Multi-View Information-Bottleneck Representation Learning

Zhibin Wan,[1] **Changqing Zhang,**[1,2] **Pengfei Zhu,**[1,2]* **Qinghua Hu**[1,2]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Tianjin Key Lab of Machine Learning, Tianjin, China
{wanzhibin, zhangchangqing, zhupengfei, huqinghua}@tju.edu.cn

Abstract

In real-world applications, clustering or classification can usually be improved by fusing information from different views. Therefore, unsupervised representation learning on multi-view data becomes a compelling topic in machine learning. In this paper, we propose a novel and flexible unsupervised multi-view representation learning model termed Collaborative Multi-View Information Bottleneck Networks (CMIB-Nets), which comprehensively explores the common latent structure and the view-specific intrinsic information, and discards the superfluous information in the data significantly improving the generalization capability of the model. Specifically, our proposed model relies on the information bottleneck principle to integrate the shared representation among different views and the view-specific representation of each view, prompting the multi-view complete representation and flexibly balancing the complementarity and consistency among multiple views. We conduct extensive experiments (including clustering analysis, robustness experiment, and ablation study) on real-world datasets, which empirically show promising generalization ability and robustness compared to state-of-the-arts.

Introduction

For real-world applications, data are usually manifested in multiple types of features (Dhillon, Foster, and Ungar 2011) or multiple modalities (Ngiam et al. 2011; Baltrušaitis, Ahuja, and Morency 2018) that are considered as multiple views. For instance, an image can be described by color (e.g., color histogram (Novak, Shafer et al. 1992)) or texture descriptor (e.g., GIST (Oliva and Torralba 2001), SIFT (Lowe 2004), HOG (Dalal and Triggs 2005)). Basically, due to the diversity of feature extraction or data acquirement, various views are usually heterogeneous. The ubiquity of multi-view data has attracted tremendous attention to the multi-view representation learning (Sun 2013; Zhang et al. 2020). Furthermore, integrating different views into a compact representation is essential for the downstream specific tasks since the intact representation could be easily developed by off-the-shelf algorithms (Zhang et al. 2018; Liu et al. 2018). Generally, when information from different views complements each other, it can be expected that the multi-view representation learning approaches can improve performance (Tao et al. 2019). To effectively explore the multi-view data, a series of methods have been proposed in recent years (Wang et al. 2015). The representative ways are Canonical Correlation Analysis (CCA) (Hotelling 1992) and its variants, which mainly maximize the consistency of multiple views by projecting different views into a common subspace. However, the main drawback of the CCA-based algorithms is that it overemphasizes exploring common information, while the view-specific intrinsic information of each view is also important, which may degrade the quality of the learned representation.

Consequently, it is still a long-term challenge to jointly exploit the view-specific information of each view and the complex relationships among different heterogeneous views under the context of multi-view representation learning. In this work, to address the above issues, we propose an unsupervised multi-view representation learning method termed Collaborative Multi-View Information Bottleneck Networks (CMIB-Nets) based on the *information bottleneck* principle. The proposed CMIB-Nets aims to collaboratively encode and integrate the view-specific intrinsic information and shared latent structure from heterogeneous views into a comprehensive representation. Moreover, the model can also adaptively balance the complementarity and consistency among different views.

Specifically, we realize this by maximizing the mutual information between multi-view information-bottleneck representation and shared representation (learning the common structure), while at the same time maximizing the mutual information between multi-view information-bottleneck representation and view-specific representations (learning the view-specific information). We also minimize the mutual information between original views and multi-view information-bottleneck representation simultaneously (compressing data to remove useless information). The resulting multi-view information-bottleneck representation can integrate the advantages of other multi-view representations, which not only encodes the view-specific information and cross-view underlying structure but also reduces the influence of superfluous information in the multi-view data. Therefore, it can be more robust to downstream tasks and improve the generalization capability. Furthermore, by utilizing the deep neural networks, the obtained represen-

*Corresponding author

tation can explore complex relationships among different views. Compared with the other unsupervised multi-view representation learning algorithms, the proposed CMIB-Nets achieves impressive performance on various tasks with different settings.

The main contributions of this work are summarized as:

- We propose an unsupervised multi-view representation learning method - Collaborative Multi-View Information Bottleneck, which extends information bottleneck to unsupervised multi-view setting and flexibly learns a representation from heterogeneous data.
- The proposed model can integrate various representations with the help of information bottleneck, making the multi-view information-bottleneck representation to collaboratively learn intra-view intrinsic information and inter-view shared structure, and also reducing the influence of superfluous information in the data. The resulting representation can improve the robustness and generalization ability of downstream tasks.
- With the introduction of neural network and variational inference, the general correlations among multiple views could be researched and the mutual information can be approximated through the variational bound.
- Extensive experiments under various conditions verify the advantages of our proposed model. Compared with existing state-of-the-art unsupervised representation learning algorithms, our model achieves impressive performances and exhibits satisfactory generalization ability.

Related Work

Information Bottleneck

Information bottleneck (Tishby, Pereira, and Bialek 2000) is an approach based on information theory, which formally describes meaningful and relevant information in the data. The theory states that if the obtained representation discards information from the input which is not useful for a given task, it will increase robustness for the downstream tasks. Specifically, given the original data $\mathbf{x}$ with the label $\mathbf{y}$, the information bottleneck can obtain a compact and effective representation $\mathbf{z}$ of data $\mathbf{x}$. And the objective of the information bottleneck principle is as follows:

$$\max_{\mathbf{Z}} I(\mathbf{Y}, \mathbf{Z}) - \beta I(\mathbf{X}, \mathbf{Z}), \tag{1}$$

where β is a trade-off factor to balance $I(\mathbf{Y}, \mathbf{Z})$ and $I(\mathbf{X}, \mathbf{Z})$.

Besides, the information bottleneck principle is used in multi-view representation learning. (Xu, Tao, and Xu 2014) uses this theory to learn a multi-view representation. To explore the nonlinear relationships of multiple views, (Wang et al. 2019) proposes a deep information bottleneck model with deep neural networks. Recently, (Federici et al. 2020) analyzes the redundant information in multiple views to obtain a robust representation.

Multi-View Representation Learning

Multi-view representation learning is designed to explore the information from multiple views for better performances (Li, Yang, and Zhang 2018). To obtain a unified representation among multiple views, CCA-based (Hotelling 1992) algorithms project different views into a common subspace through maximizing correlations among different views. To explore nonlinear correlations, DCCA (Andrew et al. 2013) extends CCA with neural networks, DCCAE (Wang et al. 2015) combines CCA and autoencoders. (Zhao, Ding, and Fu 2017) utilizes matrix factorization to obtain a representation from multi-view data with specific constraints. Recently, AE^2-Nets (Zhang, Liu, and Fu 2019) learns an intact multi-view representation by a nested autoencoder.

Proposed Approach

In this work, we propose the CMIB-Nets for learning a robust and intact representation given a multi-view dataset $\mathcal{X} = \{\mathbf{X}^{(1)}, ..., \mathbf{X}^{(V)}\}$, where $\mathbf{X}^{(v)} \in \mathbb{R}^{d_v \times n}$ is the feature matrix of the vth view with V, n and d_v being the number of views, number of samples and dimensionality of feature space for the vth view, respectively. As shown in Fig. 1, we construct a multi-view information-bottleneck representation $\mathbf{Z}$ linked with the shared representation $\mathbf{H}$ and the view-specific representation $\mathbf{S}$ to explore the inter-view complex relationship and intra-view intrinsic information of different views, and then collaboratively learn a complete representation to improve the discriminative ability of the approach.

Consistency and Complementarity

Multi-view data generally contains consistent relationships and complementary information, (i.e., *consistency* and *complementarity*), which are necessary to improve the performance of the model. In our proposed model, the shared representation across different views can reveal inter-view common structural correlations, while the view-specific representation of multiple views can indicate intra-view exclusive intrinsic information.

Consistency: Shared Representation. To explore the complex associations among different views, inspired by the reconstruction method (Lee 1996; Zhang et al. 2017), our method assumes that the multiple views are originated from an underlying latent representation (White et al. 2012; Guo 2013). Given N multi-view observations which consist of V views, our method is to infer a shared representation $\mathbf{H}$ for each view. Intuitively, we can reconstruct each view in a stable way from the shared representation (e.g., $\mathbf{x}_i{}^{(v)} = f_v(\mathbf{h}_i)$), which can essentially describe the underlying structure shared by different views. Generally, we also assume that the shared latent representation $\mathbf{h}_i$ of an arbitrary sample $\mathbf{x}_i$ in each view is conditionally independent. For convenience, $\{\mathbf{x}_i^{(v)}\}_{v=1}^{V}$ is replaced by $\mathcal{D}_i$, and then we have

$$P(\mathcal{D}_i|\mathbf{h}_i) = \prod_{v=1}^{V} P(\mathbf{x}_i^{(v)}|\mathbf{h}_i), \tag{2}$$

which denotes the joint distribution of $P(\mathbf{x}_i^{(v)}|\mathbf{h}_i)$ in all views. We model the likelihood with condition $\mathbf{h}_i$ as

$$P(\mathbf{x}_i^{(v)}|\mathbf{h}_i) \propto e^{-\Delta(\mathbf{x}_i^{(v)}, f_v(\mathbf{h}_i; \mathbf{\Theta}_v))}, \tag{3}$$

where $\Delta(\mathbf{x}_i^{(v)}, f_v(\mathbf{h}_i; \mathbf{\Theta}_v))$ represents the reconstruction loss and $\mathbf{\Theta}_v$ are parameters of f_v. Accordingly, by assuming the data is independent and identically distributed (IID), we can get the log-likelihood function as follows:

$$\mathcal{L}(\mathbf{H}; \mathbf{\Theta}) = \sum_{i=1}^{N} lnP(\mathcal{D}_i|\mathbf{h}_i). \tag{4}$$

Since maximizing the likelihood is equivalent to minimizing the Δ loss function, we can obtain the following objective function $\mathcal{L}_h(\cdot)$ for learning the shared representation part:

$$\begin{aligned} \mathcal{L}_h &= \min_{\mathbf{H}, \mathbf{\Theta}} \sum_{v=1}^{V} \sum_{i=1}^{N} \Delta(\mathbf{x}_i^{(v)}, f_v(\mathbf{h}_i; \mathbf{\Theta}_v)) \\ &= \min_{\mathbf{H}, \mathbf{\Theta}} \sum_{v=1}^{V} ||\mathbf{X}^{(v)} - f_v(\mathbf{H}; \mathbf{\Theta}_v)||_F^2, \end{aligned} \tag{5}$$

where we use a reconstruction network for the transformation function f_v in vth view.

In this way, $\mathbf{H}$ can encode consistent information of multiple views, and different views are projected into a common space. Since the shared representation integrates information from different views, it could reveal the common latent structure shared by different views.

Complementarity: View-specific Representation. Apart from modeling the consistency across multiple views, it is also important to preserve the complementary information. Since it is basically difficult to separate the private information from common or shared information for each view, here we alternatively learn representations for each view independently to guarantee the private information is contained. Under the unsupervised setting, it is natural to obtain the embeddings via auto-encoders. Therefore, we should minimize the following reconstruction loss $\mathcal{L}_s$:

$$\mathcal{L}_s = \min_{\mathbf{\Theta}_{AE_v}} \left\| \mathbf{X}^{(v)} - Dec_v\left(Enc_v\left(\mathbf{X}^{(v)}\right)\right) \right\|_F^2, \tag{6}$$

where Enc_v and Dec_v denote the encoder and decoder networks respectively, and $\mathbf{\Theta}_{AE_v}$ is the parameters for each view-specific auto-encoder. The introduced AE networks can extract unique intrinsic information and encode it into a low-dimensional representation instead of directly processing the original high-dimensional data. We concatenate the reconstructed representation for each view to obtain the view-specific representation $\mathbf{S}$.

Multi-view Information-Bottleneck Representation

In order to obtain a compact and comprehensive multi-view information-bottleneck representation $\mathbf{Z}$, we extend the information bottleneck principle to the unsupervised multi-view setting. By integrating the advantages of view-specific representation and shared representation, the proposed model can effectively encode multiple views into a complete representation, thereby improving the generalization capability and adaptively balancing the complementarity and consistency among different views.

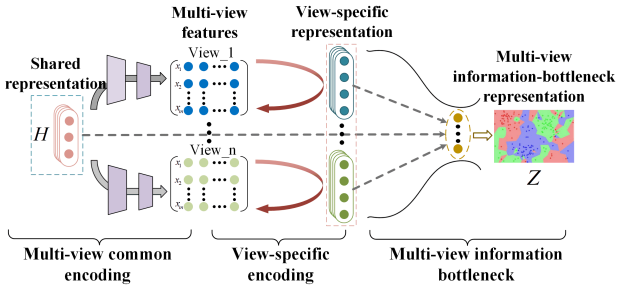


Figure 1: Illustration of the Collaborative Multi-View Information Bottleneck Networks for multi-view representation learning. The proposed CMIB-Nets collaboratively considers guidance of multi-view coding and view-specific coding, which can integrally explore the intra-view intrinsic information and inter-view latent correlations through "bottleneck" to filter worthless information and learn an intact multi-view fusion representation, thus improving generalization ability.

Specifically, in this work, we encourage to maximize the mutual information between multi-view information-bottleneck representation and shared representation $I(\mathbf{Z}, \mathbf{H})$ to explore the underlying structure shared by different views. Simultaneously, we also encourage to learn the intrinsic information of each single view by maximizing the mutual information between multi-view information-bottleneck representation and view-specific representation $I(\mathbf{Z}, \mathbf{S})$. Meanwhile, the mutual information between original views and multi-view information-bottleneck representation $I(\mathbf{Z}, \mathbf{X})$ is minimized to reduce the superfluous and unnecessary information by compressing the description of the data. Accordingly, the objective is induced as

$$\max_{\mathbf{Z}} \; I(\mathbf{Z}, \mathbf{H}) + I(\mathbf{Z}, \mathbf{S}) - \sum_{v=1}^{V} \beta_v I\left(\mathbf{Z}, \mathbf{X}^{(v)}\right), \tag{7}$$

where the task is to maximize $I(\mathbf{Z}, \mathbf{H})$ and $I(\mathbf{Z}, \mathbf{S})$, while, as in *rate-distortion theory* (Davisson 1972), simultaneously compress the description of data.

The main challenge of optimizing the above objective function Eq. (7) is that the mutual information is computationally intractable. Recently, some variational methods (Fabius and van Amersfoort 2014; Alemi et al. 2016) have been used to deal with the problem. The variational approaches can optimize the variational lower bounds of the objective function to find an approximate solution to the original objective function.

For the first term $I(\mathbf{Z}, \mathbf{H})$, according to the definition of mutual information, we have

$$\begin{aligned} I(\mathbf{Z}, \mathbf{H}) &= \int d\mathbf{h} d\mathbf{z} p(\mathbf{h}, \mathbf{z}) \, log \frac{p(\mathbf{h}, \mathbf{z})}{p(\mathbf{h}) \, p(\mathbf{z})} \\ &= \int d\mathbf{h} d\mathbf{z} p(\mathbf{h}, \mathbf{z}) \, log \frac{p(\mathbf{h}|\mathbf{z})}{p(\mathbf{h})}. \end{aligned} \tag{8}$$

Since it is intractable in our case, let $q(\mathbf{h}|\mathbf{z})$ be a variational approximation to $p(\mathbf{h}|\mathbf{z})$. Using the fact that the Kull-

back Leibler divergence is always positive, we have

$$\begin{aligned}&\mathbf{KL}\left[p\left(\mathbf{h}|\mathbf{z}\right),q\left(\mathbf{h}|\mathbf{z}\right)\right]\geq 0\Rightarrow\int d\mathbf{h}p\left(\mathbf{h}|\mathbf{z}\right)log\frac{p\left(\mathbf{h}|\mathbf{z}\right)}{q\left(\mathbf{h}|\mathbf{z}\right)}\geq 0\\&\Rightarrow\int d\mathbf{h}p\left(\mathbf{h}|\mathbf{z}\right)logp\left(\mathbf{h}|\mathbf{z}\right)\geq\int d\mathbf{h}p\left(\mathbf{h}|\mathbf{z}\right)logq\left(\mathbf{h}|\mathbf{z}\right),\end{aligned}\tag{9}$$

and hence we have,

$$\begin{aligned}I\left(\mathbf{Z},\mathbf{H}\right)&\geq\int d\mathbf{h}d\mathbf{z}p\left(\mathbf{h},\mathbf{z}\right)log\frac{q\left(\mathbf{h}|\mathbf{z}\right)}{p\left(\mathbf{h}\right)}\\&=\int d\mathbf{h}d\mathbf{z}p\left(\mathbf{h},\mathbf{z}\right)\log q\left(\mathbf{h}|\mathbf{z}\right)+H\left(\mathbf{h}\right)\\&\geq\int d\mathbf{h}d\mathbf{z}p\left(\mathbf{h},\mathbf{z}\right)\log q\left(\mathbf{h}|\mathbf{z}\right)\\&=\int d\mathbf{h}p\left(\mathbf{h}\right)\int d\mathbf{z}p\left(\mathbf{z}|\mathbf{h}\right)\log q\left(\mathbf{h}|\mathbf{z}\right).\end{aligned}\tag{10}$$

Similarly, for the second term $I\left(\mathbf{Z},\mathbf{S}\right)$, let $t\left(\mathbf{s}|\mathbf{z}\right)$ be a variational approximation, we have

$$I\left(\mathbf{Z},\mathbf{S}\right)\geq\int d\mathbf{s}p\left(\mathbf{s}\right)\int d\mathbf{z}p\left(\mathbf{z}|\mathbf{s}\right)\log t\left(\mathbf{s}|\mathbf{z}\right).\tag{11}$$

Then for the third term, we have

$$I\left(\mathbf{Z},\mathbf{X}^{(v)}\right)=\int d\mathbf{z}d\mathbf{x}p\left(\mathbf{x}^{(v)},\mathbf{z}\right)log\frac{p\left(\mathbf{z}|\mathbf{x}^{(v)}\right)}{p\left(\mathbf{z}\right)}.\tag{12}$$

Basically, calculating the marginal distribution of $p\left(\mathbf{z}\right)$ might be difficult. So let $r\left(\mathbf{z}\right)$ be a variational approximation to this marginal. Since $\mathbf{KL}\left[p\left(\mathbf{z}\right),r\left(\mathbf{z}\right)\right]\geq 0\Rightarrow\int d\mathbf{z}p\left(\mathbf{z}\right)logp\left(\mathbf{z}\right)\geq\int d\mathbf{z}p\left(\mathbf{z}\right)logr\left(\mathbf{z}\right)$, we have the following upper bound:

$$\begin{aligned}I\left(\mathbf{Z},\mathbf{X}^{(v)}\right)&\leq\int d\mathbf{x}d\mathbf{z}p\left(\mathbf{x}^{(v)}\right)p\left(\mathbf{z}|\mathbf{x}^{(v)}\right)log\frac{p\left(\mathbf{z}|\mathbf{x}^{(v)}\right)}{r\left(\mathbf{z}\right)}\\&=\int d\mathbf{x}p\left(\mathbf{x}^{(v)}\right)\int d\mathbf{z}p\left(\mathbf{z}|\mathbf{x}^{(v)}\right)log\frac{p\left(\mathbf{z}|\mathbf{x}^{(v)}\right)}{r\left(\mathbf{z}\right)}.\end{aligned}\tag{13}$$

Combining the above inequalities, we have

$$\begin{aligned}&I\left(\mathbf{Z},\mathbf{H}\right)+I\left(\mathbf{Z},\mathbf{S}\right)-\sum_{v=1}^{V}\beta_{v}I\left(\mathbf{Z},\mathbf{X}^{(v)}\right)\\&\geq\int d\mathbf{h}p\left(\mathbf{h}\right)\int d\mathbf{z}p\left(\mathbf{z}|\mathbf{h}\right)\log q\left(\mathbf{h}|\mathbf{z}\right)\\&+\int d\mathbf{s}p\left(\mathbf{s}\right)\int d\mathbf{z}p\left(\mathbf{z}|\mathbf{s}\right)\log t\left(\mathbf{s}|\mathbf{z}\right)\\&-\sum_{v=1}^{V}\beta_{v}\int d\mathbf{x}p\left(\mathbf{x}^{(v)}\right)\int d\mathbf{z}p\left(\mathbf{z}|\mathbf{x}^{(v)}\right)log\frac{p\left(\mathbf{z}|\mathbf{x}^{(v)}\right)}{r\left(\mathbf{z}\right)}.\end{aligned}\tag{14}$$

In practice, the integral over $\mathbf{h}$, $\mathbf{s}$ and $\mathbf{x}^{(v)}$ can be approximated by Monte Carlo sampling (Shapiro 2003). Therefore, we have

$$\begin{aligned}&I\left(\mathbf{Z},\mathbf{H}\right)+I\left(\mathbf{Z},\mathbf{S}\right)-\sum_{v=1}^{V}\beta_{v}I\left(\mathbf{Z},\mathbf{X}^{(v)}\right)\\&\approx\frac{1}{N}\sum_{i=1}^{N}\{\int d\mathbf{z}p\left(\mathbf{z}|\mathbf{h}_{i}\right)\log q\left(\mathbf{h}_{i}|\mathbf{z}\right)\\&\qquad+\int d\mathbf{z}p\left(\mathbf{z}|\mathbf{s}_{i}\right)\log t\left(\mathbf{s}_{i}|\mathbf{z}\right)\\&\qquad-\sum_{v=1}^{V}\beta_{v}\int d\mathbf{z}p(\mathbf{z}|\mathbf{x}_{i}^{(v)})log\frac{p(\mathbf{z}|\mathbf{x}_{i}^{(v)})}{r\left(\mathbf{z}\right)}\},\end{aligned}\tag{15}$$

where N is the size of total sampled data.

Then we employ the reparameterization trick (Kingma and Welling 2013; Alemi et al. 2016) to rewrite $p\left(\mathbf{z}|\mathbf{h}\right)d\mathbf{z}=p\left(\varepsilon_{1}\right)d\varepsilon_{1}$ and $p\left(\mathbf{z}|\mathbf{s}\right)d\mathbf{z}=p\left(\varepsilon_{2}\right)d\varepsilon_{2}$, where $\mathbf{z}=g_{1}\left(\mathbf{h},\varepsilon_{1}\right)$ and $\mathbf{z}=g_{2}\left(\mathbf{s},\varepsilon_{2}\right)$ with the Gaussian random variable ε_{1} and ε_{2}. Accordingly, we can obtain the following objective function $\mathcal{L}_{r}\left(\cdot\right)$ for learning the multi-view information-bottleneck representation, which is to be minimized:

$$\begin{aligned}\mathcal{L}_{r}=\frac{1}{N}\sum_{i=1}^{N}\{&-\mathbb{E}_{\varepsilon_{1}}log\,q\left(\mathbf{h}_{i}|g_{1}\left(\mathbf{h}_{i},\varepsilon_{1}\right)\right)\\&-\mathbb{E}_{\varepsilon_{2}}log\,t\left(\,\mathbf{s}_{i}\,|g_{2}\left(\mathbf{s}_{i},\varepsilon_{2}\right)\right)\\&+\sum_{v=1}^{V}\beta_{v}D_{KL}[p(\mathbf{z}|\mathbf{x}_{i}^{(v)}),r(\mathbf{z})]\},\end{aligned}\tag{16}$$

where $\beta_{v}>0$ are trade-off factors. The first and the second terms are reconstruction loss, and the third term is the KL divergence. In this way, the irrelevant and superfluous information in the original view can be reduced, and we can obtain a compact and comprehensive *multi-view information-bottleneck representation* which encodes view-specific intrinsic information and the across-view latent relationships.

Experiments

In the experiments, we compare our proposed CMIB-Nets with existing state-of-the-art multi-view representation learning algorithms on real-world multi-view datasets. Specifically, we evaluate the performances on clustering in terms of common metrics, and verify the generalization and robustness of the model through a variety of experiments.

Datasets

We evaluate our model on six multi-view benchmark datasets in the experiments, including:

1) Handwritten[1] contains 2000 examples of numbers 0-9 with 200 samples per class. These images are represented with two different types of features. **2) ORL**[2] face dataset includes 40 different categories, and each category has 10 different facial images. The intensity and Gabor are used as different views. **3) COIL-20**[3] consists of 1440 pictures

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

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

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

Datasets	Metrics	FeatConcate	CCA	DCCA	DCCAE	DMF	MIB	AE^2-Nets	Ours
Handwritten	ACC	76.08±2.14	66.42±4.81	66.16±1.16	69.29±1.02	71.85±3.55	81.52±2.37	85.62±1.82	**89.72±1.06**
	NMI	75.74±1.44	69.66±4.06	66.04±0.49	66.95±0.91	73.11±2.23	76.70±0.82	76.39±1.50	**81.78±1.17**
	F-score	70.95±2.05	62.06±4.77	59.09±0.38	60.50±1.30	66.67±2.97	72.93±2.23	74.58±1.85	**81.64±0.23**
	RI	93.92±0.42	91.87±1.34	91.36±0.08	91.76±0.22	92.86±1.01	94.11±0.47	95.69±0.38	**96.25±0.13**
ORL	ACC	61.11±1.50	56.96±2.04	59.64±2.20	59.42±2.06	65.36±2.88	66.69±2.01	68.85±2.11	**72.07±1.81**
	NMI	79.29±0.73	76.01±0.79	77.82±0.86	77.54±0.83	82.86±1.21	83.45±1.26	85.74±0.78	**88.23±0.48**
	F-score	47.02±2.11	45.10±1.87	47.71±2,05	46.69±2.27	52.03±3.34	56.50±2.59	59.93±1.31	**68.32±1.05**
	RI	97.12±0.26	97.29±0.10	97.40±0.14	97.37±0.13	97.32±0.22	97.58±0.28	97.94±0.11	**98.39±0.10**
COIL20	ACC	67.15±3.79	58.64±1.39	63.71±1.08	62.72±1.41	53.93±5.06	74.25±2.56	73.42±1.90	**77.58±1.17**
	NMI	79.96±1.63	70.60±0.75	75.99±1.15	76.32±0.66	72.35±2.33	82.43±1.73	82.55±1.01	**84.61±0.79**
	F-score	64.85±3.67	53.09±1.40	58.74±0.57	57.56±1.12	46.39±4.39	71.45±2.56	69.38±1.98	**74.95±1.26**
	RI	96.29±0.40	95.15±0.22	95.57±0.10	95.27±0.32	92.57±1.28	96.86±0.22	96.19±0.14	**97.16±0.17**
Caltech101	ACC	47.22±0.22	45.35±0.13	56.50±3.05	62.17±2.78	55.67±2.67	47.65±0.67	66.45±2.55	**71.02±0.98**
	NMI	57.12±0.62	50.52±0.13	57.64±3.75	**64.38±4.12**	45.56±2.18	55.43±0.56	60.93±1.73	62.47±1.86
	F-score	52.28±0.28	53.51±0.19	62.32±5.07	65.24±2.17	57.70±2.25	53.46±0.56	73.32±2.73	**75.92±1.44**
	RI	73.47±0.12	73.25±0.16	76.31±2.46	79.36±1.78	73.43±2.73	73.30±0.40	83.13±2.17	**85.48±1.15**
BBCSport	ACC	41.94±3.26	39.51±2.36	69.39±1.80	72.98±3.13	41.31±3.09	51.29±3.65	62.28±3.25	**77.13±1.63**
	NMI	15.94±2.19	12.45±1.88	50.36±1.83	54.55±4.03	15.57±1.57	35.11±2.44	54.42±3.49	**71.47±3.12**
	F-score	41.53±2.23	40.42±1.88	61.20±2.49	66.46±1.89	41.16±2.19	46.56±1.81	58.39±3.34	**73.45±2.78**
	RI	37.45±1.13	34.08±3.26	80.61±0.39	83.35±1.51	36.49±2.50	49.37±2.27	69.76±2.85	**84.35±1.98**
CUB	ACC	73.81±0.10	45.85±1.46	54.49±0.29	66.72±1.52	37.55±2.61	79.27±3.28	77.75±1.63	**80.37±2.06**
	NMI	71.48±0.41	46.60±0.58	52.51±1.09	65.77±1.36	37.84±2.03	77.34±0.75	**78.61±1.62**	77.40±1.02
	F-score	61.08±0.17	39.90±1.28	45.85±0.31	58.21±1.12	28.96±1.61	72.65±1.73	70.96±2.03	**73.89±1.29**
	RI	91.98±0.05	87.41±0.46	88.63±0.09	91.24±0.25	85.56±0.30	93.95±0.43	93.39±0.63	**94.32±0.24**

Table 1: Performance comparison on clustering task.

of 20 categories. In the experiments, the intensity and Gabor features are extracted as two different perspectives. **4) Caltech101-7**[4] is a subset of the original Caltech101 image dataset. This subset selected 1,474 images in seven views. And HOG and GIST are used as two types of features. **5) BBCSport**[5] is a collection of 544 documents associated with two views of sports articles from 5 subject areas. **6) Caltech-UCSD Birds (CUB)**[6] has 200 different categories, including 11788 images of birds with the corresponding textual descriptions (Reed et al. 2016). The image features are extracted by GoogLeNet, and the text features are extracted by Doc2Vec. The two kinds of features are used as different views.

Compared Methods

We compare our approach CMIB-Nets with the following multi-view algorithms:

- **FeatConcate**: This method directly concatenates the features of multiple views.
- **CCA** (Canonical Correlation Analysis) (Hotelling 1992): This method maps different types of features onto a projection subspace by maximizing the correlations.
- **DCCA** (Deep Canonical Correlation Analysis) (Andrew et al. 2013): This method extends CCA applying deep neural networks, and then maximizes the correlations among the different views.
- **DCCAE** (Deep Canonically Correlated AutoEncoders) (Wang et al. 2015): This method uses autoencoder to maximize the correlation of the learned representations, and then assembles the representations together.
- **DMF-MVC** (Deep Semi-NMF for MVC) (Zhao, Ding, and Fu 2017): This method takes advantage of the seminonnegative matrix factorization to obtain a representation included consistent information of multiple views.
- **AE^2-Nets** (Autoencoder in Autoencoder Networks) (Zhang, Liu, and Fu 2019): This method integrates information from multiple views into a representation by nested autoencoder.
- **MIB** (Multi-view Information Bottleneck) (Federici et al. 2020): This method identifies redundant information as that which is not shared by both views, and discards the superfluous information through information bottleneck.

Performance Evaluation

We evaluate the proposed CMIB-Nets on clustering task. Specifically, we conduct the k-means algorithm by using the learned representations from different algorithms. The reason for applying k-means is that the algorithm is relatively simple and can reflect the quality of the learned representations based on Euclidean distance.

In the experiments, we 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 per-

[4]http://www.vision.caltech.edu/Image Datasets/Caltech101/
[5]http://mlg.ucd.ie/datasets/
[6]http://www.vision.caltech.edu/visipedia/CUB-200

Datasets	Noise (μ)	CCA	DCCA	DCCAE	MIB	AE^2-Nets	Ours
MNIST	0	37.55±0.03	43.30±2.61	49.43±0.71	54.92±4.01	57.44±2.12	**60.92±2.36**
	0.1	36.35±0.01	40.59±0.07	40.84±0.38	51.71±2.94	52.50±1.49	**57.33±1.84**
	0.2	33.28±0.02	22.62±0.17	10.46±0.07	48.62±1.49	51.08±0.87	**55.83±1.06**
Fashion-MNIST	0	47.46±0.53	46.48±2.99	47.28±0.87	57.60±1.21	53.19±0.58	**62.82±1.17**
	0.1	46.58±0.94	45.76±1.52	43.96±1.86	56.82±1.40	51.79±0.36	**61.56±1.25**
	0.2	44.29±1.18	44.86±2.01	40.43±2.27	53.98±1.64	50.78±0.90	**60.10±1.53**

Table 2: Performance comparison on datasets with noise in terms of accuracy.

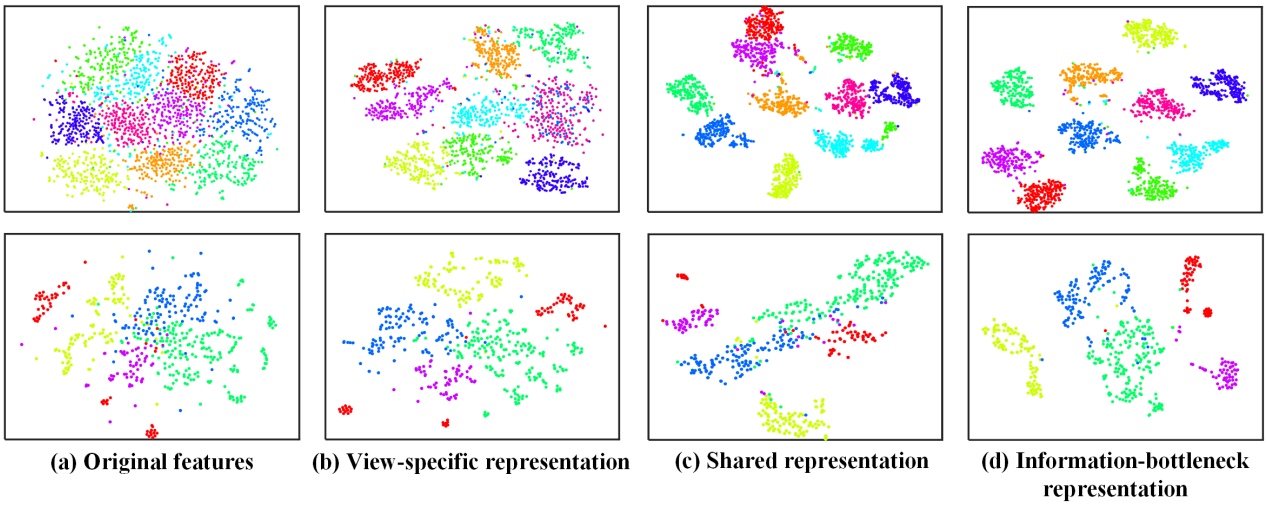


Figure 2: t-SNE visualizations of original features and various representations on Handwritten (top) and BBCSport (bottom).

formance. To reduce the effect of randomness, we run each method 30 times and report the average performances.

Table 1 reports the performances of different multi-view methods on clustering task. On the whole, our algorithm almost outperforms other compared methods on all datasets. Obviously, the performance of CCA is not unsatisfactory than other methods, especially on Caltech101 and CUB datasets, since it only focuses on linear correlations, which is difficult to deal with complex relationships. Instead, DCCA and DCCAE perform better, where the deep neural networks are effective to handle complex correlations. However, they all map different views onto a projection subspace by maximizing the correlations but cannot learn the complementary information of multiple views, which may degrade the quality of learned representation. Moreover, compared with the recent methods of AE^2-Nets and MIB, our method still obtains clear improvements. Our method captures intra-view intrinsic information and inter-view common latent structure, thus effectively improves the generalization ability of our model and shows obvious superiority.

Robustness Assessment

We verify the robustness of the algorithms by adding the noise to the datasets. Specifically, we conduct clustering experiments on two datasets, i.e., MNIST[1] and Fashion-MNIST[2]. For both datasets, each image is equally split into left and right parts as two views. In our experiment, a generated noise matrix is produced by randomly sampling from the range $[0, 1]$. Then, we multiply the noise matrix with a scalar $0 < \mu < 1$ to adjust the noise level.

According to the clustering results in Table **??**, our algorithm performs more robustly by increasing the level of noise. It can be clearly observed that as the noise level increases, the performances of all algorithms decrease, however, the performance of our algorithm is always the top performer. These results indicate that when the datasets contain noise, our algorithm can suppress the noise and ensure promising clustering performance. Although the noise is involved, the intrinsic information and the latent structure across different views are still encoded into the multi-view information-bottleneck representation, resulting in robustness and superior generalization performance.

Ablation Study

In this part, we conduct the ablation study to further demonstrate the effectiveness of the proposed CMIB-Nets.

Comparison of performances. In order to investigate the effectiveness of different components of the proposed approach, we conduct a series of clustering experiments

[1]http://yann.lecun.com/exdb/mnist

[2]https://github.com/zalandoresearch/fashion-mnist

Datasets	Methods	ACC	NMI
Handwritten	View1	73.01±1.62	70.73±2.15
	View2	70.80±1.81	65.86±3.13
	S	76.24±1.81	71.39±1.50
	H	81.85±1.57	75.94±1.36
	Z	**89.72±1.06**	**81.78±1.17**
BBCSport	View1	42.31±2.55	18.24±3.39
	View2	46.56±2.82	23.99±4.10
	S	52.92±3.65	40.24±3.19
	H	60.34±2.57	52.39±3.73
	Z	**77.13±1.63**	**71.47±3.12**
CUB	View1	64.25±2.11	64.83±1.32
	View2	61.10±1.82	58.57±3.15
	S	74.23±0.81	72.08±1.20
	H	72.75±1.36	70.69±1.58
	Z	**80.37±2.06**	**77.40±1.02**

[1] **S** represents view-specific representation, **H** denotes the shared representation, and **Z** means the multi-view information-bottleneck representation.

Table 3: Ablation study of the original views and different representations with clustering task.

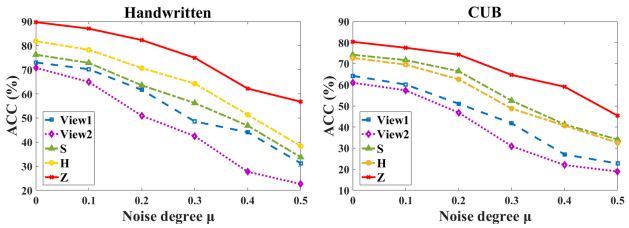
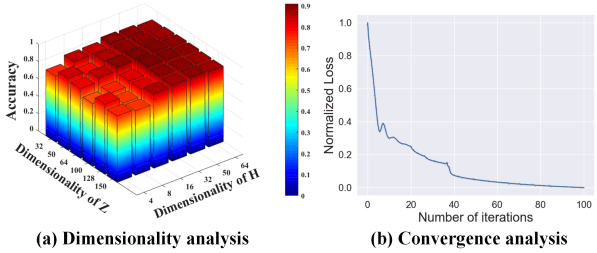


Figure 3: Comparison of the robustness of the original views and various representations on the noisy datasets.

by using the original features, shared representation **H**, view-specific representation **S**, and multi-view information-bottleneck representation **Z**, respectively. The results of the ablation study are shown in Table 3. It can be observed that the multi-view information-bottleneck representation **Z** learned by our model achieves the best results.

Comparison of visualizations. Furthermore, we also visualize the original features and different multi-view representations with t-SNE (Maaten and Hinton 2008) on Handwritten and BBCSport datasets in Fig. 2, to intuitively illustrate the advantages of our model. Obviously, it can be seen that the learned representation can better express the structural relationship of the original data. In addition, the multi-view information-bottleneck representation **Z** can jointly learn the intra-view intrinsic information and the inter-view latent structure, thereby making the clustering structure more clear.

Comparison of robustness. Moreover, the multi-view information-bottleneck representation **Z** is not only more discriminative, but also more robust than other features or representations. Specifically, we perform clustering ablation experiments on the noisy datasets, and the results are shown in Fig. 3. It can be observed that as the level of noise



(a) Dimensionality analysis (b) Convergence analysis

Figure 4: Model analysis: (a) Dimensionality analysis (i.e., Z and H); (b) Convergence analysis.

increases, the performances of all representations will decrease, but the multi-view information-bottleneck representation **Z** is always promising and is more stable.

Model Analysis

In this work, we select the dimensionality of shared representation **H** from $\{4, 8, 16, 32, 50, 64\}$ and the dimensionality of multi-view information-bottleneck representation **Z** from $\{32, 50, 64, 100, 128, 150\}$. As shown in Fig. 4(a), we show the performances of our proposed CMIB-Nets with different dimensions of the representations on Handwritten dataset. Moreover, if the dimensionality decreases too much, the representation may not have enough capacity to encode information from all views. Too larger dimensionality also produces lower performance, where high-dimensional representation tends to overfit and may contain possible noise.

To demonstrate the convergence of the proposed method, we conduct the convergence analysis as shown in Fig. 4(b). The optimization process of our model is relatively stable, where the loss decreases rapidly and converges within a number of iterations on Handwritten dataset in practice.

Conclusion

In this paper, we propose a novel multi-view representation learning method, which relies on multiple views to produce a reliable representation for downstream tasks. The proposed model explores the latent relationships and intrinsic information among different views, and exploits the information bottleneck principle to discard the useless information from the multi-view data. The resulting multi-view information-bottleneck representation can retain intrinsic information, and adaptively balance the consistency and complementarity among multiple views. The experimental results indicate that the proposed CMIB-Nets achieves superior performances on various real-world datasets for different tasks compared to state-of-the-art methods, showing better generalization performance and robustness. In future work, we will extend the current method to an end-to-end model.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (No. 61976151, 61732011 and 61876127), the Natural Science Foundation of Tianjin of China (No. 19JCYBJC15200).

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410* .

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, 1247–1255.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41(2): 423–443.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, 886–893. IEEE.

Davisson, L. D. 1972. Rate-distortion theory and application. *Proceedings of the IEEE* 60(7): 800–808.

Dhillon, P.; Foster, D. P.; and Ungar, L. H. 2011. Multi-view learning of word embeddings via cca. In *Advances in neural information processing systems*, 199–207.

Fabius, O.; and van Amersfoort, J. R. 2014. Variational recurrent auto-encoders. *arXiv preprint arXiv:1412.6581* .

Federici, M.; Dutta, A.; Forré, P.; Kushman, N.; and Akata, Z. 2020. Learning Robust Representations via Multi-View Information Bottleneck. *arXiv preprint arXiv:2002.07017* .

Guo, Y. 2013. Convex subspace representation learning from multi-view data. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Hotelling, H. 1992. Relations between two sets of variates. In *Breakthroughs in statistics*, 162–190. Springer.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Lee, T. S. 1996. Image representation using 2D Gabor wavelets. *IEEE Transactions on pattern analysis and machine intelligence* 18(10): 959–971.

Li, Y.; Yang, M.; and Zhang, Z. 2018. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering* 31(10): 1863–1883.

Liu, X.; Zhu, X.; Li, M.; Wang, L.; Tang, C.; Yin, J.; Shen, D.; Wang, H.; and Gao, W. 2018. Late fusion incomplete multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence* 41(10): 2410–2423.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60(2): 91–110.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689–696.

Novak, C. L.; Shafer, S. A.; et al. 1992. Anatomy of a color histogram. In *CVPR*, volume 92, 599–605.

Oliva, A.; and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision* 42(3): 145–175.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 49–58.

Shapiro, A. 2003. Monte Carlo sampling methods. *Handbooks in operations research and management science* 10: 353–425.

Sun, S. 2013. A survey of multi-view machine learning. *Neural computing and applications* 23(7-8): 2031–2038.

Tao, Z.; Liu, H.; Li, S.; Ding, Z.; and Fu, Y. 2019. Marginalized multiview ensemble clustering. *IEEE transactions on neural networks and learning systems* 31(2): 600–611.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057* .

Wang, Q.; Boudreau, C.; Luo, Q.; Tan, P.-N.; and Zhou, J. 2019. Deep multi-view information bottleneck. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, 37–45. SIAM.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On deep multi-view representation learning. In *International Conference on Machine Learning*, 1083–1092.

White, M.; Zhang, X.; Schuurmans, D.; and Yu, Y.-l. 2012. Convex multi-view subspace learning. In *Advances in neural information processing systems*, 1673–1681.

Xu, C.; Tao, D.; and Xu, C. 2014. Large-margin multi-viewinformation bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(8): 1559–1572.

Zhang, C.; Cui, Y.; Han, Z.; Zhou, J. T.; Fu, H.; and Hu, Q. 2020. Deep Partial Multi-View Learning. *IEEE transactions on pattern analysis and machine intelligence* .

Zhang, C.; Fu, H.; Hu, Q.; Cao, X.; Xie, Y.; Tao, D.; and Xu, D. 2018. Generalized latent multi-view subspace clustering. *IEEE transactions on pattern analysis and machine intelligence* .

Zhang, C.; Hu, Q.; Fu, H.; Zhu, P.; and Cao, X. 2017. Latent multi-view subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4279–4287.

Zhang, C.; Liu, Y.; and Fu, H. 2019. AE2-Nets: Autoencoder in Autoencoder Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2577–2585.

Zhao, H.; Ding, Z.; and Fu, Y. 2017. Multi-view clustering via deep matrix factorization. In *Thirty-First AAAI Conference on Artificial Intelligence*.