Semi-supervised Multi-view Deep Discriminant Representation Learning

Xiaodong Jia*, Xiao-Yuan Jing†, Xiaoke Zhu‡, Songcan Chen, Bo Du, Ziyun Cai, Zhenyu He, and Dong Yue

Abstract—Learning an expressive representation from multi-view data is a key step in various real-world applications. In this paper, we propose a Semi-supervised Multi-view Deep Discriminant Representation Learning (SMDDRL) approach. Unlike existing joint or alignment multi-view representation learning methods that cannot simultaneously utilize the consensus and complementary properties of multi-view data to learn inter-view shared and intra-view specific representations, SMDDRL comprehensively exploits the consensus and complementary properties as well as learns both shared and specific representations by employing the shared and specific representation learning network. Unlike existing shared and specific multi-view representation learning methods that ignore the redundancy problem in representation learning, SMDDRL incorporates the orthogonality and adversarial similarity constraints to reduce the redundancy of learned representations. Moreover, to exploit the information contained in unlabeled data, we design a semi-supervised learning framework by combining deep metric learning and density clustering. Experimental results on three typical multi-view learning tasks, i.e., webpage classification, image classification, and document classification demonstrate the effectiveness of the proposed approach.

Index Terms—Semi-supervised multi-view deep representation learning, consensus and complementarity, redundancy, adversarial training, Siamese network, density clustering.

1 INTRODUCTION

There are two characteristics in human perception of surroundings. First, human beings perceive the external world via multiple senses, such as vision, hearing, and touch. This implies that: the information faced by the brain is multi-view; the brain is capable of multi-view learning [1], [2]. Second, this perceptual process is always performed under semi-supervised conditions. One important reason is obtaining sufficient labeled data is often highly expensive. An example is medical diagnostic data, which requires not only expensive machinery but also time-consuming analysis by multiple experts. Because of these two characteristics, it is important for those intelligent machines, which are designed to tackle real-world problems for human beings, to have the ability to learn representations from semi-supervised multi-view data. Semi-supervised multi-view representation learning, which is devised to learn representations form semi-supervised multi-view data, unsurprisingly, is widely applied in real-world applications [3], [4].

Multi-view data has three basic characteristics. On the one hand, multi-view data has two inherent advantages: consensus property and complementary property [2], [5]. Consensus property instructs that classifications or representations learned from different views should be consistent with each other. From information perspective, it points out that each view contains information that shared by all views (inter-view shared information). Complementary property indicates that each view also contains information that is unique to itself (intra-view specific information). On the other hand, multi-view data has one disadvantage—redundancy [6], [7]. Different views bring not only diversity but also redundancy. Considering these advantages and disadvantages, the key of multi-view representation learning lies in how to effectively utilize the consensus and complementary property of multi-view data while properly handling the redundancy.

1.1 Motivation

Driven by numerous practical applications, a myriad of semi-supervised or supervised multi-view representation learning methods have been proposed, and they can be mainly categorized into two paradigms: joint representation and alignment representation [3], [4]. Joint representation methods fuse multiple views through concatenation, such as graph-based methods [8], [9], [10] and neural network-based methods [11], [12], [13], [14]. Alignment representation methods maximize the agreement among representations learned from different views via alignment. Common alignment strategies include Canonical Correlation Analysis (CCA) based alignment [15], [16], [17], [18], [19], which maximizes the correlations between representations, and...
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2020.2973634, IEEE Transactions on Pattern Analysis and Machine Intelligence

1.2 Contribution

The contributions of our work are summarized as follows:

1. We investigate the field of multi-view representation learning and present a new method taxonomy. This taxonomy fills the blanks of existing mainstream taxonomy with regard to the emerging shared and specific methods.

2. We propose the Multi-view Deep Discriminant Representation Learning (MDDRL) approach. With the designed orthogonality and adversarial similarity constraints, MDDRL not only comprehensively exploits the consensus and complementary properties of multi-view data but also reduces the redundancy of learned representations. MDDRL, to the best of our knowledge, is the first multi-view deep representation learning approach that has taken all the consensus, complementarity, and redundancy characteristics of multi-view data into consideration.

3. By combining deep metric learning and density clustering, we design a novel Semi-supervised learning framework for MDDRL, and thus propose the SMDDRL approach. In this framework, deep metric learning and density clustering complement each other and jointly improve their performance. Specifically, deep metric learning benefits density clustering by making the decision boundaries clearer; density clustering improves deep metric learning by bringing more labeled data. Moreover, they cooperate with each other and effectively exploit the unlabeled data to enhance the representation learning performance.

4. We conduct extensive experiments on three typical multi-view learning tasks to evaluate SMDDRL. Experimental results demonstrate that SMDDRL outperforms state-of-the-art multi-view representation learning methods. Moreover, we conduct extensive experiments to further investigate SMDDRL, such as ablation study, loss effectiveness investigation, and comparison with other possible implementation schemes of our idea. Experimental results not only verify the effectiveness and reliability of each component that we have incorporated in SMDDRL, but also indicate the superiority and robustness of SMDDRL on multi-view representation learning.
1.3 Organization

The rest of this paper is organized as follows: we will start by reviewing the related works on multi-view representation learning, then move on to introduce our SMDDRL approach in Section 3. In Section 4, we compare SMDDRL with other state-of-the-art methods. In Section 5, we conduct more experiments to investigate SMDDRL. Finally, we conclude the paper in Section 6.

2 RELATED WORK

2.1 Classical Multi-view Feature Learning Works

Representation learning is a prerequisite step in many multi-view learning tasks. In recent years, a variety of classical multi-view representation learning methods have been proposed. These methods follow the previously presented taxonomy, i.e., joint representation [8], [10], [38], [39], alignment representation [15], [16], [40], [41], [42], [43], as well as shared and specific representation [7], [24], [27], [28], [29], [30], [44], [45], [46]. For example, based on Markov network, Chen et al. [39] presented a large-margin predictive multi-view subspace learning method, which joints features learned from multiple views. Jing et al. [41] proposed an intra-view and inter-view supervised correlation analysis method for image classification, in which CCA was applied to align multi-view features. Luo et al. [30] presented a consistent and specific multi-view subspace clustering algorithm. It learns a representation subspace, in which multi-view data are represented by a set of view-specific representations and a view-consistent representation that shared by all views.

Although these methods have promoted the development of multi-view representation learning, they are limited to shallow features. Nowadays, deep learning is recognized as the state-of-the-art representation learning technique, consequently, more and more deep multi-view representation learning methods appeared.

2.2 Deep Multi-view Representation Learning Works

Deep multi-view representation learning works also follow the joint representation [11], [12], [13], [14], alignment representation [17], [18], [19], [20], [21], [22], [47], as well as shared and specific representation [25], [26] classification paradigm. For example, Kan et al. [48] proposed a multi-view deep network for cross-view classification. This network first extracts view-specific features with a sub-network, then concatenates and feeds these features into a common network, which is designed to project them into one uniform space. Harwath et al. [22] presented an unsupervised audio-visual matchmap neural network, which applies similarity metric and pairwise ranking criterion to align visual objects and spoken words. Hu et al. [25] introduced a sharable and individual multi-view deep metric learning method. It leverages view-specific networks to extract individual features from each view and employs a common network to extract shared features from all views.

Table 1 summarizes the recent presented classical and deep multi-view representation learning works. Different from classical works, SMDDRL exploits the state-of-the-art representation learning technique—deep learning; Different from deep learning works, SMDDRL comprehensively considers all the three characteristics of multi-view data.

### Table 1: Multi-view representation learning methods

<table>
<thead>
<tr>
<th>Categories</th>
<th>Representation Learning Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Classical Non-Deep</td>
</tr>
<tr>
<td>Joint Representation</td>
<td>[8], [10], [38], [39]</td>
</tr>
<tr>
<td>Alignment Representation</td>
<td>[15], [40], [41], [42]</td>
</tr>
<tr>
<td>Shared and Specific Representation</td>
<td>[7], [24], [27], [28]</td>
</tr>
</tbody>
</table>

2.3 Semi-supervised Learning works

2.3.1 General Semi-supervised Learning Works

Due to the lack of labeled data, Semi-Supervised Learning (SSL) has always been a hot topic in machine learning. A myriad of SSL methods have been proposed. For example, Co-training is a well-know disagreement-based SSL method, which trains different learners to exploit unlabeled data [49]. Pseudo-label style methods label unlabeled data with pseudo labels [50], [51], [52], [53]. Graph-based methods aim to construct a similarity graph, through which label information propagates to unlabeled nodes [54], [55], [56], [57]. Local-smoothness-regularization-based methods are another popular group of SSL methods. Different methods apply different regularizers, such as Laplacian regularization [58], manifold regularization [59], virtual adversarial regularization [60]. For example, Miyato et al. [60] proposed a smooth regularization method called virtual adversarial training, which enables the model to output a smooth label distribution for local perturbations of a given input. There are other popular methods, e.g., Ladder Network [61].

2.3.2 Semi-supervised Multi-view Learning Works

In recent years, various semi-supervised multi-view learning methods have been presented, such as co-training based methods [11], [62], [63], alignment-based methods [15], [16], [19], [24], graph-based methods [10], [38], and other traditional SSL strategy based methods. Among them, co-training based methods and alignment-based methods are the most popular ones. For example, Cheng et al. [11] presented a semi-supervised multimodal deep learning framework for RGB-D object recognition. This framework employs a diversity preserving Co-training component to label unlabeled data. Noroozi et al. [19] presented a semi-supervised multi-view representation learning network, which utilizes unlabeled training data by maximizing the correlation between them with CCA. In addition, there are some works that apply conventional SSL strategy to utilize the unlabeled data, such as Laplacian regularization [64] and manifold smoothness regularization [65].

Different from all these methods, we present a pseudo-label-style SSL framework by combining deep metric learning and density clustering. This framework provides a new choice for using unlabeled data.
3 Semi-supervised Multi-view Deep Discriminant Representation Learning (SMDDRL)

As discussed in Section 1, the key to multi-view representation learning lies in effectively utilizing consensus and complementary properties while handling the redundancy properly. In this paper, we propose a Semi-supervised Multi-view Deep Discriminant Representation Learning (SMDDRL) approach that fulfills this purpose. SMDDRL contains two parts: a) Multi-view Deep Discriminant Representation Learning (MDDRRL, introduced in Section 3.1); b) deep metric and density clustering based semi-supervised learning framework. Building on the representation learning and classification backbone, MDDRRL adds two components: a) deep metric learning for better discriminability; b) orthogonality and adversarial similarity constraints for disentangling shared and specific features. Semi-supervised learning framework contains deep metric learning (in MDDRRL) and density peak clustering, and they are trained alternatively (Step1-Step4).

3.1 Multi-view Deep Discriminant Representation Learning (MDDRRL)

Building on the foundation of representation learning and classification backbone, MDDRRL incorporates two components to enhance its performance, i.e., deep metric learning as well as orthogonality and adversarial similarity constraints. Accordingly, the loss function of MDDRRL contains three parts, and it is formulated as the weighted sum of the classification loss $L_c$, the constraint losses ($L_{diff} + L_{Adv}$), and the contrastive loss of deep metric learning $L_{con}$:

$$L = L_c + \lambda_1(L_{diff} + L_{Adv}) + \lambda_2L_{con},$$  

where $\lambda_1$, $\lambda_2$ are trade-off parameters. Fig. 3 illustrates the architecture of MDDRRL. Subsequently, we will first introduce the representation learning and classification backbone, then detail the two added components.

3.1.1 Shared and Specific Representation Learning

Shared and specific representation learning seeks to handle the inter-view shared and intra-view specific information separately and employs multiple feature extractors for each view. Suppose $X = \{x_i \in \mathbb{R}^d\}_{i=1}^N$ is $N$ examples of training data, where $d$ is the dimensionality of example $x_i$, and each example has $M$ views. Let $X^k = \{x_{ik} \in \mathbb{R}^{d_k}\}_{i=1}^N$ be the feature set in the $k$th view, and $x_{ik}$ denotes the $i$th view of $k$th example, $d_k$ is the dimensionality of $x_{ik}$, where $k = 1, 2, \cdots, M$, and $d = \sum_{k=1}^M d_k$. For $x_{ik}$, we employ two networks to project it into a subspace, one for shared information and one for specific information. In this space, the shared information can be represented by $h_{c,i}^k = W_{c,i}^k x_{ik}$, where $W_{c,i}^k \in \mathbb{R}^{1 \times d_k}$. The specific information can be represented by $h_{s,i}^k = W_{s,i}^k x_{ik}$, where $W_{s,i}^k \in \mathbb{R}^{2 \times d_k}$. Specifically, for the $k$th view, the representation can be written as:

$$h_i^k = \begin{bmatrix} h_{c,i}^k \\ h_{s,i}^k \end{bmatrix} = \begin{bmatrix} W_{c,i}^k \\ W_{s,i}^k \end{bmatrix} x_i. \quad (2)$$

Since the shared information of different views is almost the same, it’s unnecessary to include all of them into the final representation. Instead, we use the average:

$$h_{c,i} = \frac{1}{M} \sum_{k=1}^M h_{c,i}^k. \quad (3)$$

Finally, we joint the average of shared information and specific information from all views to represent the example:

$$h_i = [h_{c,i}^T, h_{s,i}^T, h_{s,i}^T, \cdots, h_{s,i}^T, \cdots, h_{s,i}^T]^T. \quad (4)$$

After representing multi-view examples into a subspace, we classify them with a multilayer network and train this classifier and the representation learning network simultaneously. Cross-Entropy Loss is used for classification.

3.1.2 Orthogonality and Adversarial Similarity Constraint

Shared and specific methods intend to handle inter-view shared and intra-view specific information separately. These two kinds of information, however, are not automatically...
separated. To separate share and specific information as well as to reduce the redundancy between them, we design the orthogonality constraint and adversarial similarity constraint. Orthogonality constraint disentangles shared and specific information as well as prevents them from contaminating each other. Adversarial similarity constraint draws on adversarial training and can ensure the similarity of shared information. Details are described as follows:

Inspired by work [31] and [33], we apply orthogonality constraint to separate the shared and specific information of each view. Let $S^k$ and $H^k$ be the matrices whose rows are shared and specific outputs from the $k^{th}$ view, orthogonality loss of the $k^{th}$ view is defined as:

$$L_{diff} = \|S^k H^k\|_F^2,$$  \hspace{1cm} (5)

where $\| \cdot \|_F^2$ is the squared Frobenius norm.

Inspired by the work [31], [66], [67], we borrow the adversarial training—which derives from Generative Adversarial Network [68]—to design the similarity constraint. Specifically, we treat the representation learning networks as generators $G$, and take the shared information as generated results. Then we employ a $M$-classes classifier as the discriminator $D$, which is used to discriminate the origin view of each generated shared information, or in other words, to discriminate the distribution to which each generated shared information belongs. For the $k^{th}$ view of $i^{th}$ example $x_i^k$, let $G_k(x_i^h)$ be the shared information generated from it. The probability that $G_k(x_i^k)$ derives from the $k^{th}$ view is:

$$P_k^i = D(G_k(x_i^k)),$$  \hspace{1cm} (6)

where $G_k$ is the generator of the $k^{th}$ view. In training phase, discriminator $D$ seeks to figure out the origin view of each shared information, i.e., for $G_k(x_i^h)$, to maximize $P_k^i$. On the contrary, generator $G_k$ seeks to minimize this probability. Put these two goals together, we get a minmax game:

$$L_{Adv} = \min_{\theta_G} \max_{\theta_D} \left( \sum_{i=1}^{N} \sum_{k=1}^{M} I_k^i \log D(G_k(x_i^k)) \right).$$  \hspace{1cm} (7)

where $I_k^i$ denotes the ground-truth view label of $x_i^k$.

We train the discriminator and the generators simultaneously until the discriminator cannot distinguish the differences between shared information generated from different views. In other words, the shared information of different views are similar enough.

### 3.1.3 Siamese Network for Deep Metric Learning

To enhance the discriminability of learned representations, we introduce deep metric learning into SMDDRL and implement it with Siamese network [69]. Besides, we improved Siamese network, such that it converges faster. Siamese network takes paired samples as input. If the paired samples come from the same class, Siamese network seeks to pull them closer. Otherwise, Siamese network tries to push them farther than the Margin, which is a hyperparameter. Here, Siamese network employs squared Euclidean distance as distance metric. Specifically, for a pair of samples $x_i$ and $x_j$, after representing as $h_i$ and $h_j$ by the representation learning network, Siamese network projects them into a subspace, in which they are represented by $\text{code}_i$ and $\text{code}_j$, the distance between them is computed as:

$$d(x_i, x_j) = ||\text{code}_i - \text{code}_j||^2.$$  \hspace{1cm} (8)

Siamese network uses contrastive loss:

$$L = \frac{1}{2N} \sum_{n=1}^{N} \left[ y_n d_n^2 + (1 - y_n) \max(x_n - d_n, 0) \right],$$  \hspace{1cm} (9)

where $d_n$ is the distance (calculated by Eq.8) of the $n^{th}$ paired input, $y_n$ denotes whether the paired samples are from the same class. If they are, $y_n = 1$, otherwise, $y_n = 0$.

Traditional Siamese network selects paired input randomly, leading to the result that contrastive loss shakes a lot, which not only adds difficulties to getting a stable result but also hampers the performance. Thus, we improve the contrastive loss. For a mini-batch, the new loss is defined as:

$$L_{Con} = \frac{1}{2m} \sum_{i=1}^{m} \left\{ (x_i - \mu_{same})^2 + \max(x_i - \mu_{diff}, 0)^2 \right\},$$  \hspace{1cm} (10)
where $m$ is the batch-size, $x_i$ is the $i^{th}$ training sample in this mini-batch. For $x_i$, $\mu_{same}$ is the average of samples in this mini-batch that have the same class label with $x_i$, $\mu_{diff}$ is the average of samples that have different class labels with $x_i$. Besides, we shuffle the train set in each iteration.

The improved loss converges faster than the original one. The reason is two-fold: Firstly, the average of a mini-batch is more stable than the randomly selected samples. Secondly, in Euclidean space, the class center of each class converges to the average of its examples. When each sample has a stable target (the class center), the loss converges faster.

### 3.2 Deep Metric Learning and Density Clustering based Semi-supervised Learning Framework

We combine deep metric learning and density clustering for semi-supervised learning. Reasons are as follows:

Recently, several semi-supervised deep learning methods have been introduced [50], [51], [52], [53], [59], [60], [61], [70], [71]. Among them, pseudo-label style methods [50], [51], [52], [53] show promising results and outperforms other contemporary competitors. The possible reason is: Deep neural networks have tons of parameters to be tuned. More parameters require more labeled data. Pseudo-label style methods directly bring labeled data, which is more favorable for deep models. Hence, we prefer pseudo-label style semi-supervised learning strategies.

In our representation learning network, we employ deep metric learning, which aims to learn a space where examples from the same class are close, while examples from different classes are far apart. In other words, deep metric learning lowers the example density of class boundaries and makes them clearer. Coincidentally, this is exactly the cluster assumption: decision boundary should lie in low-density regions to improve the generalizability of clustering algorithms [53], [56], [72]. Thus, deep metric learning could benefit clustering by making decision boundaries clearer. Experimental results (see Fig.4) also confirm this discovery. Thus, we combine deep metric learning with clustering to label the unlabeled data.

We adopt the density peak clustering algorithm [73]. The reason is two-fold: Firstly, density peak clustering is efficient (with the time complexity of $O(N)$). Secondly, density clustering can discover clusters of arbitrary shapes.

### 3.2.1 Alternative Optimization

We take an alternative training strategy to optimize the two components (deep metric learning of MDDRL and density clustering) of our semi-supervised learning framework. As shown in Fig.2, there are four main steps:

**Step 1:** Train MDDRL with labeled training data, then calculate the validation accuracy.

**Step 2:** Encode labeled and unlabeled data with the trained MDDRL and get the representations (CODE).

**Step 3:** Cluster the CODE and get clustering result.

**Step 4:** Label unlabeled data with the clustering result.

To improve accuracy, we recheck the new labeled data with the classification network, i.e., only reserve the sample whose clustering label agrees with its classification label. Then we stack them to the original labeled training data.

We repeat these steps in each iteration until the validation accuracy does not improve anymore or all the unlabeled examples are labeled. Finally, we run the test on test set.

In this framework, MDDRL and density clustering mutually reinforce each other. On the one hand, MDDRL benefits density clustering in the following two ways: Firstly, it facilitates clustering by making the decision boundaries clearer; Secondly, it reduces the data dimension, which prevents density clustering from the curse of dimensionality.

On the other hand, density clustering label unlabeled data for MDDRL. With more labeled data, MDDRL can yield a better representation space.

### 3.2.2 Implementation Details

To implement this semi-supervised learning framework, the following two details need to be noticed.

1. We choose squared Euclidean distance as the distance metric of density clustering algorithm. Since all the deep metric learning, the density clustering, and the classification network are based on the same representation space, they share the same distance metric.

2. The original labeled training data stays unchanged through all iterations. This setting is to prevent the semi-supervised algorithm from becoming a greedy one.

### 4 Comparisons with State-of-the-art

We evaluate SMDDRL on three typical multi-view learning tasks, namely webpage classification, image classification, and document classification. We adopt several baselines:

Three semi-supervised multi-view learning methods:
- Vertical Ensemble Co-Training (VE-CoT) [62] (Co-training style method).
- Auto-weighted Multiple Graph Learning (AMGL) [38] (Graph-based method).
- Multi-view Learning with Adaptive Neighbors (MLAN) [10] (Graph-based method).

Three semi-supervised multi-view deep representation learning methods:
- Deep Canonically Correlated AutoEncoder (DCCAE) [18] (CCA-based alignment representation method).
Multi-view Discriminative Neural Network (MDNN) [19] (CCA-based alignment representation method)

The implementations of AMGL and MLAN as well as DCCAE2 are downloaded from the authors’ websites. Others are implemented by ourselves. We take classification accuracy and F1-score as our evaluation measures.

4.1 Webpage Classification

4.1.1 Datasets and Settings

In addition to baselines, we also compare another two multi-view feature learning methods: Vector-valued Reproducing Kernel Hilbert Spaces (VRKHS) [74] and Semi-supervised Multi-view Correlation Feature Learning (SMCFL) [16].

All these methods are evaluated on two widely used webpage datasets: 1) WebKB [49] contains 1051 webpages, two classes, 230 for course and 821 for no-course. Each webpage is represented by two views: 3000-dimensional page view and 1840-dimensional link view. 2) Internet Advertisements (AD for short) [75] contains 3279 webpages, two classes, 458 advertisements and 2821 no-advertisements. We use a preprocessed version of AD and exploit three views: 495-dimensional base URL, 457-dimensional image URL, and 1840-dimensional link view.

For each dataset, we randomly select 50% data for training and use the other 50% for testing. To simulate semi-supervised scenario, we randomly select certain percent (10%, 30%, 50%, 70%, 90%) of training data as labeled data, and the remaining as unlabeled data. For those methods that need validation set, i.e., VE-CoT, DCCAE, and SMDDRL, we randomly select half of the labeled data as validation set. For VE-CoT, SMDLF, DCCAE and MDNN, we use two views with the best classification accuracy for them in AD, since these methods only apply to two-view-based tasks. The two classifiers used in VE-CoT are RBF SVM and Gaussian Naive Bayes. For DCCAE, we train the autoencoders with labeled and unlabeled data, and train the classifier with labeled data. In SMDDRL, the hyperparameter $\lambda_1$, $\lambda_2$ are set as 0.45, 0.85 for WebKB, and 0.65, 1.15 for AD (Obtained from grid search, details are presented in Subsection 5.5). Margin is set as 3.0. Multilayer network used in all deep models are four-layer full connection network (input layer not included), the nodes of four layers are $\{128, 64, 64, 32\}$. We use Adam optimizer with the beginning learning rate of $3 \times 10^{-4}$. Besides, dropout (rate=0.5) and $L_2$ Regularization are used to regularize the model. For each ratio of labeled data, we repeat every method 20 times and record the average.

4.1.2 Results and Discussions

Fig. 5 demonstrates the results. As shown in Fig. 5, SMDDRL significantly outperforms other competitors. The reasons are three-fold: Firstly, compared with non-deep methods, SMDDRL employs deep learning, which ensures the qualities of learned representations. Secondly, different from joint or alignment methods, SMDDRL takes both the consensus and complementary properties into consideration. It not only learns comprehensive representations but also reduces the impact of redundancy. Thirdly, with the proposed semi-supervised learning framework, SMDDRL effectively exploits the information contained in unlabeled data.

4.2 Image Classification

4.2.1 Datasets and Settings

In addition to baselines, we also compared another state-of-the-art method—Multi-view semi-supervised classification Via Adaptive Regression (MVAR) [8].

Two large-scale datasets are used in this task. 1) NUS-Object 4 contains 30K images of 31 concepts (such as bear, birds, cow, tiger, whales). Each image is represented by five types of low-level features: 64 color histogram, 144 color correlogram, 73 edge direction histogram, 128 wavelet texture, and 225 block-wise color moment. 2) Noisy MNIST [18] contains 70K grayscale images of digits 0 to 9. Two views are created by rotating and adding random noise to the images of MNIST. For noisy MNIST, we use the default split of train and test (60K, 10K). For methods that require validation sets, we split 10K examples from the train set. Other settings are the same as that in webpage classification.

4.2.2 Results and Discussions

Figure 6 displays the results. We can notice that: a) In most of the cases, DCCAE, MDNN and SMDLF outperform VE-CoT, AMGL and MLAN. These methods exploit deep learning, which can learn better representations with enough training data. b) In deep learning methods, SMDDRL outperforms alignment methods, such as DCCAE and MDNN. The possible reason is: Alignment methods seek to
maximize the agreement among different views. In this task, they aim at maximizing the agreement between subjects of multi-view images (e.g., animals in NUS-Object), leading to the negligence of specific information (e.g., habitats of animals), which also contains discriminative information that benefits classification. SMDDRL integrates both the inter-view shared and intra-view specific information into representations, thus achieving better performance.

### 4.3 Document Classification

#### 4.3.1 Datasets and Settings

In addition to baselines, we also compared Co-Labeling [76] in this experiment. All the methods are evaluated on two datasets: a) BBC\(^5\) contains 2225 documents of five topics (business, entertainment, politics, sports, and tech). Following [77], we randomly partition each document into two views. b) Reuters\(^6\) contains 18758 multilingual documents of 6 classes. Each document is written in five languages, i.e., English, French, German, Italian, and Spanish. Following [76], the documents with multiple class label are annotated using the label of their smallest class. Following [78], we conduct two-language classification on Reuters with 50\% labeled data. Other settings remain unchanged.

#### 4.3.2 Results and Discussions

Figure 7, Table 2 and Table 3 report the document classification results, which are similar to the trend revealed in image classification. A detail that should be noticed is: On BBC datasets, the performance of SMDDRL is only slightly better than that of Co-Labeling. The main reason is that BBC dataset only contains 2225 observations. Deep learning methods cannot be trained adequately. On Reuters dataset, however, with sufficient training data, Deep learning methods (SMDLF and SMDDRL) outperform Co-Labeling.

From information theory perspective, multi-view representation learning methods should make the most of the information contained in multi-view data and try their best to improve the signal-noise ratio of learned representations. Regarding this, joint methods retain all information during the concatenation, but they ignore the redundancy, which makes it difficult for learners to extract useful information from ponderous representations. Alignment methods lose part of the view-specific information during the alignment and hurt the learned representations. Shared and specific methods not only comprehensively exploit the information of multiple views but also reduce the redundancy and improve the signal-noise ratio of learned representations. Therefore, better representations can be obtained.

In addition, handling noise is an important topic in information theory and machine learning research. Noise also exists in multi-view data. Several studies indicate that the non-realizable case caused by noise is very common in multi-view data [79], [80]. Therefore, further enhancement in performance of shared and specific methods may be achievable by investigating the non-realizable case and modeling the noisy of multi-view data.

### 5 FURTHER INVESTIGATION OF SMDDRL

#### 5.1 Ablation Study

Building on the representation learning and classification Backbone (B), SMDDRL equips three components to enhance its performance, i.e., Similarity (S) and Orthogonality (O) constraints, Deep Metric learning (DM), and Semi-Supervised Learning Framework (SSF). To evaluate the contributions made by these components, we conduct ablation study.
study. Starting from the backbone, we gradually equip it with other components. In the end, we get five competitors: B, B+S, B+SO, B+SO+DM, and SMDDRL. We compare them on WebKB, Noisy MNIST, and BBC datasets with F1-score. Other settings remain unchanged.

Fig.8 illustrates the results. We can see:

(1) The results of B, B+S, B+SO indicate: After equipped with similarity constraint, the performance improved. After adding orthogonality constraint, the performance improved further. Two constraints improve classification performance by eliminating the interference from redundant information. This finding is consistent with the results in [31], [33].

(2) The results of B+SO, B+SO+DM indicate: Deep metric learning improves classification performance. Deep metric learning enhances the discriminability of learned representations, thus boosting the performance.

(3) The results of B+SO+DM, SMDDRL indicate: semi-supervised learning framework contributes significantly to classification, but this contribution diminishes with the increase of labeled training data. A possible reason is deep learning methods overfit badly on small training sets. As labeled data increases, their performances soar.

5.2 Effect of the Improvement in Siamese Network

To investigate the effectiveness of our improvement in contrastive loss, we compare the encoding ability and convergence speed of the original contrastive loss (Eq.9) with that of the improved contrastive loss (Eq.10) on WebKB and Noisy MNIST datasets. We build two Siamese networks, where the only difference is loss function, and train them with the same number of iterations and record the losses.

Fig.9 illustrates the representations and losses obtained from two networks. We can see that: a) The representations learned from the improved loss are preferable than that from the original loss. b) The improved loss converges faster than the original loss. The reason is that we exploit average in the improved loss. Specifically, since the original contrastive loss selects paired samples randomly, the encoding performance of Siamese network can be easily affected by each pair of training data. In contrast, the improved loss exploits the average. Because the average is much more stable than randomly selected samples, the loss becomes stable too. Therefore, the loss converges faster.

This improvement can also be explained by margin theory. Compared with contrastive loss, the improved loss optimizes margin distribution more effectively. Specifically:

The optimization process of Siamese Network can be explained with margin theory. For the sake of simplicity, we take two-dimensional representation space for example. As illustrated in Figure.10, if the centers of two different classes are connected by a straight line, the length of this line should be similar to Margin, which is the hyperparameter of contrastive loss. The classification hyperplane can be denoted as the perpendicular bisector of this straight line. The margin margin of the th example can be represented by the distance between the example and this bisector. The training process of Siamese Network is logically equivalent to the optimization of margin of all training examples.

Our improved loss effectively optimizes margin distribution. Gao and Zhou [81] have proved that: in margin optimization problems, the model’s generalizability could be improved by optimizing margin distribution, i.e., simultaneously maximizing average margin and minimizing margin variance. Different from contrastive loss, which randomly pulls two samples from the same class together, the improved loss estimates the class center with the average of samples from the same class in a mini-batch, and pulls these samples to this center. The introduction of average helps the improved loss get smaller margin variance in each iteration (see Figure 10 (b)), while traditional contrastive loss needs more iterations to achieve the same level of margin variance. Thus, our improved loss converges faster.

In practice, we implement the improved loss via contrastive loss. Specifically, instead of two samples, we input contrastive loss with a sample and an average, which is computed when preparing the mini-batch. If this average is $\mu_{same}$ we set $y_n = 1$, otherwise, we set $y_n = 0$. 

<table>
<thead>
<tr>
<th>Methods</th>
<th>R:EN-FR</th>
<th>R:EN-GR</th>
<th>R:EN-IT</th>
<th>R:EN-SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>VE-CoT</td>
<td>0.60 ± 0.02</td>
<td>0.62 ± 0.02</td>
<td>0.61 ± 0.01</td>
<td>0.59 ± 0.02</td>
</tr>
<tr>
<td>AMGL</td>
<td>0.48 ± 0.06</td>
<td>0.43 ± 0.04</td>
<td>0.49 ± 0.05</td>
<td>0.45 ± 0.07</td>
</tr>
<tr>
<td>MLAN</td>
<td>0.75 ± 0.03</td>
<td>0.74 ± 0.03</td>
<td>0.77 ± 0.02</td>
<td>0.74 ± 0.03</td>
</tr>
<tr>
<td>DCCAE</td>
<td>0.80 ± 0.01</td>
<td>0.82 ± 0.01</td>
<td>0.83 ± 0.01</td>
<td>0.81 ± 0.01</td>
</tr>
<tr>
<td>MDNN</td>
<td>0.82 ± 0.01</td>
<td>0.84 ± 0.01</td>
<td>0.84 ± 0.01</td>
<td>0.80 ± 0.01</td>
</tr>
<tr>
<td>SMDFL</td>
<td>0.88 ± 0.01</td>
<td>0.90 ± 0.01</td>
<td>0.90 ± 0.01</td>
<td>0.85 ± 0.01</td>
</tr>
<tr>
<td>Co-L</td>
<td>0.86 ± 0.02</td>
<td>0.85 ± 0.01</td>
<td>0.88 ± 0.02</td>
<td>0.85 ± 0.01</td>
</tr>
<tr>
<td>SMDDRL</td>
<td>0.94 ± 0.01</td>
<td>0.93 ± 0.01</td>
<td>0.94 ± 0.01</td>
<td>0.93 ± 0.01</td>
</tr>
</tbody>
</table>

Fig.8: Ablation study on WebKB, Noisy MNIST, and BBC.
Hyperplane
Hyperplane
Hyperplane
Hyperplane
Margin
Margin
Margin
Margin

Fig. 9: Comparison of the representations and losses obtained from the original contrastive loss (a, c) and the improved contrastive loss (b, d) on WebKB and Noisy MNIST. The Margin is set as 3.0 for WebKB, and 50.0 for Noisy MNIST.

Fig. 10: Explanation of our improvement in Siamese Network from the perspective of margin theory. (a) Improved loss pulls samples from the same class to their class center, while traditional contrastive loss randomly pulls two samples from the same class together. (b) Margin distribution comparison between our improved loss and contrastive loss.

TABLE 4: Effectiveness of different alignment strategies.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Alignment strategies (Average F1-score</th>
<th>Training time : Second)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CCA</td>
<td>PRL</td>
</tr>
<tr>
<td>AD</td>
<td>0.91</td>
<td>0.95</td>
</tr>
<tr>
<td>NUS-Object</td>
<td>0.15</td>
<td>0.19</td>
</tr>
<tr>
<td>BBC</td>
<td>0.95</td>
<td>0.90</td>
</tr>
</tbody>
</table>

5.3 Impact of Different Alignment Strategies

SMDDRL aligns shared representations before fusing them into the final representation. To investigate the effectiveness of different aligners, we compare the following ones:

- CCA: The aligner that maximizes the correlation between shared representations.
- Pairwise Ranking Loss (PRL): A similarity-based aligner that measures the similarity between shared representations with dot product [22].
- Adversarial Similarity Constraint (ASC): Similarity constraint by exploiting adversarial training.
- PRL+ASC: Combination of PRL and ASC.

We compare the F1-score of these strategies on AD (3 Views), NUS-Object (5 Views), and BBC (2 Views) with 50% labeled training data. Training time is also recorded.

Table 4 reports the results. We can see: a) PRL and ASC yield similar F1-score on all datasets. However, with the increase of views, the training time of PRL is much longer than that of ASC. The reason is PRL (CCA) needs to calculate the similarity (correlation) between each pair of views, while ASC calculates once for all views. b) Using both PRL and ASC improves the performance.

5.4 Impact of Different SSL Strategies

To investigate the impact of different SSL strategies on SMDDRL, we compare the following SSL strategies:

- Manifold Tangent Classifier (MTC): SSL Strategy based on manifold smoothness regularization [59].
- Virtual Adversarial Training (VAT): Local smoothness regularization with virtual adversarial perturbation [60].
- Temporal Ensembling (TE): Pseudo-label style strategy that ensembles results from different epochs [51].
- DM+DC: Pseudo-label style strategy that combines deep metric learning with density clustering.

TABLE 5: Effectiveness of different SSL strategies.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Semi-supervised learning strategies (F1-score)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MTC</td>
</tr>
<tr>
<td>WebKB</td>
<td>0.93 ± 0.02</td>
</tr>
<tr>
<td>Noisy MNIST</td>
<td>0.95 ± 0.01</td>
</tr>
<tr>
<td>BBC</td>
<td>0.94 ± 0.02</td>
</tr>
</tbody>
</table>

We compare the F1-score of these strategies on WebKB, Noisy MNIST, and BBC with 50% labeled training data. Table 5 reports the results. We can see:

(1) Those pseudo-label style strategies get better performances than regularization-style strategies. The possible reason is that pseudo-label style strategies can directly provide more labeled data, which is preferred by deep learning methods.

(2) Our SSL strategy achieves similar results to other strategies. The reasons are three-fold: Firstly, our strategy follows the clustering assumption. We assume decision boundaries should lie in low-density regions, and exploit deep metric learning to expand the decision boundaries to improve generalizability. Secondly, to improve the accuracy of labeling, we recheck the new labeled data with the trained classifier. Thirdly, we apply $L_2$ regularization and dropout to regularize this strategy. These regularizers further ensure the generalizability of our approach.
5.5 Influence of Trade-off Parameters $\lambda_1$ and $\lambda_2$

This experiment investigates the impact of trade-off parameters $\lambda_1$ and $\lambda_2$. We conduct Grid Search to find the best trade-off parameters with 40% of labeled data. We repeat each execution 20 times and record the average F1-score.

Fig. 11 illustrates the results on WebKB, Noisy MNIST, and BBC. As shown in Fig. 11, the optimal value of $\lambda_1$ and $\lambda_2$ for WebKB are 0.45 and 0.85, for Noisy MNIST are 0.25 and 0.75, for BBC are 0.75 and 0.45. Also, the optimal value we found on AD are 0.65 and 1.15, on NUS-Object are 0.25 and 0.65, on Reuters are 0.35 and 0.95.

5.6 Training Time and Convergence

In this section, we report the training time and convergence behavior of SMDDRL. Table 6 reports the training time of SMDDRL and other baselines with 50% labeled samples. Data is collected from a computer with an Intel i7 quadcore 3.6GHz CPU, two NVIDIA GTX1080Ti GPUs, and 16GB memory. We can observe the training time of SMDDRL is acceptable. SMDDRL and SMDFL consumed similar training time, which because they are all optimized alternatively. Besides, in the testing phase, the test time for each observation is less than 0.1 seconds. In practice, the training phase is usually off-line, thus SMDDRL is capable for practical use.

In order to verify the convergence behavior of SMDDRL, we present its convergence curves on all datasets when label rate is 50%. As seen from Fig. 12, the objective values are non-dramatically-increasing during the iterations and converge to a fixed value. Additionally, the small fluctuations indicate the alternative optimization finishes clustering and starts a new iteration. Therefore, our SMDDRL scales well in practice because of the fast convergence speed.

6 Conclusion and Future Work

Multi-view representation learning is crucial task in various real-world applications. In this paper, we introduce a novel Semi-supervised Multi-view Deep Discriminant Representation Learning (SMDDRL) approach. By employing shared and specific multi-view deep representation learning network as well as designing orthogonality and adversarial similarity constraints for it, SMDDRL not only comprehensively exploits the consensus and complementary properties of multi-view data but also reduces the redundancy of learned representations. By designing the deep metric learning and density clustering based semi-supervised learning.
framework, SMDDRL effectively exploits the unlabeled data to enhance its representation learning performance. Experimental comparisons on three typical multi-view learning tasks indicate SMDDRL outperforms state-of-the-art multi-view representation learning methods. Results of investigation experiments demonstrate the superiority and robustness of SMDDRL on multi-view representation learning.

In the future, building on the foundation of modeling shared and specific information of each view, we intend to introduce the theory of minds into multi-view learning to accomplish communication between views, such that more information implied in each view can be mined, and a higher level of machine intelligence can be achieved. In addition, we also intend to study the non-realizable case under multi-view setting and propose methods that can model the noise of multi-view data.

ACKNOWLEDGMENTS

We want to thank the editors and reviewers for helpful comments and suggestions. This work was supported by the National Science Foundation of China (NSFC) Key Project (61933013, 61732006), the NSFC-Key Project of General Technology Fundamental Research United Fund (U1736211), the Natural Science Foundation of Guangdong Province (2019A1515011076), and the Key Project of Natural Science Foundation of Hubei Province (2018CFA024).

REFERENCES


Xiaodong Jia received the B.S. degree of computer science and technology from Minzu University of China, Beijing, China, in 2017. He is currently pursing the Ph.D. degree at the School of Computer Science, Wuhan University. His research interests include multi-view learning and deep learning.

Xiao-Yuan Jing received the Ph.D. degree in Pattern Recognition and Intelligent System from the Nanjing University of Science and Technology, 1998. He is a Professor with the School of Computer, Wuhan University, China. His research interests include pattern recognition, machine learning, image processing and artificial intelligence. He has published more than 100 scientific papers, such as TPAMI, TIP, TCB, TIFS, TNLS, TMM, TCSVT, CVPR, AAAI, IJCAI, WWW and PR.

Xiaoke Zhu received the Ph.D. degree from the Wuhan University, Wuhan, China, in 2017. Now he is an Associate Professor with School of Computer and Information, Henan University. His research interests include pattern recognition and machine learning. He has published more than 30 scientific papers, such as TIP, TIFS, TCSVT, CVPR and AAAI.
Songcan Chen received the Ph.D. degree in communication and information systems from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 1997. He is a Professor with the Department of Computer Science and Technology, NUAA. His research interests include pattern recognition, machine learning, and neural computing. He has published more than 100 scientific papers, such as TPAMI, TKDE, KDD, TNNLS and PR.

Bo Du received the Ph.D. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2010. He is currently a Professor with the School of Computer, Wuhan University. His research interests include pattern recognition, hyperspectral image processing, and signal processing. He has published over 100 scientific papers, such as IEEE T GEOSCI REMOTE, TIP, TCB, AAAI, and IJCAI.

Ziyun Cai received the Ph.D. degree from the University of Sheffield, U.K., in 2017. He is a Lecturer with the College of Automation, Nanjing University of Posts and Telecommunications. His research interest is computer vision. He has published papers at several international journals, such as TIP and Information Science.

Zhenyu He received the Ph.D. degree from the Hong Kong Baptist University, Hong Kong, in 2007. He is an Associated Professor with the School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China. His research interests include sparse representation, pattern recognition, and computer vision. He has published more than 70 scientific papers, such as IJCV, TIP, TCB, and TCSVT.

Dong Yue received the Ph.D. degree from the South China University of Technology, in 1995. He is a Professor and the Dean with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China, and also a Changjiang Professor with the Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan, China. His research interests include intelligent system and big data computing.