Multimodal Clustering via Deep Commonness and Uniqueness Mining

Linlin Zong[1],Faqiang Miao[1],Xianchao Zhang[1],Bo Xu[2]
[1]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province,Dalian School of Software, Dalian University of Technology
[2]School of Computer Science and Technology, Dalian University of Technology
Dalian, China
llzong@dlut.edu.cn,fqmiao@yeah.net,{xczhang,xubo}@dlut.edu.cn

ABSTRACT

Deep multimodal clustering have shown their competitiveness among different multimodal clustering algorithms. Existing algorithms usually boost the multimodal clustering by exploring the common knowledge among multiple modalities, which underutilizes the uniqueness of multiple modalities. In this paper, we enhance the mining of modality-common knowledge by extracting the modality-unique knowledge of each modality simultaneously. Specifically, we first utilize autoencoders to extract the modality-common and modality-unique features of each modality respectively. Meanwhile, the cross reconstruction is used to build latent connections among different modalities, i.e., maintain the consistency of modality-common features of each modality as well as heightening the diversity of modality-unique features of each modality. After that, modality-common features are fused to cluster the multimodal data. Experimental results on several benchmark datasets demonstrate that the proposed method outperforms state-of-art works obviously.

CCS CONCEPTS

• **Computing methodologies** → Information extraction,Cluster analysis,Neural networks.

KEYWORDS

Multimodal Clustering, Deep Clustering, Consensus, Diversity

ACM Reference Format:
Linlin Zong[1],Faqiang Miao[1],Xianchao Zhang[1],Bo Xu[2]. 2020. Multimodal Clustering via Deep Commonness and Uniqueness Mining. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3412103

1 INTRODUCTION

Multimodal clustering integrates multiple representations together to identify clusters. The key problem is to obtain the confluent features with multiple information. Existing works usually achieve the confluent features based on traditional clustering and deep clustering [4][8][12]. On account of the high effectiveness of deep clustering for feature extraction, we focus on extracting the confluent features by using multiple deep neural networks (DNN).

Recently, various DNN based multimodal clustering methods have been proposed, including methods based on Deep Bolzmann Machine (DBM) and deep autoencoder. The DBM based methods [9] learn a joint representation of different modalities by DBM. But due to the high computational costs in high-dimensional data space, the DBM based methods have not been widely studied in recent years. The autoencoder based methods use autoencoders to extract low dimensional features of each modality and fusion the multimodal features to best reconstruct the input data [6, 10].

Although autoencoder based methods have been shown effective in integrating multiple feature knowledge, there is only a little study of how multimodal knowledge should be effectively exploited to maximize clustering performance. Most existing methods simply extract knowledge that exist in each modality (modality-common knowledge). In fact, the modalities are complementary, and each modality may contain some knowledge that do not exist in other modalities (modality-unique knowledge). It is not adequate to merely explore the commonness of multiple modalities, which is difficult to make full use of the knowledge of each modality comprehensively.

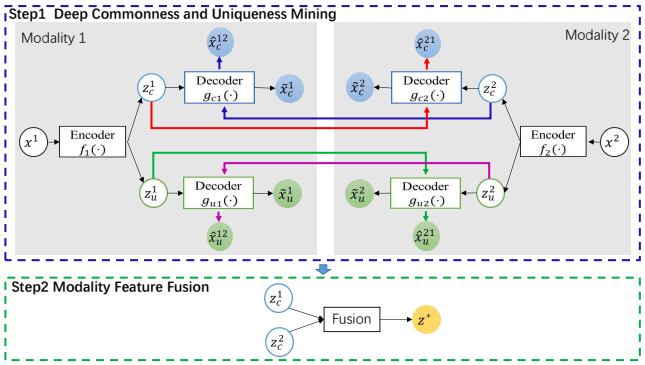

Figure 1: Flow of the proposed method.

Since the multiple modalities share the same clustering results, in this paper, we enhance the mining of modality-common knowledge by extracting the modality-unique knowledge of each modality simultaneously. Taking two modalities for example, the process is

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412103

shown in Figure 1. The proposed method consists of two steps. (1) Deep commonness and uniqueness mining. Given data x with two modalities x^1 and x^2, we extract the modality-common features z_c^j and modality-unique features z_u^j of each modality through one encoder $f_j(\cdot)$, $j \in \{1,2\}$, and then use two decoders $g_{cj}(\cdot)$ and $g_{uj}(\cdot)$ to reconstruct the modality-common data $\tilde{x}_c^j$ and modality-unique data $\tilde{x}_u^j$ of each modality. At the same time, the modality features $\hat{x}_c^{hj}$ and $\hat{x}_u^{hj}$, $h,j \in \{1,2\}, h \neq j$ are cross-reconstructed. By heightening the diversity of modality-unique features of each modality, we refine the consistency of modality-common features of each modality. (2) Modality features fusion. We obtain a final shared feature by fusing the modality-common features through the fusion layer. Experimental results show that our method outperforms state-of-the-art methods.

2 THE PROPOSED METHOD

Given multimodal data $X = \{x_1, x_2, \ldots, x_n\}$, where n is the number of instances. Each instance contains m modalities, i.e., $x_i = \{x_i^1, x_i^2, \ldots, x_i^m\}$ and $x_i^m \in R^{d_m}$ is the i-th instance in the m-th modality. We aim to cluster the multimodal data X into k clusters $C = \{c_1, c_2, \ldots, c_k\}$.

2.1 Deep Commonness and Uniqueness Mining

2.1.1 Feature Extraction. For multimodal data, each modality of the data may contain some knowledge shared by all the modalities (modality-common knowledge) and some knowledge that other modalities do not have (modality-unique knowledge). The modality-common knowledge are usually used to obtain the consensus result of the multimodal data, and would be refined by excluding the modality-unique knowledge explicitly. Then, we formulate the data x_i^j as the sum of modality-common data and modality-unique data explicitly, i.e.,

$$x_i^j = x_{ci}^j + x_{ui}^j \tag{1}$$

where x_{ci}^j and x_{ui}^j are the modality-common data and the modality-unique data of the instance x_i^j respectively.

After that, we extract the modality-common features and modality-unique features from x_{ci}^j and x_{ui}^j. To complete this task, we take advantage of the autoencoders since it can preserve the local structure of data during the clustering process[4]. The autoencoders contain two parts, one is an encoder and the other is a decoder.

In the encoder part, we employ one encoder to extract the modality-common features and modality-unique features of each modality respectively. Denote $f_j(\cdot)$ as the encoder mapping function, we have,

$$f_j(x_i^j) = f_j(x_{ci}^j + x_{ui}^j) = z_{ci}^j + z_{ui}^j \tag{2}$$

where z_{ci}^j and z_{ui}^j are the latent modality-common features and modality-unique features for x_{ci}^j and x_{ui}^j respectively.

In the decoder part, we use two decoders to decode z_{ci}^j and z_{ui}^j respectively. Denote $\tilde{x}_{ci}^j/\tilde{x}_{ui}^j$ as the reconstructed modality-common/modality-unique data of the instance x_i^j, the process of obtaining $\tilde{x}_{ci}^j$ and $\tilde{x}_{ui}^j$ through the decoder $g_{cj}(\cdot)$ and $g_{uj}(\cdot)$ can be expressed as,

$$\tilde{x}_{ci}^j = g_{cj}(z_{ci}^j), \quad \tilde{x}_{ui}^j = g_{uj}(z_{ui}^j) \tag{3}$$

Denote $\tilde{x}_i^j$ as the reconstructed x_i^j, then, $\tilde{x}_i^j = \tilde{x}_{ci}^j + \tilde{x}_{ui}^j$. Using L_2 distance, the reconstruction loss function to extract modality-common features and modality-unique features is defined as following,

$$\begin{aligned}\mathcal{L}_r &= \sum_{j=1}^{m}\sum_{i=1}^{n} ||x_i^j - \tilde{x}_i^j||^2 = \sum_{j=1}^{m}\sum_{i=1}^{n} ||x_i^j - (\tilde{x}_{ci}^j + \tilde{x}_{ui}^j)||^2 \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n} ||x_i^j - (g_{cj}(z_{ci}^j) + g_{uj}(z_{ui}^j)))||^2\end{aligned} \tag{4}$$

2.1.2 Cross-Modality Reconstruction. Cross-modality reconstruction is used to establish the relationship among multiple modalities. Specifically, it maps the modality-common features and modality-unique features of one modality to other modalities. Cross-modality reconstruction is expressed as,

$$\hat{x}_{ci}^{jh} = g_{cj}(z_{ci}^h), \quad \hat{x}_{ui}^{jh} = g_{uj}(z_{ui}^h) \tag{5}$$

where $\hat{x}_{ci}^{jh}/\hat{x}_{ui}^{jh}$ is the cross reconstructed data obtained from the common/unique features of the h-th modality z_{ci}^h/z_{ui}^h through the j-th modality decoder $g_{cj}(\cdot)/g_{uj}(\cdot)$, $j,h \in [1,m]$ and $h \neq j$.

Theoretically, there are high similarity between modality-common features of different modalities and a great difference between the modality-unique features of different modalities. Then, we deal with the modality-common features and the modality-unique features respectively.

(1) For the modality-common features, the reconstruction data $\hat{x}_{ci}^{jh}$ should be similar to the data $\tilde{x}_{ci}^j$ reconstructed by the decoder $g_{cj}(z_{ci}^j)$, that is $\hat{x}_{ci}^{jh} \approx \tilde{x}_{ci}^j$. Using L_2 distance, the loss for the cross-modality reconstruction of modality-common features can be expressed as,

$$\begin{aligned}\mathcal{L}_c &= \sum_{j=1}^{m}\sum_{h=1,h\neq j}^{m}\sum_{i=1}^{n} ||x_i^j - (\hat{x}_{ci}^{jh} + \tilde{x}_{ui}^j)||^2 \\ &= \sum_{j=1}^{m}\sum_{h=1,h\neq j}^{m}\sum_{i=1}^{n} ||x_i^j - (g_{cj}(z_{ci}^h) + g_{uj}(z_{ui}^j)))||^2\end{aligned} \tag{6}$$

The purpose of $\mathcal{L}_c$ is to make $\hat{x}_{ci}^{jh} + \tilde{x}_{ui}^j$ approach x_i^j infinitely, which can reflect that $\hat{x}_{ci}^{jh}$ is very similar to $\tilde{x}_{ci}^j$ from the side, and further, it can be concluded that z_{ci}^h, the common feature of the modality, is similar to z_{ci}^j.

(2) For the cross-modality unique features, the reconstruction data $\hat{x}_{ui}^{jh}$ should be very different from the data $\tilde{x}_{ui}^j$ reconstructed by the decoder $g_{uj}(z_{ui}^j)$. We use radial basis function as the regularization item to constrain $\hat{x}_{ui}^{jh}$ stay away from $\tilde{x}_{ui}^j$. The loss for the cross-modality reconstruction of modality-unique features can be expressed as,

$$\mathcal{L}_u = \sum_{j=1}^{m}\sum_{h=1,h\neq j}^{m}\sum_{i=1}^{n} \frac{\exp(-\|g_{uj}(z_{ui}^h) - g_{uj}(z_{ui}^j)\|^2)}{2\sigma^2} \tag{7}$$

where σ is the width parameter of the function, which controls the radial range of the function. We set σ as the median of the pairwise Euclidean distances between the data points.

2.1.3 Joint Loss. By synthesizing the above objectives, the overall optimization problem of obtaining the modality-common and modality-unique features is formulated as:

$$\min \mathcal{L} = \mathcal{L}_r + \beta \mathcal{L}_c + \gamma \mathcal{L}_u \tag{8}$$

where $\beta, \gamma > 0$ are hyper-parameters. β controls the weight of the cross reconstruction regularization term and γ controls the weight of the radial basis function regularization term.

2.2 Modality Feature Fusion

Since multiple modalities share the same clustering results, we fuse the modality-common features of different modalities to obtain a consistent feature shared by all modalities, and cluster the data with this shared consistent feature. We use the fusion layer to extract the consistent feature data. The fusion layer uses a full connection layer, and $z_i^* = Fusion(z_{ci}^1, z_{ci}^2, \ldots, z_{ci}^m; w)$ is a shared consistent features of the modality fusion, where w is the parameter of the fusion layer. To train the fusion layer, we minimize the following loss function between z_i^* and the modality-common features z_{ci}^j of all modalities,

$$\min_w \sum_{j=1}^{m} \sum_{i=1}^{n} ||z_i^* - z_{ci}^j||^2 \tag{9}$$

Notice that, we only use the modality-common features of each modality to learn the fused consistent feature. Comparing with the approaches merely extracting the modality-common features, the proposed method refines the modality-common features through excluding the modality-unique knowledge explicitly.

2.3 The Algorithm

The training process of the proposed algorithm (named as DCUMC) is divided into two steps. In the first step, we use *Eq.*(8) to train the encoders and decoders to extract modality features z_{ci}^j and z_{ui}^j. In the second step, we train the fusion layer through *Eq.*(9) and obtain z_i^*.

Suppose the maximum number of neurons in each layer of the encoder/decoder/modality feature fusion network is $\widetilde{D}_1/\widetilde{D}_2/\widetilde{D}_3$, and maximum epochs for deep commonness and uniqueness mining/modality feature fusion is T_1/T_2. Then the time complexity of the feature extraction stage is $O(T_1 nm\widetilde{D}_1^2 + T_1 nm\widetilde{D}_2^2)$, the time complexity of cross-modality reconstruction stage is $O(T_1 nm^2\widetilde{D}_2^2)$ and the time complexity of the modality feature fusion stage is $O(T_2 nm\widetilde{D}_3^2)$. So the total time of DCUMC is $O((T_1\widetilde{D}_1^2 + T_1\widetilde{D}_2^2 + T_2\widetilde{D}_3^2)nm + T_1 nm^2\widetilde{D}_2^2)$ which is polynomial order to the number of modalities m and the number of examples n. Since the number of modalities in the real world is generally not very large, the DCUMC algorithm is theoretically efficient and scalable.

3 EXPERIMENT

3.1 Experimental Setup

3.1.1 Datesets. We experiment on following benchmark multimodal datasets: **AwA**[1]: It contains 5814 instances which is divided into 10 classes. The three modalities of this dataset respectively are 2000D local self-similarity feature, 2000D SIFT feature and 2000D SURF feature. **Caltech**101[2] and **Scene**-15 [3]: We extract 254D LBP, 512D GIST and 256D CENTRIST descriptors from these datasets as three modalites. Caltech101 contains 712 instances of 10 clusters and Scene-15 contains 3000 instances of 15 clusters. **CUB**[3] and **Flowers**[4]: We consider the 1024D image features extracted by Googlenet and 1024D corresponding text features [7] as two modalities. CUB contains 2889 instances of 50 clusters and Flowers contains 3235 instances of 50 clusters.

3.1.2 Baselines. (1) Single-modality clustering algorithm: DEC[12], IDEC[4] and SpectralNet [8]. We cluster each modality and reported the best performance. (2) Traditional multimodal clustering algorithm: DIMSC[2] and ECMSC[11]. (3) Deep multimodal clustering algorithm: DCCAE[10], DMF_MVC [13], DMSCN[1] and MvSCN[5]. In order to use the two-modality method DCCAE to process three modality datasets, we use all combination of two modalities to train the DCCAE. Then the mean values were calculated as the final result. (4) To show the influence of modality-unique features, we also report the result of DCUMC without modality-unique features (DCMC), which only extracts and cross-reconstructs modality-common features from each modality.

3.1.3 Parameter Settings. All the parameter settings of the baselines are based on the original papers. For the proposed DCUMC, the encoder and decoder consist of full connection layers. We use *tanh*/*sigmoid* as the activation function on the last layer of the encoder and decoder which are relevant to modality-common/modality-unique features. Meanwhile, *RelU* is used as the activation function on other layers of encoders and decoders. The depth of neural network is adjusted according to the dimension of input data. In the training process, the parameters of the autoencoders and fusion layers are randomly initialized, and the learning rate α of Adam is set as 0.001. The weight β and γ are set as 0.3 and 0.1 respectively. After the training process, we use k-means to cluster the shared consistent features z_i^*. Our implementation is based on Pytorch.

3.2 Experiment Result

We use two widely used metrics to measure the clustering performance: accuracy (ACC) and Normalized mutual information (NMI). The clustering results of the two metrics are presented in the Table 1 and Table 2 respectively. In each column of the two tables, the best result is highlighted in boldface. From the results, we find that DCUMC is superior to single modal algorithms (DEC, IDEC and SpectralNet) and traditional multimodal clustering algorithms (DIMSC and ECMSC) on each dataset in terms of ACC and NMI. The results indicate that it is reasonable to study multimodal clusteing and the deep clustering is more effective for feature extraction tasks. Comparing with the deep multimodal methods (DCCAE, DMF_MVC, DMSCN and MvSCN), DCUMC outperforms them generally. As a exceptional case, the NMI of MvSCN is slightly higher than that of DCUMC on Scene-15 dataset, but the ACC of DCUMC are superior to MvSCN. Moreover, the performance of DCUMC outperforms DCMC. In summary, we conclude that

[1] https://cvml.ist.ac.at/AwA/

[2] http://www.vision.caltech.edu/Image\Datasets/Caltech101/

[3] http://www.vision.caltech.edu/visipedia/CUB-200-2011.html

[4] http://www.robots.ox.ac.uk/~vgg/data/flowers/102/

DCUMC achieves good performance on multimodal data. The extracting of modality-unique features optimize the process of extracting modality-common features so as to obtain better clustering results.

Table 1: Clustering accuracy(%).

dataset	AwA	Caltech101	Scene-15	CUB	Flower
DEC	16.10	35.74	16.97	20.11	17.64
IDEC	20.08	43.54	25.03	30.16	28.11
SpectralNet	20.06	40.73	46.33	19.03	27.26
DIMSC	20.79	39.47	23.87	25.16	29.46
ECMSC	22.19	53.90	42.70	47.91	45.24
DCCAE	24.31	62.85	34.95	16.86	20.80
DMF_MVC	17.18	30.34	30.87	22.88	23.99
DMSCN	21.93	52.95	27.59	49.43	62.01
MvSCN	19.99	48.31	45.00	17.51	30.41
DCMC	24.52	68.25	43.93	44.44	55.67
DCUMC	**25.61**	**72.61**	**54.00**	**53.06**	**63.30**

3.3 Hyper-Parameter Analysis

In this subsection, we analyze the effect of the hyper-parameters β and γ for the performance of DCUMC. Due to space limitation, we only present the results on the Caltech101 dataset. We set β vary in the range $[0.1, 0.3, 0.5, 0.7, 0.9]$ and γ vary in the range $[0.0, 0.1, 0.3, 0.5, 0.7, 0.9]$. Table 3 shows how the performance of DCUMC varies with different β and γ. From Table 3, we find that the performance of DCUMC is not satisfactory when γ is set to 0. In this case, the constraint of the modality-unique features on the modality-common features is removed, which makes it difficult for the modality-common features to accurately express the shared consistent features of the whole data. Moreover, DCUMC achieves stably good performance when $\beta = 0.3$ and γ vary in $[0.1, 0.3, 0.5, 0.7]$. Then, we finally selected $\beta = 0.3$ and $\gamma = 0.1$ as the hyper-parameter values of the DCUMC model in the above experiment.

4 CONCLUSION

In this paper, we propose a novel multimodal clustering algorithm DCUMC which consists of two steps. In the first step, we use one encoder to extract the latent modality-common and modality-unique features for each modality, and then use two decoders to reconstruct the modality-common data and modality-unique data of each modality. At the same time, the modality features are cross-reconstructed in other decoders. The second step is to fuse the modality-common features of different modalities to obtain the shared consistent features that are used to cluster dataset subsequently. Experiments have been carried out to verify the effectiveness of the proposed method. In the future, we will analysis the common and unique distribution of each modality.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No.61806034; No.61876028; No.61602081).

Table 2: Clustering NMI(%).

dataset	AwA	Caltech101	Scene-15	CUB	Flower
DEC	4.10	22.63	16.10	33.69	34.56
IDEC	6.73	31.00	19.92	42.97	42.67
SpectralNet	3.28	31.49	47.35	38.30	41.45
DIMSC	7.61	33.74	19.39	39.26	43.41
ECMSC	7.33	51.53	41.76	56.63	61.76
DCCAE	10.13	57.96	37.82	32.58	38.93
DMF_MVC	4.43	19.78	31.89	36.95	39.91
DMSCN	5.98	43.83	27.59	59.65	74.15
MvSCN	3.11	38.60	**49.65**	32.01	44.27
DCMC	10.77	63.43	43.80	60.15	67.85
DCUMC	**10.89**	**65.82**	49.17	**65.62**	**76.63**

Table 3: Analysis Of Parameter Setting

γ \ β	0.1	0.3	0.5	0.7	0.9
0.0	66.01	64.04	66.01	63.90	54.78
0.1	63.48	**72.61**	63.06	65.31	65.73
0.3	64.61	**70.37**	60.81	66.15	62.64
0.5	71.21	**69.94**	63.90	61.52	58.71
0.7	64.04	**70.65**	69.24	55.48	64.61
0.9	61.94	61.94	60.11	61.10	65.87

REFERENCES

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

[2] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. 2015. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–594.

[3] Li Fei-Fei and Pietro Perona. 2005. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. IEEE, 524–531.

[4] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. 2017. Improved deep embedded clustering with local structure preservation.. In *IJCAI*. 1753–1759.

[5] Zhenyu Huang, Joey Tianyi Zhou, Xi Peng, Changqing Zhang, Hongyuan Zhu, and Jiancheng Lv. 2019. Multi-view Spectral Clustering Network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 2563–2569.

[6] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *ICML*.

[7] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 49–58.

[8] Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. 2018. Spectralnet: Spectral clustering using deep neural networks. *arXiv preprint arXiv:1801.01587* (2018).

[9] Nitish Srivastava and Ruslan Salakhutdinov. 2012. Multimodal Learning with Deep Boltzmann Machines. *J. Mach. Learn. Res.* 15 (2012), 2949–2980.

[10] Weiran Wang, Raman Arora, Karen Livescu, and Jeff A. Bilmes. 2015. On Deep Multi-View Representation Learning. In *ICML*.

[11] Xiaobo Wang, Xiaojie Guo, Zhen Lei, Changqing Zhang, and Stan Z Li. 2017. Exclusivity-consistency regularized multi-view subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 923–931.

[12] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*. 478–487.

[13] Handong Zhao, Zhengming Ding, and Yun Fu. 2017. Multi-view clustering via deep matrix factorization. In *Thirty-First AAAI Conference on Artificial Intelligence*.