Deep Mutual Information Maximin for Cross-Modal Clustering

Yiqiao Mao,* Xiaoqiang Yan,* Qiang Guo, Yangdong Ye†
School of Information Engineering, Zhengzhou University, Zhengzhou, China
ieyqmao@gs.zzu.edu.cn, iexqyan@zzu.edu.cn, ieqguo@gs.zzu.edu.cn, ieydye@zzu.edu.cn

Abstract

Cross-modal clustering (CMC) aims to enhance the clustering performance by exploring complementary information from multiple modalities. However, the performances of existing CMC algorithms are still unsatisfactory due to the conflict of heterogeneous modalities and the high-dimensional non-linear property of individual modality. In this paper, a novel deep mutual information maximin (DMIM) method for cross-modal clustering is proposed to maximally preserve the shared information of multiple modalities while eliminating the superfluous information of individual modalities in an end-to-end manner. Specifically, a multi-modal shared encoder is firstly built to align the latent feature distributions by sharing parameters across modalities. Then, DMIM formulates the complementarity of multi-modalities representations as a mutual information maximin objective function, in which the shared information of multiple modalities and the superfluous information of individual modalities are identified by mutual information maximization and minimization respectively. To solve the DMIM objective function, we propose a variational optimization method to ensure it converge to a local optimal solution. Moreover, an auxiliary overclustering mechanism is employed to optimize the clustering structure by introducing more detailed clustering classes. Extensive experimental results demonstrate the superiority of DMIM method over the state-of-the-art cross-modal clustering methods on IAPR-TC12, ESP-Game, MIRFlickr and NUS-Wide datasets.

Introduction

With the rapid development of information technology, massive amounts of unlabeled data from multiple sources or views are generated in real-world applications every day. Usually, data from multiple sources is exhibited in a multi-modal form. For example, web news consists of pictures and corresponding texts. In live streaming, a video contains visual appearance and acoustic signal. Although multi-modal data appears in different modalities, sources and spaces, they often have similar high-level semantic information or cluster structures (Baltrusaitis, Ahuja, and Morency 2019). Thus, it is rational to learn cluster structures in cross-modal data with the aid of the complementary information from different modalities.

In recent years, cross-modal clustering (CMC) has made significant progress in machine learning and computer vision communities (Zhang et al. 2018, 2019; Guo and Ye 2019; Xing et al. 2019; Yan, Hu, and Ye 2017; Yan et al. 2020a). One of the most prevalent CMC approaches is to learn a shared subspace such that the mutual agreement between multiple modalities is maximized (Kim, Kittler, and Cipolla 2007; Akaho 2006; Chen et al. 2020; Zhang et al. 2019; Peng et al. 2019; Xing et al. 2019). In this research direction, the early and representative one is canonical correlation analysis (CCA) (Kim, Kittler, and Cipolla 2007), which seeks the shared subspace representations of two vectors by maximizing their correlations. After that, a variety of extensions of CCA have been proposed to learn a shared low-dimensional subspace of multiple modalities, such as kernel CCA (Akaho 2006), kernel information embedding (Wang et al. 2020) and generalized multi-view analysis (Sharma et al. 2012).

However, in these CMC methods, the original feature representations are destroyed and some necessary information is lost when the original features of different modalities are compressed into a shared low-dimensional subspace. Besides of learning a shared subspace, there are many other types of CMC approaches, such as matrix factorization based methods (Xing et al. 2019), graph model based methods (Gao et al. 2020b; Yan et al. 2020b). Although these aforementioned methods have made encouraging progress, they rely on hand-crafted features and linear embedding functions and cannot capture the non-linear structure of complex cross-modal data.

Recently, deep neural networks (DNN) have made vast inroads into unsupervised clustering due to their commendable performance. Motivated by this, DNN for cross-modal clustering has been increasingly exploited with the state-of-the-art results (Andrew et al. 2013; Abavisani and Patel 2018; Zhu et al. 2019; Federici et al. 2020; Li et al. 2019; Zhou and Shen 2020). Existing multi-modal clustering methods based on DNN are usually classified into the following two categories. The first category adopts a two-stage strategy (Wang et al. 2015; Abavisani and Patel 2018; Zhang, Liu, and Fu 2019), i.e., extracting features based on DNN and then learning the final clustering results by traditional

*Yiqiao Mao and Xiaoqiang Yan are joint first authors.
†Corresponding Author.

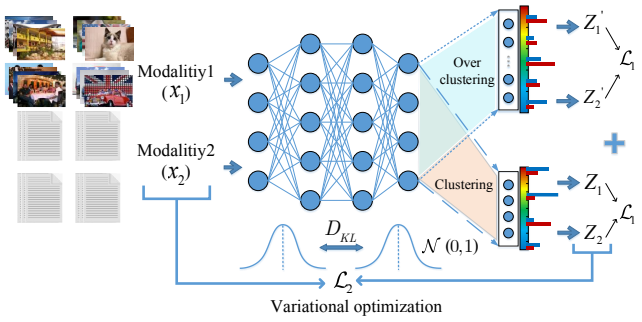


Figure 1: The pipeline of the proposed DMIM method. The $\mathcal{L}_1$ represents the preservation of shared information and $\mathcal{L}_2$ represents the elimination of superfluous information, which can be calculated by KL distance after variational optimization.

clustering method. For example, (Abavisani and Patel 2018) proposes a deep multi-modal subspace clustering network, which first utilizes multi-modal encoder to fuse modalities to a latent subspace representation and then applies spectral clustering to obtain the final results. (Zhang, Liu, and Fu 2019) presents an autoencoder in autoencoder network, which jointly performs view-specific encoding and multi-view encoding with a nested autoencoder, then K-means is applied to the multi-view representation. Obviously, this strategy may disconnect feature learning and cluster assignment, which results in the obtained feature representation is irrelevant to the later clustering task. The second category designs a clustering loss to guide the process of multi-modal feature learning (Li et al. 2019; Zhou and Shen 2020; Xu et al. 2019; Wang et al. 2018). For instance, (Zhou and Shen 2020) utilizes a adversarial network and clustering loss to optimize the pre-trained autoencoder and directly obtains the cluster structure. (Li et al. 2019) adopts an autoencoder to obtain a shared representation space between multiple views and develops a discriminator network to optimize data distribution. Although the second category has gained satisfactory results, it overly focuses on the extraction of multi-modal shared information, and does nothing to discard the irrelevant information. The resulting representations are not robust for given task as they have not eliminated modality specific nuisances.

Aiming at these challenging issues, we propose a novel deep mutual information maximin (DMIM) method for cross-modal clustering, which maximally preserves the shared information of multiple modalities while eliminating the superfluous information of individual modalities in an end-to-end manner. To this end, we first build a multi-modal shared encoder to align the latent feature distributions by sharing parameters across modalities. The multi-modal shared encoder can relieve the heterogeneous gap between different modalities since the data of different modalities obey the same coding rule by sharing parameter across modalities. Then, to maximize the shared information of multiple modalities and minimize the superfluous information of individual modalities, DMIM formulates the complementarity of multi-modalities representations as an mutual information maximin objective function. In this way, DMIM can obtain more reasonable clustering results as it has eliminated modality specific nuisances. And we also propose a variational optimization method to ensure the DMIM objective function converge to a local optimal solution. Furthermore, an auxiliary overclustering mechanism is employed to improve the clustering performance by introducing more detailed clustering classes. Figure 1 shows the pipeline of the proposed DMIM method. Extensive experimental results demonstrate the superiority of DMIM method over the state-of-the-art cross-modal clustering methods. The main contributions can be summarized as follows.

- We propose a novel deep mutual information maximin (DMIM) method for cross-modal clustering, which maximally preserves the shared information of multiple modalities while eliminating the superfluous information of individual modalities in an end-to-end manner.
- A multi-modal shared encoder is built to align the latent feature distributions by sharing parameters across modalities, which relieves the heterogeneous gap by letting the data of different modalities obey the same coding rule.
- A variational optimization method is proposed to ensure the DMIM objective function converge to a local optimal solution. Beside, an auxiliary overclustering mechanism is employed to improve the clustering performance by introducing more detailed clustering classes.

Related Work

CMC methods can effectively utilize the shared information between different modalities and have received extensive attention in recent years. We briefly introduce the relevant CMC methods in this section, which can be roughly divided into two categories: traditional CMC methods and deep learning based methods.

Cross-modal Clustering

The most representative method of traditional CMC is latent subspace learning, which maps cross-modal data into a shared latent subspace. Latent subspace learning can be roughly divided into three subcategories. The first is canonical correlation analysis (CCA) (Kim, Kittler, and Cipolla 2007), which seeks a low-dimensional latent subspace representation by maximizing the correlation between modalities. In order to improve the non-linear learning ability of CCA, (Akaho 2006) combines multiple non-linear kernel functions with CCA, and proposes the KCCA method. The second one uses matrix factorization technology to find latent factors for cross-modal data (Xing et al. 2019; Tsai et al. 2019). For instance, (Xing et al. 2019) proposes a collaborative matrix factorization method, which collaboratively factorizes relational data matrices to capture the intrinsic relations of multi-view data. The third is binary code learning (Zhang et al. 2018, 2019). For example, (Zhang et al. 2019) en codes cross-modal data into a shared binary code subspace through complementary information between modalities. In addition to latent subspace learning, anchor-based CMC have also attracted researchers attention. (Guo

and Ye 2019) employs anchors to reconstruct the correlation between instances and builds a similarity matrix between modalities. However, the aforementioned methods rely on hand-crafted features, and the representation ability of original features limits the performance of the methods.

Deep Clustering

The combination of DNN and CMC can obtain better feature representations than traditional methods, and has become a research hotspot in recent years. The CMC methods based on DNN can be roughly classified into two subcategories. The first one learns non-linear feature representations through DNN, and then uses traditional clustering methods to obtain cluster structures (Andrew et al. 2013; Abavisani and Patel 2018; Federici et al. 2020; Gao et al. 2020a). For example, (Wang et al. 2015) proposes a deep canonically correlated autoencoders (DCCAE) to learn a common representation of multiple modalities, then K-means is performed to obtain the cluster structures. (Federici et al. 2020) utilizes mutual information maximization to train a shared deep autoencoder (DAE) for cross-modal data, which can obtain cross-modal shared features. However, these methods only focus on feature learning and lack of improvement on clustering process. The second one combines feature representation with clustering process, and uses clustering loss to guide model learning (Li et al. 2019; Zhou and Shen 2020; Wang et al. 2018; Xu et al. 2019). For instance, (Wang et al. 2018) complements the common latent network of DAE by generative adversarial networks (GAN), and uses adversarial mechanisms to explore the supplementary information shared by each modality. However, the above methods only focus on finding the shared information among cross-modal data, and do not remove superfluous information in each modality, which reduces the performance of these methods.

Deep Mutual Information Maximin

In this section, the network architecture of the DMIM method is introduced and the formulation of its objective function is given, then we provide the optimization of DMIM approach.

Network Architecture

We are given two heterogeneous modalities $X_1 = \{x_1^i\} \in \mathbb{R}^{d_1 \times N}, X_2 = \{x_2^i\} \in \mathbb{R}^{d_2 \times N}$, where N is the size of cross-modal data instances, d_1 and d_2 are the feature dimension of X_1 and X_2. We adopt random variable Y to denote the class label that DMIM intends to predict about X_1 and X_2, $|Y|$ indicates the number of class label. DMIM involves predicting class label Y while maximizing the shared information of multiple modalities and minimizing the superfluous information of individual modalities simultaneously.

As shown in Figure 1, the network architecture of DMIM consists of the following components: multi-modal shared encoder Φ, clustering layer C and overclustering layer C'. Specifically, the multi-modal shared encoder Φ is designed to align the latent feature distributions by sharing parameters across modalities, while the clustering layer C and overclustering layer C' are utilized to transform the pattern structures of X_1 and X_2 into probability distribution of class label Y. Let Z_1 and Z_2 be probability distribution of class label Y of modalities X_1 and X_2. In order to share the parameters across modalities, we generate Z_1 and Z_2 through the same autoencoder Φ, which make Z_1 and Z_2 obey the same coding rule (Federici et al. 2020). Since the multiple related modalities X_1 and X_2 are characterized by Z_1 and Z_2, maximising the predictability from Z_1 to Z_2 will preserve the shared information between modalities X_1 and X_2. At the same time, the superfluous information of individual modalities is identified so as to discard modality-specific details.

When the number of class label $|Y|$ is small, the supervision effect of probability distribution of class label Y will be weaken in the process of back propagation. Thus, we also add a overclustering layer C' to the multi-modal shared encoder Φ. Usually, the number of predicted label of overclustering layer is set as much larger than true label, which can improve the clustering performance by introducing more detailed classes.

Objective Function

DMIM aims to maximally preserve the shared information of multiple modalities while eliminating the superfluous information of individual modalities. With this end in view, we devise a two-fold objective function. In the first part, we preserve the shared information of multiple modalities by maximizing the mutual information between Z_1 to Z_2 as follows

$$\mathcal{L}_1 = \max I(Z_1; Z_2) \tag{1}$$

where Z_1 and Z_2 are probability distributions of class label Y of modalities X_1 and X_2.

In the traditional entropy based loss function, maximizing entropy alone can result in degenerate solutions since entropy could be maximized trivially by setting all prediction vectors to same cluster (Caron et al. 2018). In contrast, mutual information measurement can avoid this problem. As we know, $I(Z_1; Z_2)$ can be rewritten as

$$I(Z_1; Z_2) = H(Z_1) - H(Z_1|Z_2) \tag{2}$$

where the maximum value of $H(Z_1)$ is $\log |Y|$ when each instance is allocated into every cluster with equally probability. The minimum value of the conditional entropy $H(Z_1|Z_2)$ is 0 when two clustering assignments are exactly predictable from each other. Thus, allocating all data instances to one cluster cannot maximize the mutual information $I(Z_1; Z_2)$, which naturally avoids degenerate solutions.

To eliminate modality specific nuisances, we devise the second part of the objective function as follows

$$\mathcal{L}_2 = \min[I(Z_1; X_1) + I(Z_2; X_2)] \tag{3}$$

where Z_1 can be seen as a compressed representation of original data X_1, so minimizing $I(Z_1; X_1)$ can eliminate the information in X_1. The minimal value of $I(Z_1; X_1)$ is 0, obviously, this is not our goal since the information in X_1 is totally discarded. Thus, we introduce a trade-off between shared information preservation and superfluous information elimination, and obtain the objective function of DMIM as follows

$$\begin{aligned} \mathcal{L} &= \max (\mathcal{L}_1 + \mathcal{L}_2) \\ &= I(Z_1; Z_2) - [I(Z_1; X_1) + I(Z_2; X_2)] \end{aligned} \tag{4}$$

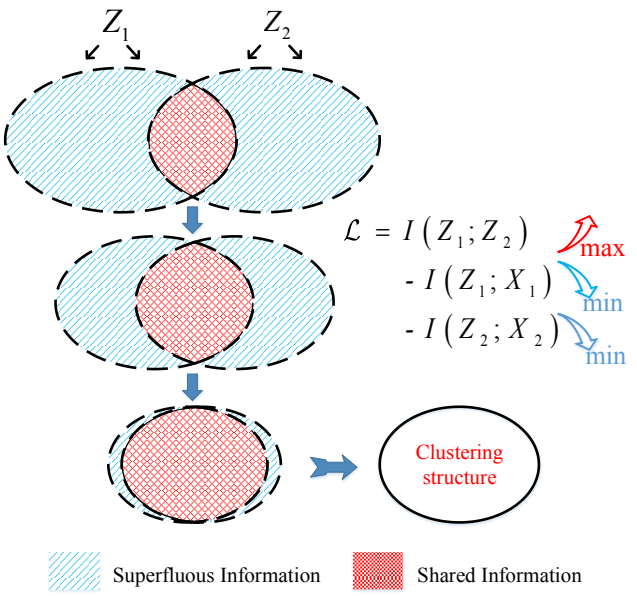


Figure 2: The illustration of shared information preservation and superfluous information elimination.

The Figure 2 shows the process of shared information preservation and superfluous information elimination.

Optimization

In this part, a variational optimization method is proposed to optimize the objective function of DMIM method. For the calculation of mutual information $I(Z_1; Z_2)$, we should first obtain the joint probability distribution $p(Z_1, Z_2)$. When the random variables Z_1 and Z_2 are independent to each other, $p(Z_1, Z_2)$ can be calculated by $p(Z_1, Z_2) = p(Z_1) \times p(Z_2)$. In this study, Z_1 and Z_2 are generated by related cross-modal data X_1 and X_2, so they are not independent. According to (Ji, Vedaldi, and Henriques 2019), we can calculate $p(Z_1, Z_2) = p(Z_1) \times p(Z_2)^T$ after marginalization over the cross-modal dataset (or batch, in practice).

Now, we present the optimization of the second part $\mathcal{L}_2$ of the DMIM objective function. In this study, we propose a variational method to fit $\mathcal{L}_2$ by its variational lower bound (Barber and Agakov 2003), in which maximizing this variational lower bound gives an unbiased estimation of the objective function. According to the definition of mutual information, we obtain

$$\begin{aligned} I(Z_1; X_1) &= \int_{z_1, x_1} p(z_1, x_1) \log \frac{p(z_1, x_1)}{p(z_1)p(x_1)} \\ &= \int_{z_1, x_1} p(z_1, x_1) \log \frac{p(z_1 \mid x_1)}{p(z_1)} \end{aligned} \tag{5}$$

We try to use the variational estimation $q(z_1)$ of $p(z_1)$ to calculate the posterior probability distribution $p(z_1|x_1)$. From the non-negativeness of $Kullback\text{-}Leibler$ (KL) divergence, we can get

$$\begin{aligned} &KL[p(z_1), q(z_1)] = \int_{z_1} p(z_1) \log \frac{p(z_1)}{q(z_1)} > 0 \\ &\Rightarrow \int_{z_1} p(z_1) \log p(z_1) > \int_{z_1} p(z_1) \log q(z_1) \end{aligned} \tag{6}$$

Now, the mutual information $I(Z_1; X_1)$ can be rewritten as follows

$$\begin{aligned} I(Z_1; X_1) &= \int_{z_1, x_1} p(z_1, x_1) \log \frac{p(z_1 \mid x_1)}{p(z_1)} \\ &< \int_{z_1, x_1} p(z_1, x_1) \log \frac{p(z_1 \mid x_1)}{q(z_1)} \\ &< \int_{z_1, x_1} p(x_1) p(z_1 \mid x_1) \log \frac{p(z_1 \mid x_1)}{q(z_1)} \end{aligned} \tag{7}$$

Similarly,

$$I(Z_2; X_2) < \int_{z_2, x_2} p(x_2) p(z_2 \mid x_2) \log \frac{p(z_2 \mid x_2)}{q(z_2)} \tag{8}$$

Now, the second part of DMIM objective function $\mathcal{L}_2$ is

$$\begin{aligned} \mathcal{L}_2 &= I(Z_1; X_1) + I(Z_2; X_2) \\ &< \int_{z_1, x_1} p(x_1) p(z_1|x_1) \log \frac{p(z_1|x_1)}{q(z_1)} \\ &+ \int_{z_2, x_2} p(x_2) p(z_2|x_2) \log \frac{p(z_2|x_2)}{q(z_2)} \end{aligned} \tag{9}$$

Next, in order to calculate the variational approximation of $\mathcal{L}_2$, unnecessary items need to be removed. Therefore we adopt Monte Carlo sampling (Shapiro 2003) to replace $p(x_1), p(x_2)$ and get

$$\begin{aligned} \mathcal{L}_2 < \frac{1}{N} \sum_i^N \{ &\int_{z_1} p(z_1 \mid x_1) \log \frac{p(z_1 \mid x_1)}{q(z_1)} + \\ &\int_{z_2} p(z_2 \mid x_2) \log \frac{p(z_2 \mid x_2)}{q(z_2)} \} \end{aligned} \tag{10}$$

Suppose $p(z_1|x_1)$ and $p(z_2|x_2)$ obey Gaussian distribution, where the mean and variances can be learned from the multi-modal shared encoder Φ, i.e., $p(z|x) = \mathcal{N}(\mu_\phi(x), \sigma(x))$, and ϕ is the parameters in the encoder. For simplicity, we re-parameterize (Kingma and Welling 2014) z_1 and z_2 as

$$\begin{aligned} z_1 &= \mu(x_1) + \sigma(x_1) * \epsilon \\ z_2 &= \mu(x_2) + \sigma(x_2) * \epsilon \end{aligned} \tag{11}$$

where ϵ is the standard normal distribution. Now, the second part of DMIM objective function $\mathcal{L}_2$ can be rewritten as

$$\begin{aligned} \mathcal{L}_2 &< \frac{1}{N} \sum_i^N \{ \mathbb{E}_{\epsilon_1} \log \frac{p(z_1 \mid x_1)}{q(z_1)} + \mathbb{E}_{\epsilon_2} \log \frac{p(z_2 \mid x_2)}{q(z_2)} \} \\ &< \frac{1}{N} \sum_i^N \mathbb{E}_\epsilon \{ KL[p(z_1 \mid x_1), q(x_1)] \\ &+ KL[p(z_2 \mid x_2), q(x_2)] \} \end{aligned} \tag{12}$$

Algorithm 1 The DMIM Algorithm

Input:
Cross-modal data X_1 and X_2, the number of clusters $|Y|$
Random initialization: Initialize the encoder Φ, layer of clustering C and over-clusteringC'
repeat
  $\mathcal{L}_c \leftarrow Z_1 = \Phi_C(X_1), Z_2 = \Phi_C(X_2)$ according to Eq. 14
  $\mathcal{L}_{c'} \leftarrow Z_1 = \Phi_{C'}(X_1), Z_2 = \Phi_{C'}(X_2)$ according to Eq. 14
  $\mathcal{L} = \mathcal{L}_c + \mathcal{L}_{c'}$
  Optimize network parameters of encoder Φ, C and C'
until converge
Output: The encoder Φ and layer of clustering C

where $p(z_1 \mid x_1)$ and $p(z_2 \mid x_2)$ are all learned from encoder Φ. Besides, to make sure that the instances are evenly allocated into all clusters, we set a uniform distribution constraint (Asano, Rupprecht, and Vedaldi 2020) to $q(x_1)$ and $q(x_2)$, i.e.,

$$\sum_{i}^{N} q(x_{1,2}) = \frac{N}{|Y|} \tag{13}$$

Finally, the overall objective function of the proposed DMIM method can be formulated as

$$\mathcal{L} \approx I(Z_1; Z_2) - \frac{1}{N}\sum_{i}^{N}\mathbb{E}_{\epsilon}\{KL[p(z_1 \mid x_1), q(x_1)] + KL[p(z_2 \mid x_2), q(x_2)]\}, with \sum_{i}^{n} q(x_{1,2}) = \frac{N}{|Y|} \tag{14}$$

In every loop of training, the DMIM method gets two probability distributions from clustering layer C and over-clustering layer C'. The overclustering layer is only utilized for training. The overall clustering process is presented in Algorithm 1, in which the $\Phi_C(X_1)$ and $\Phi_{C'}(X_1)$ denote the clustering results of X_1 from clustering layer and overclustering layer.

Experiments

In this section, extensive experiments on four real-world cross-modal datasets are conducted to verify the effectiveness of the proposed DMIM method.

Datasets and Description

The datasets used in our experiments include: 1) **IAPR-TC12** (Michael Grubinger 2006): It is a publicly image dataset with annotated tags, which contains 20,000 images with 6 classes. Each image is accompanied by a short text description. We randomly select 7855 images with its texts after removing the images which the number of tags is less than 4. 2) **ESP-Game** (von Ahn and Dabbish 2004): It is a social image collection searched from a image annotation website, which consists of 20,770 images and corresponding tags. We select 11,032 images with approximately 5 tags per image. 3) **MIRFlickr** (Huiskes and Lew 2008): It contains 25,000 images with 1386 tags. After de-noising and de-duplication, we construct a dataset with 12,154 images in total of 6 classes. 4) **NUS-Wide** (Chua et al. 2009): It consists of 269,648 social images, i.e., image with tags. By removing blank images and non-English words, we build a dataset with 20,000 images, which contains 8 classes and each image is along with 7 tags on average. We call these four datasets **IAPR, ESP, Flickr** and **NUS** for short, respectively. The details of the four datasets are shown in Table **??**.

For the image representation, we adopt the penultimate layer of VGG-16 (Simonyan and Zisserman 2015) as the 4096 dimensional features of images. To improve the computational efficiency, we use Principal Component Analysis (PCA) (Vidal, Ma, and Sastry 2005) to reduce the dimensions to 100. For the text representation, we employ BERT (Turc et al. 2019) to extract 768 dimensional features. Similarly, we also reduce the dimensions to 100 by PCA.

Dataset	Clusters	Size
IAPR	6	7855
ESP	7	11032
Flickr	6	12154
NUS	8	20000

Table 1: Statistics of the four datasets.

Experimental Setup

In the proposed DMIM method, the multi-modal shared encoder is composed of two fully connected layers. Each fully connected layer is followed by a BatchNorm layer and a ReLU layer, which are utilized to perform data normalization and non-linear activation, respectively. The clustering layer and overclustering layer also adopt fully connected layer, in which the number of hidden nodes are set as $|Y|$ and 10 $\times$ $|Y|$, respectively. At the beginning of training process, all parameters in the model are initialized randomly. After training, the clustering result is the output of clustering layer C with a softmax layer. We use Adam (Kingma and Ba 2015) optimizer with the learning rate 5×10^{-5}.

We implement our proposed DMIM method and deep clustering baselines with the public toolbox of PyTorch. Other traditional comparison baselines are conducted on Matlab 2016a. We conduct all the experiments on the platform of Windows 10 with NVIDIA 1060 Graphics Processing Units (GPUs) and 32G memory size.

Baselines and Evaluation Metrics

The performance of DMIM is compared with the state-of-the-art cross-modal clustering methods. There are six shallow and four deep cross-modal clustering algorithms.

1) K-means: It is a traditional single-modal clustering methods. We report its best clustering results on all modalities.

Methods	ESP		Flickr		IAPR		NUS	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Kmeans	0.421	0.272	0.407	0.223	0.383	0.166	0.308	0.183
CCA	0.465	0.295	0.492	0.237	0.461	0.229	0.340	0.203
MCLES	0.528	0.367	0.432	0.274	0.442	0.231	0.398	0.231
BMVC	0.523	0.351	0.531	0.321	0.421	0.227	0.467	0.293
HSIC	0.537	0.339	0.533	0.302	0.413	0.236	0.406	0.275
APMC	0.523	0.347	0.532	0.328	0.442	0.227	0.375	0.224
DCCA	0.457	0.252	0.495	0.313	0.436	0.234	0.381	0.265
DCCAE	0.461	0.311	0.429	0.307	0.413	0.227	0.357	0.273
AE2Net	0.437	0.279	0.544	0.321	0.444	0.213	0.445	0.288
DMIB	0.509	0.232	0.561	0.287	0.494	0.238	0.449	0.198
Our	**0.561**	**0.371**	**0.574**	**0.347**	**0.527**	**0.274**	**0.485**	**0.311**

Table 2: Clustering performance of different methods on four challenging datasets. The best results are highlighted in bold.

2) Canonical correlation analysis (**CCA**) (Kim, Kittler, and Cipolla 2007): it maps the original data into a shared subspace to maximize the correlations between modalities.

3) Multi-view clustering in latent embedding space (**MCLES**) (Chen et al. 2020): It learns the latent embedding representations of multi-modal data, and then obtains the clustering index matrix of global structure.

4) Binary multi-view clustering (**BMVC**) (Zhang et al. 2019): It encodes multi-modal data into a shared binary code subspace through complementary information between modalities.

5) Highly-economized scalable image clustering (**HSIC**) (Zhang et al. 2018): It jointly learns multi-modal shared binary code representations and a discrete clustering architecture.

6) Anchor-based partial multi-view clustering (**APMC**) (Guo and Ye 2019): It utilizes anchors to reconstruct the correlations between instances and builds a similarity matrix between modalities.

7) Deep canonical correlation analysis (**DCCA**) (Andrew et al. 2013): It is a non-linear extension of traditional CCA, which learns a non-linear mapping of multi-modal data to maximize the correlations of the result representations.

8) Deep canonically correlated autoencoders (**DCCAE**) (Wang et al. 2015): It employs two deep autoencoders to learn the cross-modal features, and the correlations between these features are calculated to reconstruct the data and optimize the network.

9) Autoencoder in autoencoder networks (**AE^{2}Net**) (Zhang, Liu, and Fu 2019): It jointly performs view-specific encoding and multi-view encoding with a nested autoencoder.

10) Deep multi-view information bottleneck (**DMIB**) (Federici et al. 2020): It adopts mutual information to measure the correlation of two autoencoder, and obtains fused feature representations of two views.

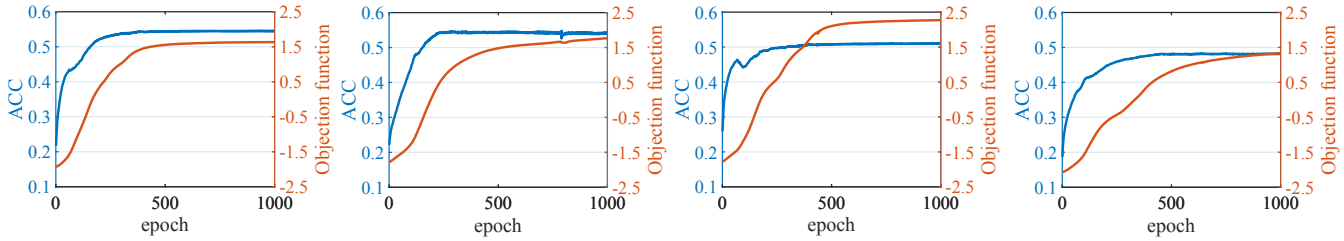
We report the average evaluation with the metrics of Clustering Accuracy (ACC) and Normalized Mutual Information (NMI). Higher values indicate better clustering performance for the metrics. For the multi-view methods, we consider different modalities as the different views. In the experiments, we run 20 times for each experiment and report the average performance.

Comparison Results

The clustering performances of DMIM and comparison algorithms on four datasets are reported in Table **??**. From the presented results, we have the following observations and discussions: 1) Our method achieves the best results on each dataset compared with other algorithms, which verifies the DMIM method can effectively explore the shared information among cross-modal data and obtain better clustering performance. 2) The DMIM method shows superior performance than traditional methods on four datasets, which demonstrates that the DMIM method based on deep neural networks can more effectively learn the non-linear relationships contained in the features. 3) At the same time, DMIM is also superior to the representative deep cross-modal clustering methods. It shows that our DMIM maximally preserves the shared information of multiple modalities while eliminating the superfluous information of individual modalities, which allows it to obtain more robust clustering structure. 4) Similar to our method, the DMIB method is also based on mutual information. Compared with the DMIB, DMIM method achieves better clustering structure. The reason is that the DMIB method separates feature learning and clustering tasks, which makes it only focus on feature learning and ignore the clustering task. In addition, our DMIM method employs an auxiliary overclustering mechanism to improve clustering performance.

Ablation Study

We perform ablation study to analyze the role of each component in DMIM method in this part. Specifically, we conduct experiments with different components ablation as follows: 1) Without overclustering and superfluous information elimination (Without Over and SIE). In this scenario, we only remain the first part of objective function 4, i.e., $I(Z_1; Z_2)$. 2) Without overclustering (Without Over). In this setting, we remove the overclustering layer compared with DMIM method. 3) Without superfluous information elimination (Without SIE). We only remove the part of superflu-



(a) ESP (b) Flickr (c) IAPR (d) NUS

Figure 3: The convergence of DMIM. The red line indicates the values of objective function, while the blue line denotes the ACC values.

Methods	ESP	Flickr	IAPR	NUS
Without Over&SIE	0.4864	0.5051	0.4606	0.4407
Without SIE	0.5162	0.5357	0.4701	0.4595
Without Over	0.5305	0.5517	0.5148	0.4678
DMIM	**0.5613**	**0.5739**	**0.5274**	**0.4855**

Table 3: Ablation study in terms of ACC.

ous information elimination from DMIM method. As shown in Table 3, we can observe: firstly, the overclustering layer has a certain impact on the clustering performance, which provides more detailed clustering classes for the target data. Secondly, the part of superfluous information elimination has a significant impact on the final clustering results, which can eliminate modality specific nuisances. These above observations indicate that all components in DMIM method are designed reasonably.

Convergence Analysis

In order to investigate the convergence of the proposed DMIM method, we report the values of objective function and the ACC values with the increasing of iteration number in Figure 3. In this figure, the red line indicates the value of objective function, while the blue line denotes the ACC values. As shown in Figure 3, we can observe that these two values increase quickly in the first approximately 300 iterations, then the values approach to be a fixed value after approximately 500 iterations. Therefore, our proposed optimization algorithm is reliable and converges quickly.

Impact of the Cluster Number of Overclustering

The proposed DMIM incorporates an auxiliary overclustering mechanism, which improves the clustering performance by introducing more detailed clustering classes. In this section, we perform experiments to explore the impact of the cluster number of overclustering on the final clustering results. Suppose the number of ground-truth label is $|Y|$, the cluster number of overclustering varies in $\{1 \times |Y|, 2 \times |Y|, \cdots, 20 \times |Y|\}$. As shown in Figure 4, we can observe that with the gradual increasing of the cluster number from $1 \times K$, the ACC values of DMIM have a raise on the four used datasets. The reason of this phenomenon is

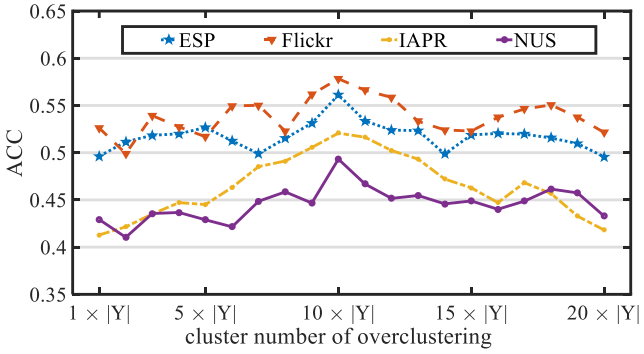


Figure 4: The impact of cluster number of the overclustering mechanism.

that small cluster number may lack the discriminative power, since it cannot provide detailed class information. As the number of clusters increases, overclustering layer can provide more and more detailed cluster structures. Then, with further increasing of the cluster number, the accuracy decreases to some extent. Note that, the DMIM method obtains best ACC when cluster number is $10 \times |Y|$ on the four used datasets. This is because too large cluster number is not only sensitive to errors, but also makes the model more inclined to optimize the overclustering layer C', while ignoring the optimization of the clustering layer C.

Conclusion

In this paper, a novel deep mutual information maximin (DMIM) method for cross-modal clustering is proposed to maximally preserve the shared information of multiple modalities while eliminating the superfluous information of individual modalities. Specifically, a multi-modal shared encoder is built to align the latent feature distributions. Then, the shared information of multiple modalities and the superfluous information of individual modalities are identified by mutual information maximization and minimization respectively. Finally, a variational optimization method is proposed to ensure it converge to a local optimal solution. Experimental results on several real-world datasets show the superiority of the proposed DMIM method over state-of-the-art cross-modal clustering methods.

Acknowledgements

This work was supported by National Natural Science Foundation of China under grant 61772475 and 61906172, the National Key Research and Development Program of China under grant 2018YFB1201403.

References

Abavisani, M.; and Patel, V. M. 2018. Deep Multimodal Subspace Clustering Networks. *JSTSP* 12(6): 1601–1614.

Akaho, S. 2006. A kernel method for canonical correlation analysis. *CoRR* abs/cs/0609071.

Andrew, G.; Arora, R.; Bilmes, J. A.; and Livescu, K. 2013. Deep Canonical Correlation Analysis. In *ICML*, 1247–1255.

Asano, Y. M.; Rupprecht, C.; and Vedaldi, A. 2020. Self-labelling via simultaneous clustering and representation learning. In *ICLR*.

Baltrusaitis, T.; Ahuja, C.; and Morency, L. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *TPAMI* 41(2): 423–443.

Barber, D.; and Agakov, F. V. 2003. The IM Algorithm: A Variational Approach to Information Maximization. In *NIPS*, 201–208.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep Clustering for Unsupervised Learning of Visual Features. In *ECCV*, 139–156.

Chen, M.; Huang, L.; Wang, C.; and Huang, D. 2020. Multi-View Clustering in Latent Embedding Space. In *AAAI*, 3513–3520.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. NUS-WIDE: A Real-world Web Image Database from National University of Singapore. In *CIVR*, 48:1–48:9.

Federici, M.; Dutta, A.; Forré, P.; Kushman, N.; and Akata, Z. 2020. Learning Robust Representations via Multi-View Information Bottleneck. In *ICLR*.

Gao, Q.; Lian, H.; Wang, Q.; and Sun, G. 2020a. Cross-Modal Subspace Clustering via Deep Canonical Correlation Analysis. In *AAAI*, 3938–3945.

Gao, Q.; Xia, W.; Wan, Z.; Xie, D.; and Zhang, P. 2020b. Tensor-SVD Based Graph Learning for Multi-View Subspace Clustering. In *AAAI*, 3930–3937.

Guo, J.; and Ye, J. 2019. Anchors Bring Ease: An Embarrassingly Simple Approach to Partial Multi-View Clustering. In *AAAI*, 118–125.

Huiskes, M. J.; and Lew, M. S. 2008. The MIR Flickr Retrieval Evaluation. In *MIR*, 39–43.

Ji, X.; Vedaldi, A.; and Henriques, J. F. 2019. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *ICCV*, 9864–9873.

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. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Li, Z.; Wang, Q.; Tao, Z.; Gao, Q.; and Yang, Z. 2019. Deep Adversarial Multi-view Clustering Network. In *IJCAI*, 2952–2958.

Michael Grubinger, Paul Clough, H. M. T. D. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *LREC*, 13–23.

Peng, X.; Huang, Z.; Lv, J.; Zhu, H.; and Zhou, J. T. 2019. COMIC: Multi-view Clustering Without Parameter Selection. In *ICML*, 5092–5101.

Shapiro, A. 2003. Monte Carlo Sampling Methods. *Handbooks in Operations Research and Management Science* 10: 353–425.

Sharma, A.; Kumar, A.; Daume, H.; and Jacobs, D. W. 2012. Generalized Multiview Analysis: A discriminative latent space. In *CVPR*, 2160–2167.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Tsai, Y. H.; Liang, P. P.; Zadeh, A.; Morency, L.; and Salakhutdinov, R. 2019. Learning Factorized Multimodal Representations. In *ICLR*.

Turc, I.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. *arXiv:1908.08962* .

Vidal, R.; Ma, Y.; and Sastry, S. 2005. Generalized Principal Component Analysis (GPCA). *TPAMI* 27(12): 1945–1959.

von Ahn, L.; and Dabbish, L. 2004. Labeling Images with a Computer Game. In *HCI*, 319–326.

Wang, Q.; Ding, Z.; Tao, Z.; Gao, Q.; and Fu, Y. 2018. Partial Multi-view Clustering via Consistent GAN. In *ICDM*, 1290–1295.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. A. 2015. On Deep Multi-View Representation Learning: Objectives and Optimization. In *ICML*, 1083–1092.

Wang, Z.; Wang, L.; Wan, J.; and Huang, H. 2020. Shared low-rank correlation embedding for multiple feature fusion. *IEEE TMM* 1–1.

Xing, Y.; Yu, G.; Domeniconi, C.; Wang, J.; Zhang, Z.; and Guo, M. 2019. Multi-View Multi-Instance Multi-Label Learning Based on Collaborative Matrix Factorization. In *AAAI*, 5508–5515.

Xu, C.; Guan, Z.; Zhao, W.; Wu, H.; Niu, Y.; and Ling, B. 2019. Adversarial Incomplete Multi-view Clustering. In *IJCAI*, 3933–3939.

Yan, X.; Hu, S.; and Ye, Y. 2017. Multi-task Clustering of Human Actions by Sharing Information. In *CVPR*, 4049–4057.

Yan, X.; Ye, Y.; Qiu, X.; Manic, M.; and Yu, H. 2020a. CMIB: Unsupervised Image Object Categorization in Multiple Visual Contexts. *IEEE TII* 16(6): 3974–3986.

Yan, X.; Ye, Y.; Qiu, X.; and Yu, H. 2020b. Synergetic information bottleneck for joint multi-view and ensemble clustering. *Information Fusion* 56: 15–27.

Zhang, C.; Liu, Y.; and Fu, H. 2019. AE2-Nets: Autoencoder in Autoencoder Networks. In *CVPR*, 2577–2585.

Zhang, Z.; Liu, L.; Qin, J.; Zhu, F.; Shen, F.; Xu, Y.; Shao, L.; and Shen, H. T. 2018. Highly-Economized Multi-view Binary Compression for Scalable Image Clustering. In *ECCV*, 731–748.

Zhang, Z.; Liu, L.; Shen, F.; Shen, H. T.; and Shao, L. 2019. Binary Multi-View Clustering. *TPAMI* 41(7): 1774–1782.

Zhou, R.; and Shen, Y. 2020. End-to-End Adversarial-Attention Network for Multi-Modal Clustering. In *CVPR*, 14607–14616.

Zhu, P.; Hui, B.; Zhang, C.; Du, D.; Wen, L.; and Hu, Q. 2019. Multi-view Deep Subspace Clustering Networks. *CoRR* abs/1908.01978.