85	64.58	47.33	61.94	62.50	64.46
LUse-21	14.29	27.52	30.57	27.49	21.81	29.00	32.24	24.43	28.33	34.43	35.05
Scene-15	17.65	26.43	38.97	38.69	18.60	41.40	44.70	25.87	44.80	46.57	56.60

Table 4. Parameter Setting results on Digits and Cal101 (%)

	Digits			Cal101		
	ACC	NMI	Purity	ACC	NMI	Purity
$\beta = 0, \gamma = 0$	69.05	68.05	73.55	54.07	47.98	55.76
$\beta = 0.5, \gamma = 0$	81.20	72.84	81.20	57.16	50.39	59.55
$\beta = 1, \gamma = 0$	85.55	74.98	85.55	62.50	57.51	62.50
$\beta = 0, \gamma = 0.5$	81.85	80.53	82.25	60.67	58.51	63.62
$\beta = 0.5, \gamma = 0.5$	90.35	82.46	90.35	63.76	63.68	63.34
$\beta = 1, \gamma = 0.5$	90.75	83.17	90.75	66.57	64.31	64.46
$\beta = 0, \gamma = 1$	80.50	81.85	84.05	58.00	59.88	63.34
$\beta = 0.5, \gamma = 1$	90.25	82.60	90.25	62.92	61.08	59.26
$\beta = 1, \gamma = 1$	90.55	82.02	90.25	58.14	58.57	61.65

The value of β and γ separately reflects how much we want to enforce the cross reconstruction regularization and the VIB regularization. The $\beta = 1, \gamma = 0$ stands for DMCR without cross reconstruction regularization, the $\beta = 1, \gamma = 0.5$ stands for DMCR without VIB regularization, and the $\beta = 1, \gamma = 0$ stands for DMCR without both cross reconstruction regularization and VIB regularization. It can be seen that the performance of DMCR without one regularization is better than that of DMCR without both regularization, but is worse than that of DMCR with both regularization, which proves the validity of the VIB regularization and cross reconstruction regularization. Furthermore, as shown in the tables, we get the best results when $\beta = 1$ and $\gamma = 0.5$.

6 Conclusion

In this paper, we propose a novel deep multimodal clustering framework called DMCR. Firstly, we control the scale of feature using VIB. Secondly, we reduce the distribution differences among multimodal features using cross reconstruction. Thirdly, we fuse the extracted features to common features. Finally, we cluster the common features using k-means. In addition, we prove that the proposed cross reconstruction method effectively reduce the distribution differences

of multimodal features. We compare our DMCR algorithm with the state-of-the-art multimodal methods on many multimodal datasets. Experimental results show that our algorithm achieves obviously improvement on multimodal clustering task.

References

1. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications. In: SIGMOD Conference (1998)
2. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. ArXiv abs/1612.00410 (2016)
3. Andrew, G., Arora, R., Bilmes, J.A., Livescu, K.: Deep canonical correlation analysis. In: ICML (2013)
4. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. ArXiv abs/1701.07875 (2017)
5. Cai, X., Nie, F., Huang, H., Kamangar, F.: Heterogeneous image feature integration via multi-modal spectral clustering. In: CVPR, pp. 1977–1984 (2011)
6. Cao, X., Zhang, C., Fu, H., Liu, S., Zhang, H.: Diversity-induced multi-view subspace clustering. In: CVPR, pp. 586–594 (2015)
7. Chen, G.: Deep learning with nonparametric clustering. ArXiv abs/1501.03084 (2015)
8. Coifman, R.R., Lafon, S.: Diffusion maps. Appl. Comput. Harmonic Anal. **21**, 5–30 (2006)
9. Fei-Fei, L., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: CVPR, vol. 2, pp. 524–531 (2005)
10. Gao, J., Han, J., Liu, J., Wang, C.: Multi-view clustering via joint nonnegative matrix factorization. In: SDM (2013)
11. Goodfellow, I.J.: NIPS 2016 tutorial: Generative adversarial networks. ArXiv abs/1701.00160 (2016)
12. Guo, X., Gao, L., Liu, X., Yin, J.: Improved deep embedded clustering with local structure preservation. In: IJCAI (2017)
13. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. CoRR abs/1312.6114 (2013)
14. von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**, 395–416 (2007)
15. Nadler, B., Lafon, S., Coifman, R.R., Kevrekidis, I.G.: Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. In: NIPS (2005)
16. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML (2011)
17. Peng, X., Xiao, S., Feng, J., Yau, W.Y., Yi, Z.: Deep subspace clustering with sparsity prior. In: IJCAI (2016)
18. Shaham, U., Stanton, K.P., Li, H., Nadler, B., Basri, R., Kluger, Y.: SpectralNet: spectral clustering using deep neural networks. ArXiv abs/1801.01587 (2018)
19. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep Boltzmann machines. J. Mach. Learn. Res. **15**, 2949–2980 (2012)
20. Tian, F., Gao, B., Cui, Q., Chen, E., Liu, T.Y.: Learning deep representations for graph clustering. In: AAAI (2014)

21. Wang, W., Arora, R., Livescu, K., Bilmes, J.A.: On deep multi-view representation learning. In: ICML (2015)
22. Wang, X., Guo, X., Lei, Z., Zhang, C., Li, S.Z.: Exclusivity-consistency regularized multi-view subspace clustering. In: CVPR, pp. 1–9 (2017)
23. Xie, J., Girshick, R.B., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: ICML (2015)
24. Yang, J., Parikh, D., Batra, D.: Joint unsupervised learning of deep representations and image clusters. In: CVPR, pp. 5147–5156 (2016)
25. Zhang, C., Fu, H., Liu, S., Liu, G., Cao, X.: Low-rank tensor constrained multiview subspace clustering. In: ICCV, pp. 1582–1590 (2015)