CROSS-VIEW EQUIVARIANT AUTO-ENCODER

Zhibin Wan[1], *Changqing Zhang*[1*], *Yu Geng*[1], *Huazhu Fu*[2], *Xi Peng*[3], *Pengfei Zhu*[1], *Qinghua Hu*[1]

[1]College of Intelligence and Computing, Tianjin University, China
[2]Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
[3]College of Computer Science, Sichuan University, China
{wanzhibin, zhangchangqing}@tju.edu.cn

2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428184

ABSTRACT

Unsupervised representation learning on multi-view data (multiple types of features or modalities) becomes a compelling topic in machine learning. Most existing methods focus on directly projecting different views into a common space to explore the consistency across different views. Although simple, the underlying relationships among different views are not guaranteed during the learning process. In this paper, we propose a novel unsupervised multi-view representation learning model termed as Cross-View Equivariant Auto-Encoder (CVE-AE), which jointly conducts data reconstruction with view-specific autoencoder for information preservation within each view, and transformation reconstruction with transformation decoder for correlations preservation across different views. Accordingly, the generalization ability of our model is promoted due to the preserved intra-view intrinsic information and underlying inter-view relationships. We conduct extensive experiments on real-world datasets, and the proposed model achieves superior performance over state-of-the-art unsupervised representation learning methods.

Index Terms— Multi-view learning, transformation equivariance, representation learning, auto-encoder, unsupervised algorithm

1. INTRODUCTION

In many real-world applications, data are usually described from different perspectives that are considered as multiple views. Recently, numerous methods have been proposed to jointly utilize multiple types of features [1] or multiple modalities of data [2]. Most multi-view representation learning algorithms [3] consider integrating different views into a unified representation, which is of vital importance for various tasks [4, 5] since unified representation could be easily exploited by off-the-shelf algorithms. Although existing methods are effective, exploring multiple views is still a long-standing challenge due to the complex correlations among different views. Furthermore, different views are usually highly heterogeneous owing to the diversity of feature extractions and the data collections. Therefore, it is important and necessary to explore the information and correlations within multiple views to learn a powerful representation that can be used for downstream tasks.

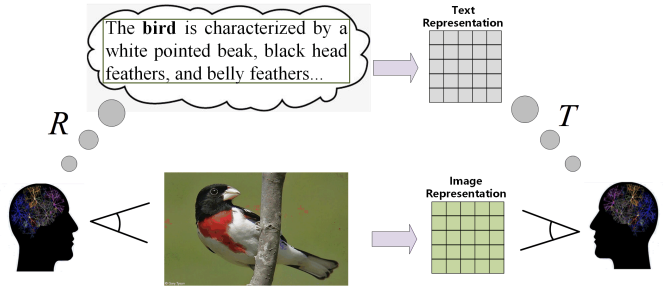

Fig. 1. Give an image, people can easily associate a textual description to the visual image due to the underlying relationships R among different modalities. However, for different representations learned from the original modalities, it is usually difficult to obtain a representation of one view according to the representation of another view because the transformation T is usually not well preserved.

However, most existing multi-view representation learning methods [6] mainly focus on maximizing the consistency of different views by projecting different views into a common subspace. Although simple, the underlying relationships (transformations) among different views are not guaranteed during the learning process. As a result, the underlying relationships are not well preserved among the learned representations. As shown in Fig. 1, for multi-view data, there are not only intra-view information, but also underlying inter-view transformations which indicate stable relationships among different views. The generalization ability of the learned model may be harmed when the underlying relationships among different views are ignored.

*Corresponding author: Changqing Zhang

This work was supported in part by National Natural Sci- ence Foundation of China (No. 61976151 and 61876127), the Natural Science Foundation of Tianjin of China (No. 19JCYBJC15200).

Therefore, we propose the Cross-View Equivariant Auto-Encoder (CVE-AE), which jointly conducts cross-view transformations reconstruction with transformation decoder to preserving the underlying relationships, and view-specific reconstruction with autoencoder to extract the intrinsic information. Specifically, our model consists of two main components, i.e., view-specific autoencoders and transformation decoder. The view-specific autoencoders are responsible for extracting intrinsic information from each view, while the transformation decoder ensures the relationships among different views to be preserved in the learned representations. Accordingly, the learned representation will be transformation equivariant which promotes cross-view consistency and reveals intrinsic relationships among different views. Compared with existing state-of-the-art unsupervised representation learning methods, the proposed CVE-AE can achieve very impressive performance on different tasks. In summary, the main contributions are summarized as follows:

- We propose a novel unsupervised multi-view representation learning framework: Cross-View Equivariant Auto-Encoder, which can flexibly learn a multi-view representation from heterogeneous views. As the best of our knowledge, this is the first work encoding the relationships among different views into learned representations with transformation equivariance.
- The proposed model jointly considers both view-specific intrinsic information with sample reconstruction by using view-specific autoencoders and underlying relationships with transformation reconstruction by using equivariant decoder, which can promote the generalization ability of the model.
- Extensive experiments verify the advantages of the proposed model. Compared with existing state-of-the-art unsupervised representation learning methods, our model achieves impressive performances on different downstream tasks.

2. RELATED WORK

Transformation equivariance which means the transformation of the obtained feature representation should be the same as the transformation of the original data. The research of transformation equivariance can be traced back to the capsule nets [7], where the capsules are designed to be equivariant to various transformations. However, it is difficult to guarantee the resultant capsules have the transformation equivariance [8]. Recently, a flexible method [9] is proposed which learns a representation by reconstructing the transformation. The method in [10] proposes an affine equivariant autoencoder to learn features that are equivariant to affine transformation.

Multi-view representation learning has attracted intensive attention recently [11]. Particularly, unsupervised algorithm is a rather challenging problem because there is no guidance during the learning process. Most methods are based on CCA [12], which basically map multiple types of features onto a common subspace by maximizing the correlation. To explore nonlinear correlations, DCCA [13] extends CCA using deep neural networks, deep CCA [6] further extends CCA with deep neural networks. Different from CCA, Multi-view Dimensionality co-reduction (MDcR) [14] applies the kernel matching to regularize the dependence across multiple views. Inspired by deep learning, semi-nonnegative matrix factorization is utilized to find a common representation including consistent information of multiple views [15]. Lately, a nested autoencoder [16] is developed which integrates information from heterogeneous sources into an intact representation for multi-view data.

3. PROPOSED METHOD

3.1. Formulation

Suppose a multi-view dataset $\mathcal{X} = \left\{X^{(v)} \in \mathbb{R}^{d_v \times n}\right\}_{v=1}^{V}$, where $X^{(v)}$ is the feature matrix of the vth view with V, d_v and n being the number of views, the dimensionality of feature space in the $v$the view and the number of samples, respectively. Given two views $X^{(i)}$ and $X^{(j)}$ from the multi-view dataset, the transformation T indicates the underlying relationships between two views. To learn the representation of each view, a encoding function $f : X^{(v)} \rightarrow Z^{(v)}$ and a decoding function $g : Z^{(v)} \rightarrow \overline{X}^{(v)}$ are defined. And the transformation $\overline{T}$ represents the underlying relationships between the reconstructed views.

In this paper, we hope that the transformations $\overline{T}$ among the reconstructed views and the transformations T of the original views are as similar as possible to more relaxedly satisfy the transformation equivariance. Specifically, the underlying transformation can be decoded from the learned representations which not only encode intrinsic view-specific information but also the underlying cross-view relationships.

3.2. Cross-View Equivariant Auto-Encoder

Based on the above motivation, we propose the Cross-View Equivariant Auto-Encoder to encode the intrinsic information within each view and transformation equivariance property. Firstly, we obtain the transformation relationship T_{ij} of any two views $X^{(i)}$ and $X^{(j)}$ with the following objective function

$$\min_{T} \left\| T_{ij} X^{(i)} - X^{(j)} \right\|_F^2 . \tag{1}$$

The goal of the proposed CVE-AE (as presented in Fig. 2) is to learn a comprehensive and powerful representation that can not only fully retain the information of each view, but also the transformations among different views. Therefore, our model mainly consists of two components. The first one

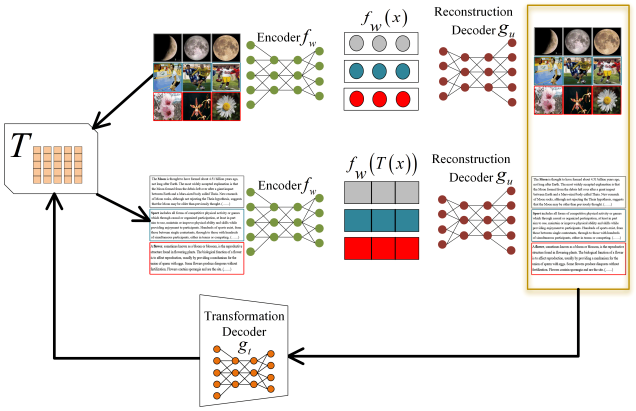

Fig. 2. Overview of the proposed model. The key components consist of the view-specific autoencoder and transformation decoder, which can obtain the low-dimensional representations preserving intrinsic information of each single view and underlying relationships across different views, respectively.

view-specific autoencoder which aims to extract the intrinsic information from each view by reconstruction. The second component is to ensure that the learned representation maintains the transformation relationships.

For the first component, we employ an encoder network $f_w^{(v)}(\cdot)$ which aims to extract the low-dimensional representation with $Z^{(v)} = f_w^{(v)}\left(X^{(v)}\right)$ for the vth view. Meanwhile, we aim to learn a decoder network $g_u^{(v)}(\cdot)$ to map $Z^{(v)}$ back to $X^{(v)}$, where w and u are the sets of weight in encoder and reconstruction decoder networks, respectively. To obtain the reconstructed representation $Z^{(v)}$, we minimize the following reconstruction loss

$$\min_{w,u} \sum_{v=1}^{V} \left\| X^{(v)} - g_u^{(v)}\left(f_w^{(v)}\left(X^{(v)}\right)\right)\right\|_F^2. \tag{2}$$

With the obtained low-dimensional representations of different views, we focus on encoding the transformation equivariance into them.

On the one hand, we consider taking advantage of same encoder network $f_w^{(v)}(\cdot)$ to extract the representation $Z^{(v)} = f_w^{(v)}\left(X^{(v)}\right)$ of each single view. On the other hand, we define a completely different decoder network $g_t(\cdot)$ parameterized with t. The transformation decoder aims to recover the transformation relationships among different views. Note that, since the reconstruction for the transformation is ensured through the reconstructed views which obtained from the representations, it forces the model to extract expressive and correlative features as a proxy of the original views.

Specifically, we reconstruct the transformation relationship with the transformation decoder network: $\overline{T}_{ij} = g_t^{(i,j)}\left(g_u^{(i)}\left(f_w^{(i)}\left(X^{(i)}\right)\right), g_u^{(j)}\left(f_w^{(j)}\left(X^{(j)}\right)\right)\right)$. The decoder network is trained by minimizing the difference between the predicted transformation $\overline{T}_{ij}$ and the original transformation T_{ij} for the purpose of learning the equivariant representations. Accordingly, the transformation autoencoder network can be optimized by minimizing the loss as

$$\begin{aligned} &\min_{w,t} \sum_{i=1}^{V}\sum_{\substack{j=1\\ j\neq i}}^{V} \left\|T_{ij} - \overline{T}_{ij}\right\|_F^2, \\ &\overline{T}_{ij} = g_t^{(i,j)}\left(g_u^{(i)}\left(f_w^{(i)}\left(X^{(i)}\right)\right), g_u^{(j)}\left(f_w^{(j)}\left(X^{(j)}\right)\right)\right). \end{aligned} \tag{3}$$

By jointly considering view-specific intrinsic information with data reconstruction and underlying relationships with transformation reconstruction, the objective is induced as

$$\min_{T,w,u,t} \left\|\mathcal{L}_t\right\|_F^2 + \mu\left\|\mathcal{L}_e\right\|_F^2 + \lambda\left\|\mathcal{L}_r\right\|_F^2, \tag{4}$$

where $\mathcal{L}_t(\cdot)$ represents the loss of learning the transformation relationships between any two views, while $\mathcal{L}_e(\cdot)$ denotes the loss functions associated with the underlying structural relationships reconstruction (transformation reconstruction), and $\mathcal{L}_r(\cdot)$ denotes the loss functions related to the view-specific intrinsic information reconstruction (data reconstruction)

$$\begin{aligned} \mathcal{L}_t &= \sum_{i=1}^{V}\sum_{\substack{j=1\\ j\neq i}}^{V}\left(T_{ij}X^{(i)} - X^{(j)}\right), \\ \mathcal{L}_e &= \sum_{i=1}^{V}\sum_{\substack{j=1\\ j\neq i}}^{V}\left(T_{ij} - g_t^{(i,j)}\left(\overline{X}^{(i)}, \overline{X}^{(j)}\right)\right), \\ \mathcal{L}_r &= \sum_{i=1}^{V}\left(X^{(i)} - g_u^{(i)}\left(f_w^{(i)}\left(X^{(i)}\right)\right)\right), \end{aligned} \tag{5}$$

where $\mu > 0$ and $\lambda > 0$ are trade-off factors to balance the satisfaction of equivariance and data reconstruction. For all views, the proposed model automatically learns low-dimensional representations by view-specific autoencoders and transformation decoders. It is worth noting that although the proposed CVE-AE is an unsupervised representation learning model, it has powerful generalization and can be extended to meet specific tasks.

4. EXPERIMENTS

In the experiments, we compare the proposed CVE-AE with other unsupervised multi-view representation learning methods on five real-world datasets with multiple views, and then evaluate the results with commonly used evaluation metrics.

4.1. Datasets

We employ five multi-view benchmark datasets in the experiments. **Handwritten**[1] contains 2000 examples from number

[1] https://archive.ics.uci.edu/ml/datasets/Multiple+Features

Table 1. Performance comparison on clustering task

Datasets	Metrics	CCA	DCCA	DCCAE	MDcR	DMF	AE^2-Nets	Ours
Handwritten	ACC	66.42±4.81	66.16±1.16	69.29±1.02	76.77±2.37	71.85±3.55	84.52±1.82	**90.35±1.28**
	NMI	69.66±4.06	66.04±0.49	66.95±0.91	76.70±0.82	73.11±2.23	76.09±1.50	**82.75±1.43**
	F-score	62.06±4.77	59.09±0.38	60.50±1.30	71.93±2.23	66.67±2.97	73.95±1.85	**81.27±1.74**
	RI	91.87±1.34	91.36±0.08	91.76±0.22	94.11±0.47	92.86±1.01	95.41±0.38	**96.87±0.39**
ORL	ACC	56.96±2.04	59.64±2.20	59.42±2.06	61.69±2.17	65.36±2.88	68.85±2.11	**71.50±2.53**
	NMI	76.01±0.79	77.82±0.86	77.54±0.83	79.45±1.26	82.86±1.21	84.05±0.78	**87.62±0.84**
	F-score	45.10±1.87	47.71±2,05	46.69±2.27	48.50±2.59	52.03±3.34	63.61±1.31	**68.14±2.33**
	RI	97.29±0.10	97.40±0.14	97.37±0.13	97.28±0.28	97.32±0.22	97.94±0.11	**98.33±0.19**
COIL20	ACC	58.64±1.39	63.71±1.08	62.72±1.41	64.25±2.88	53.93±5.06	73.42±1.90	**75.04±1.53**
	NMI	70.60±0.75	75.99±1.15	76.32±0.66	79.43±1.37	72.35±2.33	82.55±1.01	**83.56±0.57**
	F-score	53.09±1.40	58.74±0.57	57.56±1.12	63.58±2.56	46.39±4.39	69.38±1.91	**73.06±1.29**
	RI	95.15±0.22	95.57±0.10	95.27±0.32	96.09±0.29	92.57±1.28	96.86±0.22	**97.03±0.24**
Caltech101	ACC	45.35±0.13	56.50±3.05	53.93±5.78	46.57±0.67	55.67±2.67	**62.17±2.78**	60.92±1.18
	NMI	50.52±0.13	57.64±3.75	53.94±3.73	56.43±0.56	45.56±2.18	60.38±4.12	**61.89±1.68**
	F-score	53.51±0.19	62.32±5.07	57.57±3.10	51.56±0.56	57.70±2.25	**66.24±2.17**	63.69±1.92
	RI	73.25±0.16	76.31±2.46	74.12±2.78	73.30±0.40	73.43±2.73	**80.36±1.78**	78.59±1.25
CUB	ACC	45.85±1.46	54.49±0.29	66.72±1.52	73.69±3.23	37.55±2.61	73.75±1.63	**74.45±0.34**
	NMI	46.60±0.58	52.51±1.09	65.77±1.36	**74.50±0.75**	37.84±2.03	72.61±1.62	74.27±1.44
	F-score	39.90±1.28	45.85±0.31	58.21±1.12	65.73±1.23	28.96±1.61	68.96±2.03	**69.62±1.10**
	RI	87.41±0.46	88.63±0.09	91.24±0.25	92.79±0.43	85.56±0.30	92.92±0.63	**93.10±0.20**

0 to 9 with 200 samples per class. In the experiments, we extracted two types of features as two views. **ORL**[2] includes 40 classes, and each class contains 10 face images. The intensity, LBP and Gabor features are extracted as different views. **COIL-20**[3] consists of 1440 pictures of 20 categories. Three types of features (intensity, LBP and Gabor) are used as different views. **Caltech101-7**[4] contains 1,474 images in six views, which is a subset of the original Caltech101 image dataset. **CUB**[5] has 200 different bird categories, including 11788 bird pictures and corresponding text descriptions. The image features are extracted by GoogLeNet, and the text features are extracted by Doc2Vec. The image and text features are used as two different views.

4.2. Compared methods

We compare our approach with the following unsupervised multi-view representation learning methods, including: **CCA** [12], **DCCA** [13], **DCCAE** [6], **MDcR** [14], **DMF-MVC** [15], and **AE^2-Nets** [16].

4.3. Implementation details

There are two parts in the proposed CVE-AE, view-specific autoencoder and transformation decoder. The network architecture of each component employs the 3-layer fully connected network with ReLU activation function. We also employ the Adam optimizer [17] with 10^{-3} learning rate for the datasets. In addition, CVE-AE is implemented via PyTorch, and all the experiments are conducted on NVIDIA Geforce GTX TITAN Xp GPU.

[2]https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase

[3]http://www.cs.columbia.edu/CAVE/software/softlib/coil-20

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

[5]http://www.vision.caltech.edu/visipedia/CUB-200

4.4. Experimental results

We evaluate the proposed CVE-AE on clustering task. Specifically, we utilize CVE-AE and other compared methods to learn the multi-view representations, then we employ k-means algorithm to evaluate the learned representations. We also adopt four different metrics: Accuracy (ACC), Normalized Mutual Information (NMI), F-score and Rand Index (RI). Employing different metrics can reflect different clustering characteristics, while it is consistent that the higher the value, the better the clustering performance.

Table 1 shows the performances of different multi-view methods on clustering task. Obviously, our algorithm almost outperforms other methods on most datasets. The reason for this observation is that the proposed model can comprehensively learn the intrinsic view information and underlying transformation relationships among heterogeneous multiple views, which can significantly improve the accuracy and stability of the clustering task. In short, the superior performance validates the advantages of our CVE-AE.

5. MODEL ANALYSIS

In order to further illustrate our proposed method, we conduct model analysis experiments.

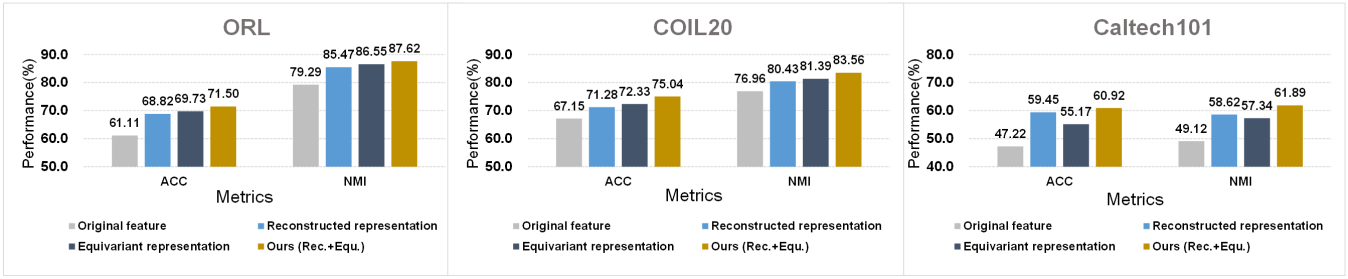

Fig. 3. Performance comparison of different representations

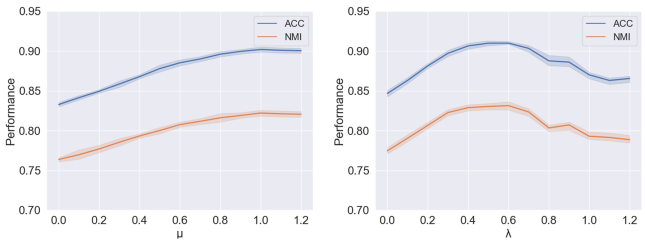

Fig. 4. The study on parameter sensitivity for CVE-AE

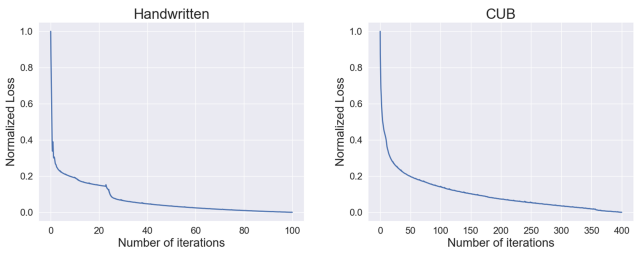

Fig. 5. Convergence experiments (where loss values are normalized to range $[0, 1]$).

5.1. Parameter Analysis

The hyperparameters μ and λ are essential for controlling the joint learning of equivariant representation and reconstructed representation. As shown in Fig. 4, we tune the hyperparameters on the handwritten dataset and illustrate the clustering performances of our model with different values.

5.2. Convergence Analysis

To demonstrate the convergence of the proposed method, we conduct the convergence experiments as shown in Fig. 5. The optimization process is basically stable, where the loss decreases quickly and converges within a number of iterations.

5.3. Ablation Study

The proposed CVE-AE jointly considers both view-specific intrinsic information with reconstructed representation and underlying relationships with equivariant representation, which can obtain a more competitive representation and promote the generalization ability of the model. In order to further analyze the proposed model and various representations, we conduct a series of clustering experiments under the conditions of original features, only reconstructed representation, only equivariant representation, and both considerations, respectively.

Table 2. Ablation study of the different representations

Datasets	Rec.	Equ.	ACC	NMI
Handwritten	√		76.85±1.81	75.94±1.57
		√	82.26±2.42	78.39±1.60
	√	√	**90.35±1.28**	**82.75±1.43**
CUB	√		68.77±0.86	70.08±1.36
		√	71.50±0.62	72.64±1.13
	√	√	**74.45±0.34**	**74.27±1.44**

[1] **Rec.** denotes the reconstructed representation, and **Equ.** means the equivariant representation.

Is the transformation decoder (equivariant representation) effective? In our model, we learn an equivariant representation that preserves the underlying structural information of multiple views. The obtained representation can reveal the latent connection among views and promote consistency across views, thereby improve the generalization ability of the model. Therefore, here we aim to prove that the equivariant representation learned with the transformation decoder can retain structural information to enhance the effectiveness of the representation and perform better on downstream tasks. It can be observed in Fig. 3 that equivariant representation achieves higher performance in terms of all metrics than the original features. And as shown in Table 2, compared with the only reconstructed representation, the performance of CVE-AE that considers the structural information among views and the view-specific intrinsic information is more satisfactory, which further empirically proves the superiority of the equivariant representation.

Is the view-specific autoencoder (reconstructed representation) effective? The reconstructed representation aims to retain the main information in the views, so as to re-

construct original views with a low-dimensional compact and lossless representation. The method has been proven effective and widely used. The Fig. 3 illustrates that the reconstructed representation is more advantageous than the original features. As shown in Table 2, it can be observed that both reconstructed representation and equivariant representation have better performance than the only equivariant representation. Therefore, the reconstructed representation containing rich information is effective.

Which of the learned representations is more competitive? In order to investigate the superiority of the above representations, we conducted comparison experiments. The results are shown in Fig. 3, where the equivariant representation is more powerful than the reconstructed representation, and that focusing on both equivariant and reconstructed representation achieve the best performance. We also note that in some cases the reconstructed representation may be more effective. The possible reason is that the intrinsic information is preserved which is the foundation of learned representations. However, our proposed CVE-AE which considers both internal connections and inherent information is much more competitive in a variety of settings.

6. CONCLUSIONS

In this paper, we have proposed a novel unsupervised representation learning model for heterogeneous data, which jointly conducts data reconstruction with autoencoder for view-specific information preservation, and transformation reconstruction with transformation decoder for the relationships preservation across different views. Therefore, the proposed model can adaptively obtain a comprehensive representation that simultaneously focuses on intra-view information and inter-view structural relationships. The experimental results indicate that the proposed CVE-AE has superior performances compared to state-of-the-art methods. In the future, we will consider the case where the transformation relationship is nonlinearity.

7. REFERENCES

[1] Paramveer Dhillon, Dean P Foster, and Lyle H Ungar, "Multi-view learning of word embeddings via cca," in *Advances in neural information processing systems*, 2011, pp. 199–207.

[2] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[3] Abhishek Kumar, Piyush Rai, and Hal Daume, "Co-regularized multi-view spectral clustering," in *Advances in neural information processing systems*, 2011, pp. 1413–1421.

[4] Meina Kan, Shiguang Shan, and Xilin Chen, "Multi-view deep network for cross-view classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4847–4855.

[5] Changqing Zhang, Huazhu Fu, Qinghua Hu, Xiaochun Cao, Yuan Xie, Dacheng Tao, and Dong Xu, "Generalized latent multi-view subspace clustering," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[6] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes, "On deep multi-view representation learning," in *International Conference on Machine Learning*, 2015, pp. 1083–1092.

[7] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang, "Transforming auto-encoders," in *International Conference on Artificial Neural Networks*. Springer, 2011, pp. 44–51.

[8] Taco Cohen and Max Welling, "Group equivariant convolutional networks," in *International conference on machine learning*, 2016, pp. 2990–2999.

[9] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo, "Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2547–2555.

[10] Xifeng Guo, En Zhu, Xinwang Liu, and Jianping Yin, "Affine equivariant autoencoder," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 2413–2419.

[11] Chang Xu, Dacheng Tao, and Chao Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[12] Harold Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*, pp. 162–190. Springer, 1992.

[13] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*, 2013, pp. 1247–1255.

[14] Changqing Zhang, Huazhu Fu, Qinghua Hu, Pengfei Zhu, and Xiaochun Cao, "Flexible multi-view dimensionality co-reduction," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 648–659, 2016.

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

[16] Changqing Zhang, Yeqing Liu, and Huazhu Fu, "Ae2-nets: Autoencoder in autoencoder networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2577–2585.

[17] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.