Computer Fraud and Security
ISSN (online): 1873-7056

Specific-information Extraction and Augmentation for Semi-supervised Multi-view Representation Learning

Wanqing Xu[1], Xiaoyuan Jing[1]*, Wei Liu[1]
[1] *School of Computer Science, Wuhan University, Wuhan, China*
Corresponding Author's Email: jingxy_2000@126.com

Abstract: In practical applications, learning accurate representations from multi-view data is a critical step. The approaches of shared-and-specific framework have consistently been a focal point in the field of multi-view classification, as they leverage both shared and complementary information through these representations. However, existing authoritative methods lack precision in extracting information from multi-view data, resulting in a significant amount of interfering redundant information. Furthermore, research on data augmentation at the level of specific information has not been fully developed. To address this issue, a novel semi-supervised multi-view learning method, SEMA (Specific-InforMation Extraction and Augmentation), is proposed. SEMA achieves more accurate specific information by incorporating orthogonal constraints and designs a data augmentation strategy tailored for specific information. This strategy provides a large number of auxiliary samples for semi-supervised multi-view learning while preventing the consistency of shared information from being repeatedly augmented. The experimental results on seven benchmark datasets demonstrate the effectiveness of SEMA.

Keywords: multi-view feature learning, specific information, data augmentation, semi-supervised learning

1. Introduction

Multi-view data, a significant form of modern data representation, seeks to capture and document the same object from various dimensions or perspectives, including but not limited to color, texture, and shape. This approach provides a richer and more comprehensive feature description of the object [1,2]. Compared to single-view data, multi-view data offers a deeper insight into the object's intrinsic properties and external characteristics, thereby facilitating a more precise and detailed analysis. The labeling of multi-view data is prohibitively costly, whereas unlabeled data is readily accessible. Consequently, semi-supervised learning [3,4] is an effective strategy. Semi-supervised multi-view learning utilize a limited amount of labeled data combined with a large volume of unlabeled data for model training, aiming to enhance learning performance by extracting latent information from the unlabeled data. This learning method has shown considerable promise across various domains and applications, such as image recognition and natural language processing [5,6].

Unlike single-view data, multi-view data is characterized by three fundamental attributes. Firstly, it encompasses shared and specific information [7]. Shared information refers to the correlated data that is common across multiple views, whereas specific information denotes the unique data contained within each individual view [8]. Furthermore, multi-view data incorporates a significant amount of redundant information. The primary challenge lies in effectively integrating these diverse representations to maximize the utility of both shared and specific information, while simultaneously mitigating the detrimental effects of redundancy [9].

In recent years, significant advancements have been made in the field of multi-view representation learning, with the emergence of numerous semi-supervised and supervised learning methods. These methodologies can be broadly categorized into two classes: joint methods and alignment methods [10]. Joint methods integrate multiple views into a single unified feature vector, with common models including graph-based models [11, 12] and neural network-based models [13]. On the other hand, alignment methods aim to project different views into a shared subspace to maximize the correlation between views, thereby ensuring feature consistency and facilitating their effective application in learning tasks [14,15]. Typical representatives of alignment methods include Canonical Correlation Analysis (CCA) [16] and its variants, such as Kernel CCA (KCCA) [17], Multi-view CCA (MCCA) [18], and DeepCCA [19].

However, existing methods exhibit certain limitations when handling multi-view data. While joint methods achieve representation fusion, they neglect the consensus attributes and information exchange between views, leading to a significant presence of redundant information [20]. Alignment methods capture relevant information between views but fail to fully exploit the unique internal information and rich complementary details of each view [21]. To address these limitations, the shared-and-specific approach has been developed, which effectively leverages the consensus and complementary characteristics of multi-view data by partitioning the information of each view into shared information and view-specific information [22]. For instance, Zhou et al. [23] developed a method to learn a shared dictionary alongside multiple view-specific dictionaries in the latent space, thereby fully exploiting consensus and complementary information. Hu et al. [24] introduced MvDML (sharable and individual multi-view metric learning), a framework that leverages individual features and shared features across all views through multi-view specific networks and a common network, respectively. Xu et al. [25] utilized deep neural networks to extract interactive information between pairwise views. Jia et al.

[26] proposed MDDRL, a multi-view deep discriminant representation learning method, which, by incorporating orthogonality and adversarial similarity constraints, simultaneously addresses the three characteristics of multi-view data, demonstrating that reducing redundancy in representation learning can effectively enhance learning performance.

While the previously discussed shared-and-specific frameworks have marginally improved the performance of multi-view representation learning, they remain deficient in accurately extracting specific information. View-specific information refers to the unique attributes not possessed by other views. However, existing methods fail to adequately eliminate redundancy, resulting in insufficiently accurate extraction of specific information. To address this issue, we propose an innovative solution: enhancing orthogonality constraints to extract more accurate specific information and reduce redundancy. Orthogonal constraints force the specific information of different views to be mutually independent, thereby ensuring that the extracted specific information truly reflects the unique attributes of each view. Furthermore, orthogonal constraints also aid in eliminating redundant information, thereby improving the compactness and efficiency of the representation. Through theoretical analysis and experimental validation, our paper demonstrates the effectiveness and superiority of enhancing orthogonal constraints in improving the performance of multi-view representation learning.

In the field of multi-view classification, semi-supervised learning techniques enhance model performance by generating additional training data from a limited set of labeled data through data augmentation methods. For example, Mixup [27] generates new samples by combining two existing samples in a proportional manner, MixMatch [28] extends label information by predicting low-entropy labels for unlabeled samples, and GVCA [29] utilizes both labeled and unlabeled data to expand feature distributions, thereby enhancing the diversity of feature representations. These approaches have effectively improved the accuracy and robustness of multi-view classification to a certain extent.

However, existing data augmentation methods suffer from significant limitations in their design and application: they are not tailored to the shared-and-specific framework and typically enhance the entire representation of all views. For instance, new representations are generated by linearly combining the representations with the sum of combination weights equal to 1, thereby achieving overall enhancement for all views. This approach overlooks the fundamental differences between shared and specific information in multi-view data, leading to repeated enhancement of shared information and consequently generating substantial redundancy. To address this issue, our paper proposes an innovative data augmentation method specifically designed for the shared-and-specific framework. When generating training samples using unlabeled data, this method only specifically enhances the specific information while keeping the shared information unchanged. This targeted augmentation strategy ensures that the unique attributes of each view are fully exploited and utilized, while avoiding redundant repetition of shared information.

Based on the discussion, two critical aspects for semi-supervised multi-view learning based on the shared-and-specific representation framework are the enhancement of the extraction method for specific information and the design of data augmentation strategies based on specific information. We introduce a novel semi-supervised multi-view learning approach SEMA. The contributions of this paper are summarized as follows:

Firstly, we introduce a novel semi-supervised multi-view learning method. This method achieves the accurate extraction of specific information by imposing orthogonality constraints on the shared and specific representations within each view and between the specific representations of different views. Secondary, a data augmentation method specifically designed for the shared-and-specific framework is proposed, offering both supervised and unsupervised information augmentation. This method fully leverages labeled and unlabeled data to generate a substantial number of reliable auxiliary samples through targeted enhancement of specific information, thereby enhancing the model's accuracy and robustness. We conducted extensive experiments on seven public datasets. The experimental results validate the effectiveness and rationality of our approach. Further ablation studies confirm the feasibility and superiority of the method in specific information extraction and data augmentation.

2. Multi-view deep partition representation learning

MDPRL (Multi-View Deep Partition Representation Learning) is an innovative method for multi-view representation learning, whose performance improvement is attributed to the strategic application of orthogonality and adversarial similarity constraints. The overall architecture is illustrated in Fig. 1. Consequently, the loss function of the proposed approach can be partitioned into four components: a classification loss L_c for categorization, the constraint loss L_{spec}, which aims to minimize redundancy in the extracted specific information; the constraint loss L_{con} for extracting consensus information, further reinforcing consistency across multiple views; and finally, the semi-supervised loss L_{semi}, which leverages unlabeled data more effectively to boost model performance. This divide-and-conquer strategy enables MDPRL to exhibit distinct advantages in the field of multi-view learning. The combined loss function can be presented as:

$$L = L_c + \alpha L_{spec} + \beta L_{con} + \gamma L_{semi} \tag{1}$$

where α, β and γ are trade-off parameters.

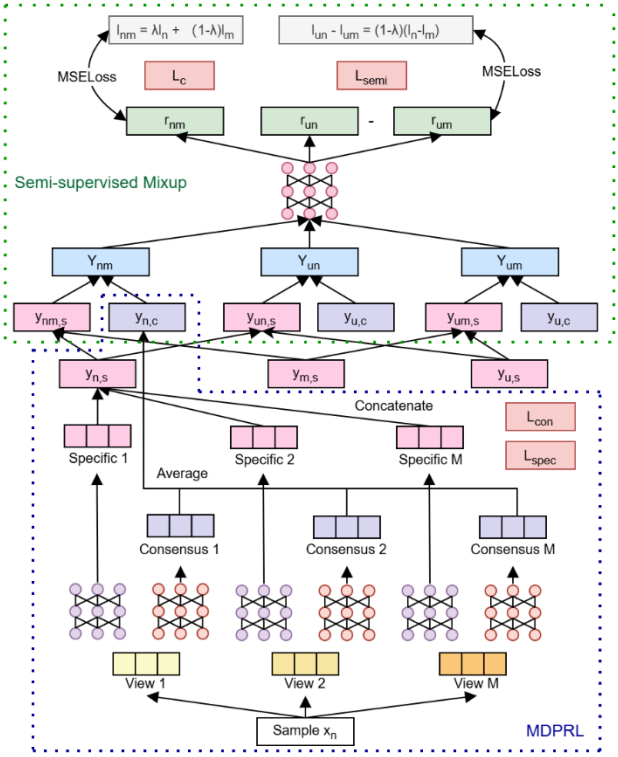


Figure. 1. The architecture of our proposed approach.

2.1 Consensus Information

Suppose the sample set $X = \left\{ x^v \in \mathrm{R}^{d_v} \right\}_{v=1}^{M}$ is collected from M views, where d_v denotes the feature dimensionality of the view v, with each view comprising N samples. The sample set originating from view v can be represented as $x^v = \left\{ x_i^v \in \mathrm{R}^{d_v} \right\}_{i=1}^{N}$, where x_i^v signifies the representation of the i^{th} sample under view v, $i = 1, 2, \cdots N$.

For consensus information, defining a set of neural networks $\left\{ f_c^v \right\}_{v=1}^{M}$ where f_c^v captures the high-level consensus representation for the v^{th} view, the step can be represented by the following:

$$y_{i,c}^v = f_c^v \left(x_i^v \right) \tag{2}$$

where $y_{i,c}^v \in \mathrm{R}_v^d$ and f_c^v is a nerual network that has a total of L layers.

$$h_{f_c^v}^l = \sigma \left(W_{f_c^v}^l h_{f_c^v}^{l-1} + b_{f_c^v}^l \right) \tag{3}$$

For layer l ($l = 1, 2, \cdots, L$), the weight matrix $W_{f_c^v}^l \in \mathrm{R}^{m_l \times m_{l-1}}$, where $m_0 = d_v$, $m_L = d$. The bias vectors $b_{f_c^v}^l \in \mathrm{R}^{m_l}$, $h_{f_c^v}^l \in \mathrm{R}^{m_l}$ are the outputs of the l^{th} layer, and σ. is the activation function.

Building on [25], we employ adversarial training to evaluate the similarity of the acquired representation $y_{i,c}^v$. We conceptualize the representation learning networks $\left\{ f_c^v \right\}_{v=1}^{M}$ as a set of generators, where the parameters are trained in parallel. We then utilize an M-class classifier as the discriminator D to differentiate the distribution of each generated representation, which can be presented as:

$$P_i^v = D \left(G_v \left(x_i^v \right) \right) \tag{4}$$

Upon the completion of training, the comprehensive consensus representations from all views are nearly identical. Consequently, we employ the average $y_{i,c}$ as a substitute, and get all the consensus information $Y_c = \left\{ y_{i,c} \in \mathrm{R}^d \right\}_{i=1}^{N}$, which can be expressed as:

$$y_{i,c} = \frac{1}{M} \sum_{v=1}^{M} y_{i,c}^{v} \tag{5}$$

2.2 Specific Information

For the extraction of specific information, we define a collection of neural networks $\left\{ f_s^v \right\}_{v=1}^{M}$ to capture the view-specific representation, which can be formalized as follows:

$$y_{i,s}^{v} = f_s^v \left(x_i^v \right) \tag{6}$$

where $y_{i,s}^{v} \in \mathrm{R}_v^d$ and f_s^v is a nerual network that has a total of L layers.

$$h_{f_s^v}^{l} = \sigma\left(W_{f_s^v}^{l} h_{f_s^v}^{l-1} + b_{f_s^v}^{l} \right) \tag{7}$$

For layer l ($l = 1, 2, \cdots, L$), the weight matrix $W_{f_s^v}^{l} \in \mathrm{R}^{m_l \times m_{l-1}}$, where $m_0 = d_v$, $m_L = d$. The bias vectors $b_{f_s^v}^{l} \in \mathrm{R}^{m_l}$, $h_{f_s^v}^{l} \in \mathrm{R}^{m_l}$ are the outputs of the l^{th} layer, and σ is the activation function.

To effectively disentangle the shared information from the specific information across different views and significantly reduce information redundancy, we introduce an orthogonal constraint mechanism. Unlike traditional methods that focus solely on ensuring the absence of shared information within view-specific information, neglecting the potential correlations and redundancies among different view-specific information, our approach innovatively incorporates a new orthogonality constraint into the model. This constraint not only enforces the independence between view-specific information and shared information but also ensures the orthogonality among different view-specific information. As a result, a profound separation of view-specific information is successfully achieved.

Let $Y_s^v = \left\{ y_{i,s}^{v} \in \mathrm{R}^d \right\}_{i=1}^{N}$ be the specific information extracted from the v^{th} view, the orthogonality loss of the consensus information and view-specific information can be defined as:

$$L_{spec1} = \sum_{v=1}^{M} \left\| Y_c^{\bullet} \, Y_s^v \right\|_F^2 \tag{8}$$

where $\left\| \cdot \right\|_F^2$ is the squared Frobenius norm and the orthogonality loss among the view-specific information itself can be defined as:

$$L_{spec2} = \sum_{v=1}^{M} \sum_{u=v+1}^{M} \left\| Y_s^{v\bullet} \, Y_s^u \right\|_F^2 \tag{8}$$

Combining these two components, the final loss for this section can be expressed as:

$$. L_{spec} = \sum_{v=1}^{M} \left\| Y_c^{\bullet} \, Y_s^v \right\|_F^2 + \sum_{v=1}^{M} \sum_{u=v+1}^{M} \left\| Y_s^{v\bullet} \, Y_s^u \right\|_F^2 . \tag{9}$$

2.3 Different kinds of information integration

Building upon the foundations established in the preceding two sections, the initial step involves concatenating the features extracted from all views. This concatenation aims to obtain a comprehensive representation of both view-specific information and shared information, as demonstrated in (11) and (12):

$$Y_c = \left[Y_c^{1\bullet}, Y_c^{2\bullet}, \cdots, Y_c^{M\bullet} \right]^{\bullet} \tag{10}$$

$$Y_S = \left[Y_S^{1\bullet}, Y_S^{2\bullet}, \cdots, Y_S^{M\bullet} \right]^{\bullet} \tag{11}$$

Subsequently, the final representation is achieved through concatenation, as illustrated in (13):

$$Y = \left[Y_S^{\bullet}, Y_C^{\bullet} \right]^{\bullet} \tag{12}$$

We designed a neural network φ for classification, that is:

$$z_i = \varphi(y_i) \tag{13}$$

where z_i denotes the probability distribution over the possible classes. For classification, the cross-entropy loss is utilized, which is formulated as follows:

$$L_C = -\sum_{c=1}^{C} l_i \log z_i \tag{14}$$

where l_i represents the one-hot encoding of the sample label for sample y_i and C represents the total number of classes.

3. Semi-supervised mixup

The insufficiency of training data constitutes a prevalent challenge in machine learning, particularly within the context of multi-view learning. Mixup [27], recognized as an effective data augmentation technique, has inspired the development of an innovative semi-supervised blending approach in this study. This method aims to delve into the feature distribution between labeled and unlabeled data, thereby generating additional samples with reliable labels. The strategy comprises two components: labeled data augmentation and unlabeled data augmentation.

3.1 Labeled data augmentation

For labeled samples, our method begins by sampling a random variable λ' from the Beta distribution. We calculate the final weight parameter by setting $\lambda = \max(\lambda', 1-\lambda')$. Subsequently, for the specific information and labels derived from the labeled samples, augmentation is performed separately according to the following formulas:

$$\begin{aligned} y_{nm,s}^{v} &= \lambda y_{n,s}^{v} + (1-\lambda) y_{m,s}^{v}, v = (1,2,\cdots,M) \\ y_{nm,c}^{v} &= \lambda y_{n,c}^{v} + (1-\lambda) y_{m,c}^{v}, v = (1,2,\cdots,M) \\ l_{nm} &= \lambda l_n + (1-\lambda) l_m \end{aligned} \tag{15}$$

where $y_{n,s}^{v}$ and $y_{m,s}^{v}$ represent the specific information extracted from two randomly chosen labeled samples x_n^{v} and x_m^{v} of the v^{th} view according to (6), $y_{n,c}^{v}$ and $y_{m,c}^{v}$ represent the corresponding shared information. l_n and l_m denote their labels. Subsequently, the components are integrated according to (11), (12), and (13) to generate the final representation. During the prediction phase, rather than relying on (15), the objective is to ensure that the predicted results closely resemble the target representation, a process described by the following formula:

$$L_C = \left\| \varphi(Y_{nm}) - l_{nm} \right\|_F^2 \tag{16}$$

3.2 Unlabeled data augmentation

The aforementioned method can also be applied to unlabeled samples. Randomly select an unlabeled sample x_u and two labeled samples x_n and x_m, and the corresponding labels for the labeled samples are l_n and l_m. According to (16), the augmented specific representation and labels obtained by mixing x_u and x_n are:

$$\begin{aligned} y_{un,s}^{v} &= \lambda y_{u,s}^{v} + (1-\lambda) y_{n,s}^{v}, v = (1,2,\cdots,M) \\ l_{un} &= \lambda l_u + (1-\lambda) l_n \end{aligned} \tag{17}$$

Similarly, the augmented specific representation and label obtained by mixing x_u and x_m can be expressed as:

$$\begin{aligned} y_{um,s}^{v} &= \lambda y_{u,s}^{v} + (1-\lambda) y_{m,s}^{v}, v = (1,2,\cdots,M) \\ l_{um} &= \lambda l_u + (1-\lambda) l_m \end{aligned} \tag{18}$$

In (18)(19), l_u is unknown and therefore cannot participate directly in the training. To circumvent this unknown label, we can subtract the two augmented labels and conduct training by calculating the difference between the two samples. This step can be expressed as:

$$l_{un} - l_{um} = (1-\lambda)(l_n - l_m) \tag{19}$$

According to (12)(13), the final representation after augmentation can be expressed by the following formulas:

$$Y_{un,s} = \left[Y_{un,s}^{1\bullet}, Y_{un,s}^{2\bullet}, \cdots, Y_{un,s}^{M\bullet} \right]^{\bullet} \tag{20}$$

$$Y_{um,s} = \left[Y_{um,s}^{1\bullet}, Y_{um,s}^{2\bullet}, \cdots, Y_{um,s}^{M\bullet} \right]^{\bullet} \tag{21}$$

$$Y_{un} = \left[Y_{un,s}^{\bullet}, Y_{u,c}^{\bullet} \right]^{\bullet} \tag{22}$$

$$Y_{um} = \left[Y_{um,s}^{\bullet}, Y_{u,c}^{\bullet} \right]^{\bullet} \tag{23}$$

Based on the above, we adopt prediction differences in the calculation of the objective function, rather than precise prediction results, which can be presented by the formula:

$$L_{semi} = \sum_{v=1}^{M} \| \left(\varphi(Y_{un}) - \varphi(Y_{um}) \right) - (1-\lambda)(l_n - l_m) \|_F^2 \tag{24}$$

The enhancement process is specifically tailored to view-specific information, rather than raw features or shared information. Experimental results indicate that shared information tends to remain consistent across the entire sample set, while view-specific representations become increasingly distinct as noise levels are reduced. This approach yields samples of high quality and distinctiveness. Consequently, the decision boundaries of the trained classifiers are refined to be more distinct, enhancing the generalization and performance of the model. To prioritize the influence of sample n on the overall performance, the parameter λ is set above 0.5.

4. Experiment

4.1 Experimental Setup

1) Datasets

To comprehensively validate the effectiveness of the model, experiments were conducted on seven datasets:

Caltech101 [30]: Comprises 101 object categories with a total of 9144 samples, employing the standard 102-class setting. Additionally, a subset (Caltech101-7) consisting of 7 categories was selected from Caltech101 for more refined evaluation.

Internet Advertisements (AD) [31]: Focuses on the task of web page advertisement classification, containing two categories: advertisement and non-advertisement web pages. This dataset aims to test the model's classification capability in complex web content.

NUSWIDEOBJ [32]: A multi-view web image dataset with 30,000 images, covering 31 categories and providing rich view features.

Flowers-102: A large flower image dataset comprising 102 categories with a total of 8189 images.

HW [33]: Includes 2,000 images divided into 10 categories. Each image has been extracted with 6 types of features.

Reuters: Contains feature representations of documents written in five different languages (English, French, German, Spanish, and Italian), covering 6 categories.

These datasets, characterized by their varying number of categories, sample sizes, and view counts, collectively constitute a comprehensive testing platform. This platform is designed to validate the model's effectiveness and robustness across different scenarios and tasks. Table 1 provides a detailed description of the seven datasets.

Table 1 Information of datasets

Dataset	# of samples	# of views	# of classes
Caltech101	9 144	6	102
Caltech101-7	1 474	6	7
AD	3 279	3	2
NUSWIDEOBJ	30 000	5	31
Flowers-102	8 189	4	102
HW	2 000	6	10
Reuters	1 8758	5	6

2) Baselines

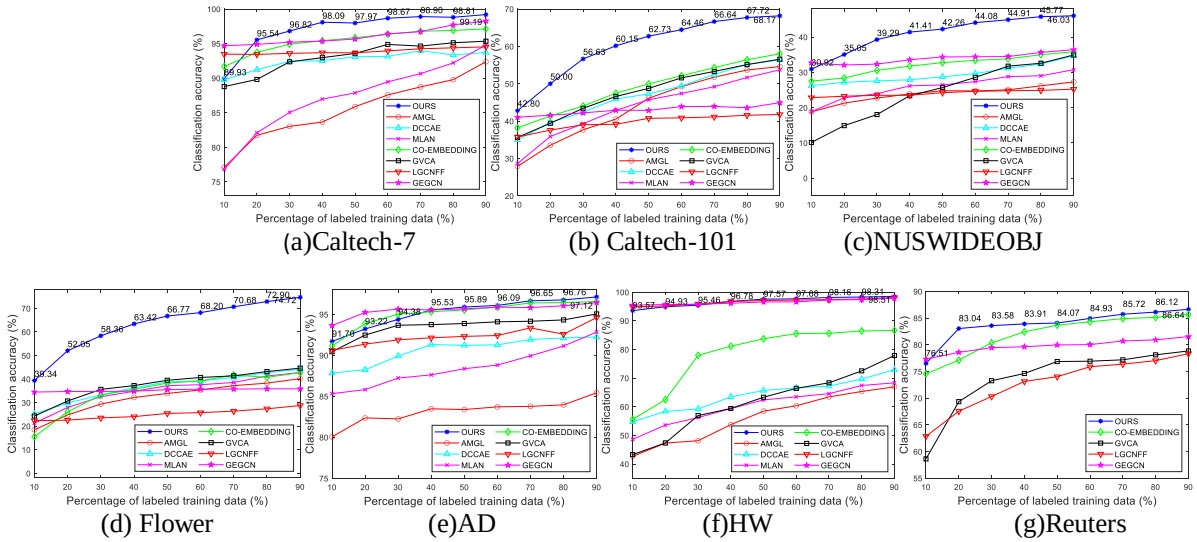
To validate the effectiveness of the proposed method, we compared it with several state-of-the-art approaches for multi-view classification tasks. Specifically, we employed two semi-supervised multi-view learning methods: Auto-weighted Multiple Graph Learning (AMGL) [34] and Multi-view Learning with Adaptive Neighbors (MLAN) [35]. Additionally, we considered five semi-supervised multi-view deep representation learning methods: Deep Canonically Correlated AutoEncoder (DCCAE) [36], Generative View-Correlation Adaptation (GVCA) [29], Co-embedding [2], Learnable Graph Convolutional Network and Feature Fusion (LGCNFF) [37] and Generative Essential Graph Convolutional Network (GEGCN) [38].

3) Implementation Details

In the experiments, the feature embedding networks f_c^v and f_s^v were implemented as three-layer fully connected networks with hidden and output layer sizes set to 512 and 256, respectively, and both employing the ReLU activation function. The classification network φ comprised two hidden layers with sizes set to 256 and 128. The batch size was set to 64, and the Adam optimizer was used with a learning rate of 0.0001. The network depth and the number of neurons per layer for the comparative methods were set according to their original versions. For DCCAE and GVCA, which are designed for two-view problems, we utilized a concatenation strategy to combine multiple views into two, and then selected the concatenation that achieved the highest classification accuracy. All experiments were conducted on the same computer with the following configuration: Intel i7 quad-core 3.6GHz CPU, two NVIDIA GTX1080Ti GPUs, and 16GB of RAM.

Classification accuracy and F1-score were the primary evaluation metrics for these experiments. Each experiment was repeated 20 times, and the average value was taken as the result. Each dataset was randomly divided into training and testing sets, each comprising 50% of the data. To simulate a semi-supervised learning environment, a certain proportion of the training data (ranging from 10% to 90%) was randomly selected as labeled data, with the remainder serving as unlabeled data. For the three parameters in (1), they were set to α =0.5, β =0.7 and γ =1.0.

4.2 Results



(a)Caltech-7 (b) Caltech-101 (c)NUSWIDEOBJ

(d) Flower (e)AD (f)HW (g)Reuters

Figure. 2 Performance comparison of our approach and compared methods on all the datasets

Table 2 F1-score and training time for all datasets across different methods

Method	Dataset (F1-socre|Training time: s)						
	Caltech101-7	Caltech101	NUSWIDEOBJ	Flower	AD	HW	Reuters
AMGL	80.46|27.27	39.44|133.45	11.84|598.23	20.42|392.48	80.63|52.53	45.59|23.31	-
DCCAE	90.28|43.94	41.29|491.88	16.23|1631.19	23.57|689.99	85.98|64.63	51.18|41.53	-
MLAN	79.83|39.52	38.31|274.37	14.26|1289.31	22.56|481.23	82.66|81.92	52.67|28.15	-
CO-EMBEDDIING	90.77|248.75	45.76|2831.25	23.17|2293.24	29.71|3674.22	92.78|392.56	79.31|273.82	81.94|3928.31
GVCA	91.83|381.20	42.12|3493.05	22.99|3294.53	31.34|5929.47	91.34|521.12	58.72|682.91	74.87|22124.26
LGCNFF	90.92|41.69	37.42|312.31	20.21|1461.82	21.49|1589.22	92.65|96.06	96.75|56.89	61.24|1942.39
GEGCN	92.20|74.07	32.89|593.29	21.63|2194.08	24.35|1733.93	93.64|118.35	95.25|92.03	62.54|2364.40
SEMA	94.44|223.16	60.73|1120.58	28.31|2632.44	53.13|1702.05	94.18|187.42	97.36|213.17	83.25|18327.64

The experimental results presented in Fig. 2 thoroughly validate the effectiveness of the proposed method. On the Reuters dataset, due to computational complexity limitations imposed by the algorithm and machine resources, certain models encountered timeouts or memory insufficiency errors and are accordingly not displayed. When addressing datasets with already high classification accuracy (such as Caltech101-7, AD, HW, and Reuters), although further improvements present challenges, the method consistently demonstrates a stable advantage. For datasets with relatively lower classification accuracy (such as Caltech101, NUSWIDEOBJ, and Flower), the method notably outperforms other competing approaches.

Table 2 presents the F1-score and training time of SEMA and other comparative methods across various datasets with 50% labeled data. The results indicate that SEMA significantly outperforms other methods in terms of F1-score, while its training time remains within an acceptable range.

This significant advantage primarily stems from two key factors: Firstly, the method, in its information processing, not only emphasizes the integration of shared and complementary information but also endeavors to minimize redundancy to the greatest extent, thereby effectively enhancing the precision of the representation. Secondly, the implemented data augmentation strategy, by combining labeled and unlabeled data, generates a more diverse and robust set of auxiliary samples. This not only strengthens the model's robustness but also provides substantial support for performance improvement.

To visually demonstrate the classification performance of SEMA, we obtained the representations of all methods on four datasets and projected them into a two-dimensional space using the t-SNE (t-distributed stochastic neighbor embedding) dimensionality reduction technique. Subsequently, the mapped two-dimensional data were color-coded according to the true labels. As illustrated in Fig. 3-6, these visualizations clearly show that the feature representations generated by SEMA are more compact within classes and exhibit clear inter-class boundaries, further validating the effectiveness and superiority of the model.

4.3 Ablation Study

To verify the effectiveness of each component, a series of ablation experiments were designed. In these experiments, the proportion of labeled data was fixed at 50%, and a total of 20 repeated experiments were conducted across multiple datasets to ensure the stability and reliability of the results. The average classification accuracy was used as the primary evaluation metric, with detailed results presented in Table 3.

Table 3 Ablation Study

Method	B+complete-A	C+complete-A	SEMA
Caltech101-7	96.53	97.68	97.79
Caltech101	59.52	60.80	62.73
NUSWIDEOBJ	41.69	42.14	42.26
Flower	64.88	65.31	66.77
AD	94.46	95.53	95.89
HW	96.77	97.48	97.57
Reuters	83.46	83.84	84.07

B+complete-A: Employing the conventional shared-specific framework, as utilized in MDDRL, this approach ensures that specific information is only unrelated to shared information during extraction. In the semi-supervised learning module, data augmentation operations are applied to the entire representation.

C+complete-A: Based on the shared-specific framework, a new orthogonality constraint is introduced to enhance the orthogonality between specific information. In the semi-supervised learning module, the global enhancement strategy is still employed.

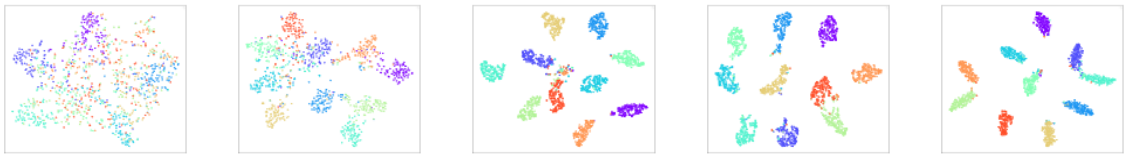
The experimental results demonstrate that the reinforced constraints effectively reduce redundant information in the representation, thereby enhancing the model's representational capacity. Additionally, the introduced specific data augmentation strategy has shown its superiority by significantly improving the model's performance in a semi-supervised learning setting.

(a)GVCA (b)CO-EMBEDDING (c) LGCNFF (d)GEGCN (e)SEMA

Figure. 3 T-SNE visualization results on Caltech101-7 dataset

(a)GVCA (b)CO-EMBEDDING (c) LGCNFF (d)GEGCN (e)SEMA

Figure. 4 T-SNE visualization results on AD dataset



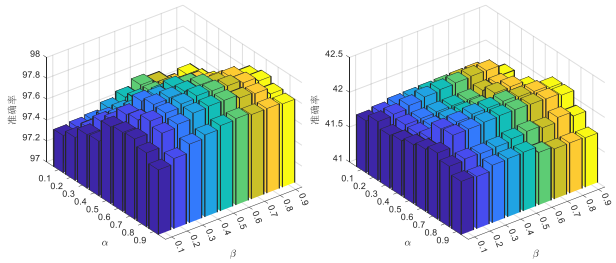
(a)GVCA (b)CO-EMBEDDING (c) LGCNFF (d)GEGCN (e)SEMA

Figure. 5 T-SNE visualization results on HW dataset

(a)GVCA (b)CO-EMBEDDING (c) LGCNFF (d)GEGCN (e)SEMA

Figure. 6 T-SNE visualization results on Reuters dataset

4.4 Effect of Trade-off Parameters



(a)Caltech101-7 (b) NUSWIDEOBJ

Figure. 7 Influence of trade-off parameters on model

This experiment deeply explores the impact of trade-off parameters α and β on model performance. To find the optimal combination of these two parameters, a grid search method was used, and a thorough search was conducted under the condition of 50% labeled data. Each experiment was repeated 20 times to ensure result

stability and reliability, with average classification accuracy serving as the evaluation metric. Fig. 7 clearly illustrates the experimental outcomes on the Caltech101-7 and NUSWIDEOBJ datasets. For Caltech101-7, the optimal values for α and β were identified as 0.7 and 0.5, respectively; meanwhile, for NUSWIDEOBJ, the optimal values were 0.6 for both parameters. The results indicate a significant decline in model performance when α and β values are either too high or too low, underscoring the criticality of parameter selection.

Furthermore, similar experiments were performed on additional datasets. The optimal values for α and β were 0.8 and 0.5 on Caltech101, 0.7 and 0.6 on AD, 0.9 and 0.5 on Flower, 0.4 and 0.3 on HW, and 0.8 and 0.4 on Reuters. These findings further substantiate the substantial impact of trade-off parameters on model performance and reveal the variability in optimal values across different datasets.

5. Conclusion

In this paper, we propose an innovative unified semi-supervised multi-view feature learning method SEMA. SEMA leverages a shared-and-specific framework and significantly enhances orthogonality constraints. By deeply exploring shared and view-specific information, this approach effectively reduces redundancy in the learned feature representations, thereby generating more precise and efficient feature representations. Furthermore, a meticulously designed specific data augmentation strategy, which utilizes labeled and unlabeled data to generate reliable auxiliary samples, further enhances the model's performance in a semi-supervised learning environment. Extensive experimental evaluations consistently demonstrate the superior effectiveness of this method across seven public datasets, fully validating its substantial potential and value in practical applications. More comprehensive ablation experiments further reveal that each module within the method plays an indispensable role, collectively contributing to a significant enhancement in the final performance. Future research can further explore more complex and diverse data augmentation strategies to better accommodate various types of datasets and task requirements. Additionally, investigating how to more effectively integrate enhanced orthogonality constraints with deep learning models to achieve higher-level feature representation and classification performance remains an important direction worthy of in-depth research and exploration.

Data sharing agreement

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, author-ship, and publication of this article.

Funding

The research was supported by the NSFC Project under Grant No. 62176069 and 61933013; the Natural Science Foundation of Guangdong Province under Grant No. 2023A1515012653; the Innovation Group of Guangdong Education Department under Grant No. 2020KCXTD014.

References

[1] L. Wang, B. Sun, J. Robinson, T. Jing, and Y. Fu (2020). EV-action: Electromyography-vision multi-modal action dataset. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina: IEEE, pp. 160-167.

[2] X. Jia, X.-Y. Jing, X. Zhu, Z. Cai, and C.-H. Hu (2021). Co-embedding: a semi-supervised multi-view representation learning approach. Neural Computing and Applications, 34(6): 4437–4457.

[3] L. Wang, Y. Liu, C. Qin, G. Sun, and Y. Fu (2020). Dual relation semi-supervised multi-label learning. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(04): 6227-6234.

[4] Y. Cheng, X. Zhao, R. Cai, Z. Li, K. Huang, and Y. Rui (2016). Semi-supervised multimodal deep learning for RGB-D object recognition. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16), AAAI Press, pp. 3345–3351.

[5] S. Savarese and L. Fei-Fei (2010). Multi-view object categorization and pose estimation. In Studies in Computational Intelligence, Springer, pp. 205-231.

[6] D. Du, L. Wang, H. Wang, K. Zhao, and G. Wu (2019). Translate-to-recognize networks for RGB-D scene recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11836–11845.

[7] X. Xue, F. Nie, S. Wang, X. Chang, B. Stantic, and M. Yao (2017). Multi-view correlated feature learning by uncovering shared component. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2810–2816.

[8] G. Chao and S. Sun (2016). Consensus and complementarity based maximum entropy discrimination for multi-view classification. Inf. Sci., 367: 296–310.
[9] T. Baltruaitis, C. Ahuja, and L.-P. Morency (2019). Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2): 423-443.
[10] Y. Li, M. Yang, and Z. Zhang (2019). A Survey of multi-view representation learning. IEEE Transactions on Knowledge and Data Engineering, 31(10): 1863-1883.
[11] S. El Hajjar, F. Dornaika, and F. Abdallah (2022). One-step multi-view spectral clustering with cluster label correlation graph. Inf. Sci., 592: 97-111.
[12] Q. Qiang, B. Zhang, F. Wang, and F. Nie (2021). Fast multi-view discrete clustering with anchor graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(11), pp. 9360-9367.
[13] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li (2021). Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1407-1417.
[14] X.-Y. Jing, F. Wu, X. Dong, S. Shan, and S. Chen (2017). Semi-supervised multi-view correlation feature learning with application to webpage classification. In Proceedings of the AAAI Conference on Artificial Intelligence, 31(1), pp. 1374-1381.
[15] Y. Peng, J. Qi, and Y. Yuan (2018). Modality-specific cross-modal similarity measurement with recurrent attention network. IEEE Transactions on Image Processing, 27(11), pp. 5585-5599.
[16] H. Hotelling (1936). Relations between two sets of variates. Biometrika, 28(3-4), 321-377.
[17] S. Akaho (2006). A kernel method for canonical correlation analysis. Journal of Machine Learning Research, 7, 1483-1506.
[18] J. Rupnik and J. Shawe-Taylor (2010). Multi-view canonical correlation analysis. In Proceedings of the 12th Slovenian International Conference on Knowledge Discovery and Data Mining (SiKDD), pp. 1-4.
[19] G. Andrew, R. Arora, J. Bilmes, and K. Livescu (2013). Deep canonical correlation analysis. In Proceedings of the 30th International Conference on Machine Learning (ICML), pp. 1247-1255.
[20] W. Guo, J. Wang, and S. Wang (2019). Deep multimodal representation learning: a survey. IEEE Access, 7, 63373-63394.
[21] Q. Zheng, J. Zhu, and Z. Li (2022). Collaborative unsupervised multi-view representation learning. IEEE Transactions on Circuits and Systems for Video Technology, 32(7), 4202-4210.
[22] S. Luo, C. Zhang, W. Zhang, and X. Cao (2018). Consistent and specific multi-view subspace clustering. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI), pp. 3730-3737.
[23] T. Zhou, C. Zhang, C. Gong, H. Bhaskar, and J. Yang (2020). Multiview latent space learning with feature redundancy minimization. IEEE Transactions on Cybernetics, 50(4), 1655-1668.
[24] J. Hu, J. Lu, and Y.-P. Tan (2018). Sharable and individual multi-view metric learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(9), 2281-2288.
[25] J. Xu, W. Li, X. Liu, D. Zhang, J. Liu, and J. Han (2020). Deep embedded complementary and interactive information for multi-view classification. In Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI), pp. 6494-6501.
[26] X. Jia, X. Jing, X. Zhu, S. Chen, B. Du, Z. Cai, Z. He, and D. Yue (2021). Semi-supervised multi-view deep discriminant representation learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(7), 2496-2509.
[27] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz (2018). Mixup: Beyond empirical risk minimization. In Proceedings of the 6th International Conference on Learning Representations (ICLR), pp. 1-13.
[28] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel (2019). Mixmatch: A holistic approach to semi-supervised learning. In Advances in Neural Information Processing Systems, pp. 1-14.
[29] Y. Liu, L. Wang, Y. Bai, C. Qin, Z. Ding, and Y. Fu (2020). Generative view-correlation adaptation for semi-supervised multi-view learning. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 318-334.
[30] L. Fei-Fei, R. Fergus, and P. Perona (2007). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. Computer Vision and Image Understanding, 106(1), 59-70.
[31] N. Kushmerick (1999). Learning to remove Internet advertisements. In Proceedings of the Third Annual Conference on Autonomous Agents (AGENTS'99), Association for Computing Machinery, pp. 175-181.
[32] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng (2009). NUS-WIDE: A real-world web image database from National University of Singapore. In Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR'09), Association for Computing Machinery, 48(1), 1-9.
[33] H. Tao, C. Hou, F. Nie, and J. Zhu (2017). Scalable multi-view semi-supervised classification via adaptive regression. IEEE Transactions on Image Processing, 26(9), 4283-4296.

[34] F. Nie, J. Li, and X. Li (2016). Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16), AAAI Press, pp. 1881-1887.
[35] F. Nie, G. Cai, J. Li, and X. Li (2018). Auto-weighted multi-view learning for image clustering and semi-supervised classification. IEEE Transactions on Image Processing, 27(3), 1501-1511.
[36] W. Wang, R. Arora, K. Livescu, and J. Bilmes (2015). On deep multi-view representation learning. In Proceedings of the 32nd International Conference on Machine Learning (ICML), PMLR 37, pp. 1083-1092.
[37] Z. Chen, L. Fu, J. Yao, W. Guo, C. Plant, and S. Wang (2023). Learnable graph convolutional network and feature fusion for multi-view learning. Information Fusion, 95, 109-119.
[38] J. Lu, Z. Wu, L. Zhong, Z. Chen, H. Zhao, and S. Wang (2024). Generative essential graph convolutional network for multi-view semi-supervised classification. IEEE Transactions on Multimedia, 26(1), 7987-7999.