Progressive Deep Multi-View Comprehensive Representation Learning

Cai Xu[1], Wei Zhao[1]*, Jinglong Zhao[1], Ziyu Guan[1], Yaming Yang[1], Long Chen[2], Xiangyu Song[3]

[1] School of Computer Science and Technology, Xidian University, China
[2] Xi'an University of Posts and Telecommunications, China
[3] Swinburne University of Technology, Melbourne, Australia
{cxu@, ywzhao@mail., jinglong@stu., zyguan@, yym@}xidian.edu.cn, chenlong@xupt.edu.cn, xsong@swin.edu.au

Abstract

Multi-view Comprehensive Representation Learning (MCRL) aims to synthesize information from multiple views to learn comprehensive representations of data items. Prevalent deep MCRL methods typically concatenate synergistic view-specific representations or average aligned view-specific representations in the fusion stage. However, the performance of synergistic fusion methods inevitably degenerate or even fail when partial views are missing in real-world applications; the aligned based fusion methods usually cannot fully exploit the complementarity of multi-view data. To eliminate all these drawbacks, in this work we present a Progressive Deep Multi-view Fusion (PDMF) method. Considering the multi-view comprehensive representation should contain complete information and the view-specific data contain partial information, we deem that it is unstable to directly learn the mapping from partial information to complete information. Hence, PDMF employs a progressive learning strategy, which contains the pre-training and fine-tuning stages. In the pre-training stage, PDMF decodes the auxiliary comprehensive representation to the view-specific data. It also captures the consistency and complementarity by learning the relations between the dimensions of the auxiliary comprehensive representation and all views. In the fine-tuning stage, PDMF learns the mapping from the original data to the comprehensive representation with the help of the auxiliary comprehensive representation and relations. Experiments conducted on a synthetic toy dataset and 4 real-world datasets show that PDMF outperforms state-of-the-art baseline methods. The code is released at https://github.com/winterant/PDMF.

Introduction

Many real-world applications involve multiple views. For example, doctors diagnose the disease of a patient according to his the blood test, the radiology test, etc. These views often exhibit consistent and complementary information of the same data. Synthesizing multi-view data could boost the performance of many tasks. In recent years, the rapid growth of deep learning researches leads to many popular deep multi-view learning research topics, such as deep multi-view clustering (Wen et al. 2022, 2020b,a), trusted deep multi-view classification (Han et al. 2021; Xu et al. 2022), deep multi-view contrastive learning (Lin et al. 2021; Li et al. 2023). This work is concerned with the fundamental problem of most deep multi-view learning methods, the deep Multi-view Comprehensive Representation Learning (MCRL) problem, which aims to *fuse the consistent and complementary information of all views to obtain the comprehensive representation*.

Prevalent deep MCRL methods can be roughly divided into Multi-View Synergistic Representation Learning (MSRL) and Multi-View Aligned Representation Learning (MARL). MSRL usually first learns separate view-specific representations and synergizes them through some criteria, then concatenates them to construct the multi-view comprehensive representation (Vendrov et al. 2016). However, the performances of most MSRL methods inevitably degenerate or even fail when partial views are missing for some instances in real-world applications. Another line is MARL, which usually first learns aligned view-specific representations of all views, then (weighted) averages them to construct multi-view comprehensive representation (Wen et al. 2020b). In general, deep MARL methods use the Deep Neural Networks (DNNs) to model the mappings from view-specific data to aligned view-specific representations. However, most existing deep MARL methods cannot fully exploit the complementarity of multi-view data, i.e., some views may contain information that other views do not have. For example, in Figure 1, the image view contains the lesion size and location information, but can not provide the content of leucocyte. Hence the dimensions of the comprehensive representation reflecting the content of leucocyte should not connect with the image view.

Recently, in order to explicitly consider the complementarity of multi-view data, researchers propose some new deep MCRL paradigms. Wang *et al.* (Wang et al. 2020) propose a novel dimension exchange strategy for MARL. They regard a dimension of the aligned view-specific representations as the complementary dimension if it is less than a preset threshold. The complementary dimensions would be replaced by the average of the corresponding dimensions of other views. However, this method may unstable since an inferior initialization would extremely influent the model training. In (Nagrani et al. 2021), Nagrani *et al.* propose a bottleneck-based fusion method, which can be deem as the

*Corresponding author

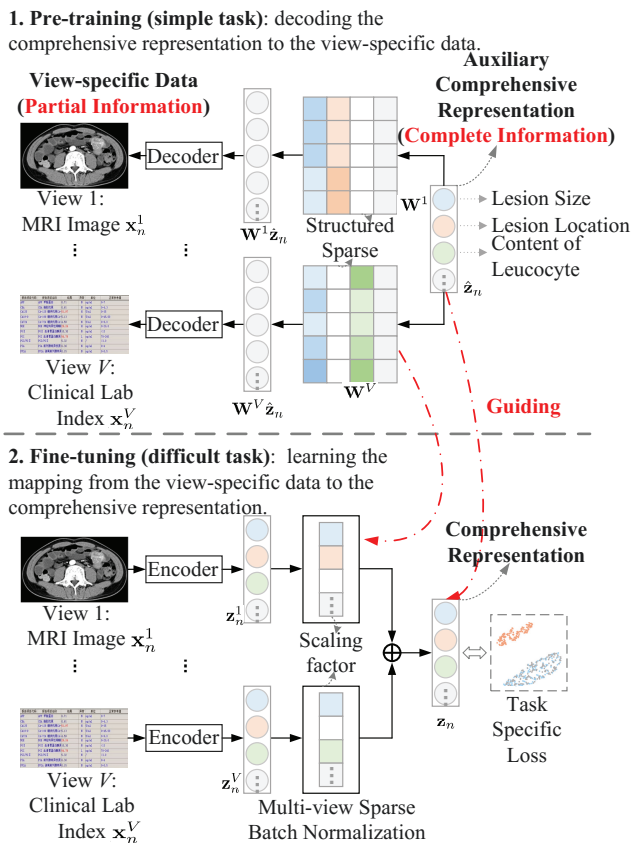


Figure 1: Illustration of PMDF. In the pre-training stage, we perform a simple task, i.e., learning the mappings from the auxiliary multi-view comprehensive representation $\hat{\mathbf{z}}_n$ (*complete information*) to view-specific data $\mathbf{x}_n^v$ (*partial information*). We also establish the view-specific relation matrices $\{\mathbf{W}^v\}_{v=1}^V$ to explicitly model the consistency and complementarity among different views. In the fine-turning stage, we learn the mappings from $\mathbf{x}_n^v$ to $\mathbf{z}_n$ under the guidance of $\{\mathbf{W}^v\}_{v=1}^V$ and $\hat{\mathbf{z}}_n$.

integration of MSRL and MARL. It divides the multi-view comprehensive representation into the aligned part (bottleneck) and view-specific parts, which are forced to collect the consistent information and complementary information, respectively. However, this paradigm is concentrated on the two-view case because the model complexity exponentially increases with the increase of views.

To eliminate the above limitations and drawbacks, we propose a new MARL method, named Progressive Deep Multi-view Fusion (PDMF). Considering the multi-view comprehensive representation $\mathbf{z}$ should contain complete information and the view-specific data $\mathbf{x}^v$ contain partial information, we deem that it is difficult to directly learn the mapping from partial information to complete information. Inspired by human learning process, we propose a progressive learning strategy. As shown in Figure 1, in the pre-training stage, PDMF aims to solve a simple task, i.e., decoding the auxiliary comprehensive representation $\hat{\mathbf{z}}$ to view-specific data $\mathbf{x}^v$. PDMF establishes view-specific relation matrix $\mathbf{W}^v$ to capture the relation of $\hat{\mathbf{z}}$ and $\mathbf{x}^v$. PDMF also requires $\mathbf{W}^v$ to be sparse in term of columns to explicitly model the consistency and complementarity among different views. In the fine-tuning stage, PDMF learns the mapping from $\mathbf{x}^v$ to $\mathbf{z}$. A Multi-view Sparse Batch Normalization (MSBN) layer is established to fuse the aligned view-specific representations. The sparse scaling factors of MSBN contain the dimension-specific correlations of the aligned view-specific representations and $\mathbf{z}$. We also use the $\mathbf{W}^v$ and $\hat{\mathbf{z}}$ to guild the learning of sparse scaling factors and $\mathbf{z}$.

The contributions of this work are as follows: (1) we propose a progressive multi-view comprehensive learning strategy to explicitly consider the consistency and complementarity of multi-view data; (2) we develop the general MSBN layer to fuse the aligned view-specific representations, which could be used in most prevalent MARL methods. This layer facilitates the fusion model to integrate inter-view information and reserve intra-view information in a flexible fashion; (3) we empirically evaluate PDMF on a synthetic toy dataset and 4 real-world datasets to show its superiority over state-of-the-art baseline methods.

Related Work

In this section, we brief review three lines of related works about MCRL, Multi-view Synergistic Representation Learning (MSRL), Multi-view Aligned Representation Learning (MARL) and Multi-modal Pre-training. We also show the structure comparisons of MSRL, MARL and the proposed PDMF in Fig. 2.

Multi-View Synergistic Representation Learning

Deep MSRL first learns view-specific representations $\{\mathbf{z}_s^v\}_{v=1}^V$ of multi-view data $\{\mathbf{x}^v\}_{v=1}^V$ with separate DNNs $\{f_s^v\}_{v=1}^V$. MSRL also synergizes $\{\mathbf{z}_s^v\}_{v=1}^V$ through some criteria, such as maximizing correlation (Andrew et al. 2013), enforcing a partial order (Vendrov et al. 2016) between view-specific representations and maximizing reconstruction accuracy (Wan et al. 2021; Radford et al. 2021). Then, MSRL constructs the multi-view synergistic representation $\mathbf{z}_s$ by the aggregation network f_s, such as concatenation (Andrew et al. 2013; Zeng et al. 2019). The whole process is shown in:

$$\mathbf{z}_s^v = f_s^v(\mathbf{x}^v), \mathbf{z}_s = f_s(\mathbf{z}_s^1, \cdots, \mathbf{z}_s^V). \tag{1}$$

The pioneer deep MSRL method is Deep Canonical Correlation Analysis (DCCA) (Andrew et al. 2013). DCCA learns two separate DNNs for two views, with the objective that the learned high-level view-specific representations are as correlated as possible. Hazirbas *et al.* (Hazirbas et al. 2016) point out the performance of MSRL is highly affected by the choice of which layer to fuse. Vendrov *et al.* (Vendrov et al. 2016) propose to capture a partial order of the representations of text and image views, i.e., enforcing a hierarchy on the representation. For example, the semantic of an image ("woman walking her dog") is hierarchical and transitive: "woman walking her dog" $\rightarrow$ "woman walking" $\rightarrow$ "woman" $\rightarrow$ "person". Xu *et al.* (Xu et al. 2020) establish

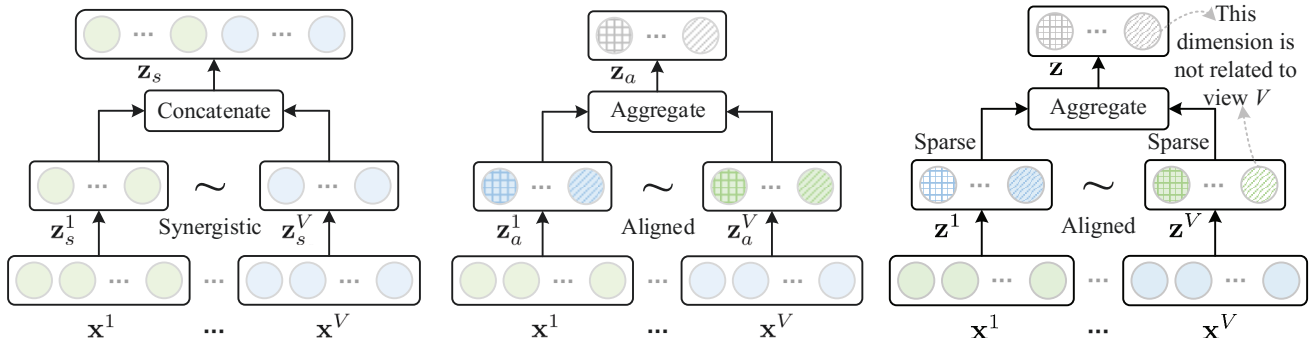


Figure 2: Structures of *Multi-view Synergistic Representation Learning* (*MSRL*), *Multi-view Aligned Representation Learning* (*MARL*) and the proposed *PDMF*. MSRL learns separate view-specific representations and synergizes them through some criteria. MARL aligns view-specific representations. PDMF explicitly establishes the sparse connections between the multi-view comprehensive representation and the aligned view-specific representations. Therefore, the dimensions of the multi-view comprehensive representation can flexibly connect to all view. This complies with the complementarity of multi-view data.

multi-view interactive maps through the cross-correlations of $\{\mathbf{z}_s^v\}_{v=1}^V$. Then they impose supervision on the interactive maps. Recently, Wan *et al.* (Wan et al. 2021) require the view-specific representations should reconstruct the transformation relationship in original data. Radford *et al.* (Radford et al. 2021) reconstruct some parts of one view by the representation of another view, which can mine the cross-view interaction in a self-supervised manner.

Unfortunately, MSRL usually assumes that all of the views are complete. However, in real-life cases some views could be missing for some data instances. Due to the concatenation based fusion paradigm, MSRL can not well applied in this case. Our PDMF can naturally solve this problem since it aligns view-specific representations.

Multi-View Aligned Representation Learning

MARL requires the view-specific representations $\{\mathbf{z}_a^v\}_{v=1}^V$ are aligned. In the fusion stage, MARL (weighted) averages [1] $\{\mathbf{z}_a^v\}_{v=1}^V$ to learn multi-view aligned representation $\mathbf{z}_a$. The whole process is shown in:

$$\mathbf{z}_a^v = f_a^v(\mathbf{x}^v), \mathbf{z}_a = \sum_{v=1}^{V} w^v \mathbf{z}_a^v, \tag{2}$$

where w^v is the weight of the v-th view. Representative earlier MARL method is Bimodal Deep Autoencoder (Ngiam et al. 2011), which extracts aligned representations by training a bimodal deep autoencoder. Srivastava and Salakhutdinov (Srivastava and Salakhutdinov 2012) propose to learn aligned representation of images and texts by Deep Boltzmann Machines. These methods require the view-specific representations are fully aligned ($\{w^v = 1\}_{v=1}^V$) (Zhang et al. 2021, 2022). There are also some works directly impose aligned constraints to $\{\mathbf{z}_a^v\}_{v=1}^V$, such as Maximum-Mean-Discrepancy (Gretton et al. 2012), Minimizing Distribution Divergence (Wen et al. 2020b; Xu et al. 2019).

However, most existing MARL methods cannot fully exploit the complementarity of multi-view data, i.e., some views may contain information that other views do not have. Therefore, some dimensions of the aligned view-specific representation should not relate to the all views. Some works (Nagrani et al. 2021) provide a way to alleviate this issue, which learns multi-view aligned representation (consistent information) while simultaneously maintaining view-specific representations (complementary information). But this paradigm is concentrated on the two-view case. Wang *et al.* (Wang et al. 2020) regard a dimension of $\{\mathbf{z}_a^v\}_{v=1}^V$ as the complementary dimension if it is less than a preset threshold. Then they replace it by the average of the corresponding dimensions of other views. However, the method may unstable in the training process since a inferior initialization would extremely influent the model training. Compared to this method, PDMF establishes a easier pre-training task to better initialize the consistency and complementary dimensions.

Multi-Modal Pre-training

This work is also related to Multi-modal[2] Pre-training (Radford et al. 2021), which also uses the progressive learning strategy. In the pre-training stage, Multi-modal Pre-training aims to pre-train multi-modal transformers with less high-quality labelled data. It extracts region features (Anderson et al. 2018), CNN-based grid features or patch features for vision modal and word embedding for language model (Kenton and Toutanova 2019). Then it sends the vision or language modal features to transformers. Finally, it

[1] There are some other strategies to fuse $\{\mathbf{z}_a^v\}_{v=1}^V$, such as element-wise multiplication. In this work, we focus on the representative (weighted) averages fusion strategy.

[2] Modalities usually refer to different information sources of the same object, such as text, image, video, etc. View is a broader concept, such as various low-level features of an image. Multiple views of an instance describe same object, while this relation may not existed in some multi-modal datasets.

uses different pre-training objectives to train the transformers, such as reconstructing the masked element by leverage the unmasked remainders. The pre-trained transformers can achieve surprising effectiveness by fine-tuning with only a tiny amount of manually-labelled data on downstream tasks. PDMF is superior to Multi-modal Pre-training from two perspectives: 1) Most Multi-modal Pre-training methods can only deal two specific modals (Chen et al. 2022), i.e., vision modal (image or video) and language modal data. PDMF can handle various kinds of multi-view data with arbitrary views; 2) Multi-modal Pre-training requires large scale training data, while PDMF applies to different data sizes.

The Method

In this section, we present Progressive Deep Multi-view Fusion (PDMF) in detail.

Notations and Problem Statement

In the MCRL problem, an instance is characterized by multiple views. Suppose we are given a multi-view dataset with V views and N instances. We use $\mathbf{x}_n^v \in \mathbb{R}^{D^v}$ to denote the feature vector for the v-th view $(v = 1, \ldots, V)$ of the n-th instance $(n = 1, \ldots, N)$, where D^v is the dimensionality of the v-th view. y_n is the class label for the n-th instance. The multi-view data contain consistency and complementarity properties: consistency denotes different views exhibit common information and complementarity implies that some views may contain knowledge that other views do not have. Our target is to learn the multi-view comprehensive representation $\mathbf{z}_n \in \mathbb{R}^D$, which should contain the consistent and complementary information of $\{\mathbf{x}_n^v\}_{v=1}^V$.

Framework

Considering the multi-view comprehensive representation ($\mathbf{z}_n$) should contain complete information and the view-specific data ($\mathbf{x}_n^v$) contain partial information, we deem that it is difficult to directly learn the mapping from $\mathbf{x}_n^v$ to $\mathbf{z}_n$. Motivated by the curriculum learning (Bengio et al. 2009), which solves easier easier subtasks first, and then increase the difficulty level, we propose a progressive learning method, PDMF. As shown in Figure 1, PDMF contains pre-training (simple task) and fine-tuning (difficult task) stages. In the pre-training stage, PDMF decodes the auxiliary comprehensive representation $\hat{\mathbf{z}}_n$ to view-specific data $\mathbf{x}_n^v$. In addition, PDMF explicitly learns the relations of $\hat{\mathbf{z}}_n$ and $\{\mathbf{x}_n^v\}_{v=1}^V$ in the relation matrices, $\{\mathbf{W}^v\}_{v=1}^V$. In the fine-tuning stage, PDMF trains view-specific encoders to learn aligned view-specific representations $\{\mathbf{z}_n^v\}_{v=1}^V$. Considering $\mathbf{z}_n^v$ contains partial information while $\mathbf{z}_n$ should contain complete information, we deem some dimensions of $\mathbf{z}_n^v$ should not related to $\mathbf{z}_n$. This also complies with the complementarity of multi-view data. Therefore, we elaborate a Multi-view Sparse Batch Normalization (MSBN) layer to fuse $\{\mathbf{z}_n^v\}_{v=1}^V$. The sparse scaling factors of MSBN model the relations of $\mathbf{z}_n^v$ and $\mathbf{z}_n$. We also use the relation matrices and the auxiliary comprehensive representation $\hat{\mathbf{z}}_n$ to guild the learning of sparse scaling factors and $\mathbf{z}_n$. Details regarding each component will be elaborated as below.

Pre-training In the pre-training stage, we aim to solve a simple task, i.e., decoding the auxiliary comprehensive representation $\hat{\mathbf{z}}_n$ (which should contain complete information) to view-specific data $\mathbf{x}_n^v$ (containing partial information). The decoding process contains two parts: 1) the relation matrices $\{\mathbf{W}^v\}_{v=1}^V$ which measure the relations of all elements of $\hat{\mathbf{z}}_n$ and each view; 2) the decoder neural networks $\{g^v(\cdot)\}_{v=1}^V$ that learn the complex hierarchical nonlinear mapping of view-specific data $\mathbf{x}_n^v$ and high-level representation $\mathbf{W}^v\hat{\mathbf{z}}_n$. The target of the decoding process is:

$$L_n^r = \sum_{n=1}^{N}\sum_{v=1}^{V} \|\mathbf{x}_n^v - g^v(\mathbf{W}^v\hat{\mathbf{z}}_n)\|_2^2, \quad \hat{\mathbf{z}}_n \in \mathbb{R}^D, \mathbf{W}^v \in \mathbb{R}^{D_W^v \times D}, \tag{3}$$

where L_n^r denotes the reconstruction loss. For each $\mathbf{W}^v$, we add a structured sparseness regularizer on it to encourage some column vectors in $\mathbf{W}^v$ to become 0:

$$\|\mathbf{W}^v\|_{1,\infty} = \sum_{j=1}^{D} \max_{1\le i\le D_W^v} |W_{ij}^v|. \tag{4}$$

This makes the v-th view independent of the latent dimensions corresponding to these zero valued vectors. For example, in Fig. 1 the third column of $\mathbf{W}^1$ are zero columns, which means the third latent dimension of $\hat{\mathbf{z}}_n$ is not associated with view 1. This dimension could represent complementary information in multi-view data. On the contrary, if none of the related vectors of a latent dimension are completely zero, this dimension would correlate with all the views and capture the consistent information across all the views. The correlation setting for each dimension is determined automatically by optimization.

Such an unsupervised framework is hard to guarantee that the learned auxiliary comprehensive representation captures the conceptual structures in multi-view data. Therefore, we introduce the supervision information to the learning process. A directly strategy is to learn a classification function based on $\hat{\mathbf{z}}_n$. However, the generalization ability may be affected since $\hat{\mathbf{z}}_n$ and classifier are jointly learnt, which is likely an under-constrained problem. This strategy may find representation that can well fit the training data but not well reflect the underlying patterns. We establish a clustering-like classification scheme for prediction:

$$\hat{y}_n = \arg\max_{y\in\mathcal{Y}} \frac{1}{|\mathcal{T}(y)|} \sum_{\hat{\mathbf{z}}\in\mathcal{T}(y)} \phi(\hat{\mathbf{z}})^T\phi(\hat{\mathbf{z}}_n), \tag{5}$$

where $\mathcal{Y}$ denotes the whole label set, $\mathcal{T}(y)$ is the set of comprehensive representation $\hat{\mathbf{z}}$ of class y, and $\phi(\cdot)$ denotes the feature mapping function for $\hat{\mathbf{z}}$. We set $\phi(\hat{\mathbf{z}}) = \hat{\mathbf{z}}$ for simplicity and effectiveness in implementation. By jointly considering classification and conceptual structures learning, the clustering-like classification loss is specified as:

$$L_n^c = \max\left\{0, \Delta(y_n, \hat{y}_n) + \frac{1}{|\mathcal{T}(\hat{y}_n))|} \sum_{\hat{\mathbf{z}}\in\mathcal{T}(\hat{y}_n)} \phi(\hat{\mathbf{z}})^T\phi(\hat{\mathbf{z}}_n) - \frac{1}{|\mathcal{T}(y_n))|} \sum_{\hat{\mathbf{z}}\in\mathcal{T}(y_n)} \phi(\hat{\mathbf{z}})^T\phi(\hat{\mathbf{z}}_n)\right\}. \tag{6}$$

where $\Delta(y_n, \hat{y}_n) = 0$ when $\hat{y}_n = y_n$, otherwise, $\Delta(y_n, \hat{y}_n) = 1$. The clustering-like classification loss not only penalizes the misclassification but also ensures structured representation. If the n-th instance is correctly classified, i.e., $\hat{y}_n = y_n$, the loss degrades to normal $\Delta(y_n, \hat{y}_n)$. Otherwise, the last two terms of L_n^c enforce the similarity between $\hat{\mathbf{z}}$ and the center of class y_n larger than that between $\hat{\mathbf{z}}$ and the center of class $\hat{y}_n$ with a margin $\Delta(y_n, \hat{y}_n)$. Therefore, this loss keeps the instances from belonging to the same class are near each other in the comprehensive representation, while keeping instances from different classes as distant as possible.

By synthesizing the above objectives, the overall optimization problem in the pre-training stage is formulated as:

$$\min_{\{\mathbf{W}^v, g^v\}_{v=1}^V, \{\hat{\mathbf{z}}_n\}_{n=1}^N} \sum_{n=1}^{N} (L_n^r + \alpha_1 L_n^c) + \alpha_2 \sum_{v=1}^{V} \|\mathbf{W}^v\|_{1,\infty}, \tag{7}$$

where $\alpha_1, \alpha_2 > 0$ are hyper-parameters.

The pre-training model learns the comprehensive representation $\{\hat{\mathbf{z}}_n\}_{n=1}^N$ of the training set. However, it faces out-of-the-sample problem since an additional optimization problem should be established for new data. Therefore, we introduce the fine-tuning stage to solve this problem.

Fine-tuning In the fine-turning stage, we aim to map the view-specific data $\mathbf{x}_n^v$ (containing partial information) to the multi-view comprehensive representation $\mathbf{z}_n$ (which should contain complete information). This is difficult especially when the views are highly independent. Therefore we involve the auxiliary $\hat{\mathbf{z}}_n$ and relation matrices $\{\mathbf{W}^v\}_{v=1}^V$ in the learning process.

Firstly, we establish encoders $\{f^v(\cdot)\}_{v=1}^V$ to learn aligned view-specific representations:

$$\mathbf{z}_n^v = f^v(\mathbf{x}_n^v). \tag{8}$$

In order to align the representation of multiple views, we add view-specific Batch Normalization (BN) in each layer. BN whitens activations within a mini-batch of N^b instances for each dimension and further transforms the whitened activations using affine parameters γ and β. Therefore, BN is conducive to eliminate covariate shift among different views. We denote the c-th dimension of output of the l-th layer as $z_{n,l,c}^v$, where $c \in (1, \cdots, D_l^v)$. The BN process is:

$$z_{n,l+1,c}^v = \gamma_{l,c}^v \frac{z_{n,l,c}^v - \mu_l^v}{\sqrt{{\sigma_l^v}^2 + \epsilon}} + \beta_{l,c}^v, \tag{9}$$

where $\gamma_{l,c}^v$ and $\beta_{l,c}^v$ are the trainable scaling factor and offset, respectively. ϵ is a small constant to avoid zero value. μ_l^v and σ_l^v denote the mean and the standard deviation, respectively, of all representation dimensions for the current mini-batch data:

$$\mu_l^v = \frac{\sum_n \sum_c z_{n,l,c}^v}{N^b \cdot D_l^v}, \tag{10a}$$

$${\sigma_l^v}^2 = \frac{\sum_n \sum_c (z_{n,l,c}^v - \mu_l^v)^2}{N^b \cdot D_l^v}. \tag{10b}$$

Considering the complementarity of multi-view data, some dimensions of the L^v-th (highest) layer view-specific representation $\mathbf{z}_n^v$ should not related to the multi-view comprehensive representation $\mathbf{z}_n$. We elaborate a Multi-view Sparse Batch Normalization (MSBN) layer to model this sparse connection. The sparsity constraint is imposed on the scaling factors $\{\gamma_{L^v,c}^v\}_{v=1}^V$ of the MSBN layer. $\gamma_{L^v,c}^v$ evaluates the correlation between the input $z_{n,L^v-1,c}^v$ and the output $z_{n,L,c}^v$. When $\gamma_{L^v,c}^v \to 0$, $z_{n,L^v,c}^v$ will lose its influence to the v-th view. In addition, the sparsity constraint causes that once $\gamma_{L^v,c}^v \to 0$ at a certain training step, it will almost do henceforth.

To better initialize the learning process, we involve relation matrices $\{\mathbf{W}^v\}_{v=1}^V$ to guide the learning. Specifically, we extract the relation of $\mathbf{z}_n$ and view v from $\mathbf{W}^v$:

$$\bar{w}_c^v = \sum_{j=1}^{D_W^v} \frac{1}{D_W^v} |W_{jc}^v|. \tag{11}$$

By minimizing $\|\gamma_{L^v,c}^v - \sigma(a\bar{w}_c^v + b)\|_2^2$, we achieve a better $\gamma_{L^v,c}^v$ by the assistant of $\bar{w}_c^v$. a, b are trainable parameters. $\sigma(\cdot)$ is the sigmoid function which forces $\gamma_{L^v,c}^v$ to be in (0,1) and facilitates more sparse values.

Then, since view-specific representations are well aligned through the above strategies, we directly average them to obtain the multi-view comprehensive representation:

$$\mathbf{z}_n = \frac{1}{V} \sum_{v=1}^{V} \mathbf{z}_n^v. \tag{12}$$

We also use auxiliary comprehensive representation $\hat{\mathbf{z}}_n$ to guild the learning of $\mathbf{z}_n$. The whole auxiliary loss is defined as:

$$L_n^a = \|\mathbf{z}_n - \hat{\mathbf{z}}_n\|_2^2 + \frac{\zeta}{V \cdot D} \sum_{v,c} \|\gamma_{L^v,c}^v - \sigma(a\bar{w}_c^v + b)\|_2^2), \tag{13}$$

where $\zeta > 0$ is hyper-parameter. The overall optimization problem in the fine-turning stage is summarized as:

$$\min_{\{f^v\}_{v=1}^V} \frac{1}{N^b} \sum_n (\lambda_1 L^t(\mathbf{z}_n) + \delta_t L_n^a) + \lambda_2 \sum_{v,c} |\gamma_{L^v,c}^v|, \tag{14}$$

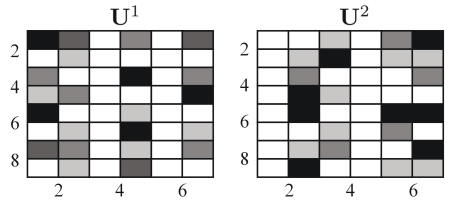
where the first term is the task specific loss. λ_1, λ_2 is hyper-parameter. $\delta_t = \max(0, 1 - t/10)$ is the annealing coefficient, t is the index of the current training epoch.

Experiments

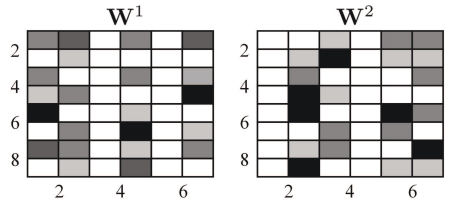
We evaluate the performance of PDMF on a synthetic toy dataset and four real-world datasets.

A Toy Example

We first evaluate PDMF on a synthetic toy example to investigate it explicitly models the consistency and complementarity. In the toy example, we generate multi-view data from underlying comprehensive representation and force some dimensions of the representation are not related to partial views. Specifically, the toy dataset consists of 2 views

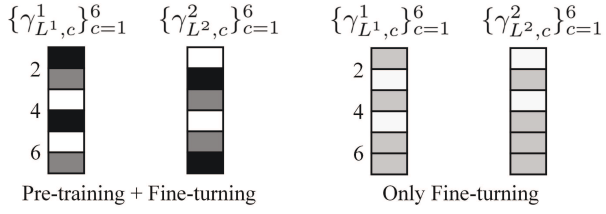


(a) Mapping matrices $\mathbf{U}^v$



(b) Relation matrices $\mathbf{W}^v$

Figure 3: The correlations of view-specific data and multi-view comprehensive representation in the toy example.



(a) (b)

Figure 4: The learnt scaling factors on the toy example.

of 1000 data instances $\{\mathbf{x}_n^v\}_{n=1}^{1000}$, which belong to 2 categories with 500 data instances in each category. The instances are generated from the multi-view comprehensive representation $\{\mathbf{z}_n^v\}_{n=1}^{1000}$ with 6 dimensions, with 3 for each category. Each element of $\mathbf{z}_n^v$ is the sum of a number sampled from gamma distributed $\Gamma(1, 0.9)$, the noise sampled from Gaussian distribution $N(0, 0.1)$ and the consistent 0.5. We use $\mathbf{x}_n^v = \mathbf{U}^v\mathbf{z}_n^v + \mathbf{p}^v$ to generate data instances, where $\mathbf{U}^v \in \mathbb{R}^{8\times 6}$ and $\mathbf{p}^v \in \mathbb{R}^8$ denote the view-specific mapping matrix and noise, respectively. The elements of $\mathbf{U}^v$ are produced by a uniform distribution $U(0.4, 1)$. We randomly set 30 percent elements to be zero to simulate the real-world multi-view mapping pattern. We also set some columns of $\mathbf{U}^v$ to be 0 to model the complementarity of multi-view data. The elements of $\mathbf{p}^v$ are produced by the Gaussian distributions $N(0, 0.1)$. The generated mapping matrices $\mathbf{U}^v$ are shown in Fig. 3(a).

We perform PDMF on the toy dataset. Figs. 3(b) and 4(a) show the correlation matrices and scaling factors of PDMF, respectively. We can see that the correlation matrices are very similar to the ground truth mapping matrices. In addition, the sparse relations of the comprehensive representation and two views are explicitly captured by the scaling factors. This indicates the consistency and complementarity properties can be recovered by PDMF. We further test the necessity of the pre-training stage. We remove the auxiliary loss in Eq. (14) and directly train this fine-tuning model (PDMF-F). The scaling factors of this model are shown in Fig. 4(b). We can see that all elements of the scaling factors tend to be small values, which is not comply the real sparse relations. The reason might directly solving the hard task, i.e., learning complete information from partial information, is hard to achieve satisfactory performance.

Experiments on Real-World Datasets

Datasets We use 4 real-world datasets to evaluate PDMF:

Handwritten Dataset [3] consists of features of handwritten numbers. It contains 10 categories (handwritten numbers '0'-'9') with 200 images in each category and 6 types of image features, which are used as 6 views in our experiments.

CUB (Caltech-USD Birds) Dataset [4] contains 11788 bird images associated with text descriptions of 200 categories. We use the first 10 categories, and select 60 instances in each category to construct our dataset. We extract text features and visual features as two views.

Scene15 Dataset (Fei-Fei and Perona 2005) contains 4485 images from 15 indoor and outdoor scene categories. Three kinds of features, i.e., 1536D GIST description, 3780D HOG histogram and 4096D LBP feature are extracted as three views.

UCIA (UCI Activity) Dataset [5] is a sequential multi-sensors dataset. It consists of sensor data for 19 different activities such as standing, sitting, etc. It contains 9120 instances with 5 views. The instances contain 9(dimension)*125(timestamps) features for each view.

Important statistics are summarized in Table 1.

Evaluation Methodology We compare PDMF with the following MCRL baselines: **Best Single View (BSV)** classifies on each view, and reports the best result. **Concat** concatenates feature vectors of different views to apply classification. **DCCAE** (Wang et al. 2015) is a representative MSRL method which employs autoencoders and Canonical Correlation constraint to learn the comprehensive representation. Since it is for two views, we run DCCAE on all two-view combinations of a multi-view data set and report the best results. **DIMC** (Wen et al. 2020b) is a representative MARL method which weighted averages the view-specific representations to obtain the comprehensive representation. **CEN** (Wang et al. 2020) is a novel MARL method which uses dimension exchange strategy to learn the complementarity of multi-view data. **AE^2-Nets** (Zhang et al. 2022) is the state-of-the-art MARL method. It establishes the inner-AE-networks to extract view-specific intrinsic information, while the outer-AE-networks to integrate this view-specific intrinsic information from different views into multi-view comprehensive representation.

We compare the classification performance of PDMF and baseline methods. For the multi-view comprehensive rep-

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

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

[5] https://archive.ics.uci.edu/ml/datasets/daily+and+sports+activities

Dataset	Size	# of categories	Dimensionality
Handwritten	2000	10	6/47/64/76/216/240
CUB	600	10	300/1024
Scene15	4485	15	1536/3780/4096
UCIA	9120	19	9*125 for all 5 views

Table 1: Datasets summary.

Method	Handwritten	CUB	Scene15	UCIA
BSV	92.32±0.71	75.37±0.42	47.66±0.27	66.31±0.51
Concat	94.84±0.51	78.34±0.60	48.54±0.64	67.32±0.74
DCCAE	96.74±0.41	81.48±0.45	54.94±0.21	70.65±0.54
DIMC	91.28±0.93	80.39±0.83	58.32±0.47	77.38±0.49
CEN	93.37±1.03	87.42±0.67	64.25±0.59	79.24±0.72
AE^2-Nets	96.65±0.32	85.75±0.59	67.19±0.42	80.06±0.39
PDMF	**98.97±0.31**	**91.26±0.29**	**69.33±0.55**	**83.36±0.47**

Table 2: Classification accuracy on different datasets (accuracy±standard deviations,%).

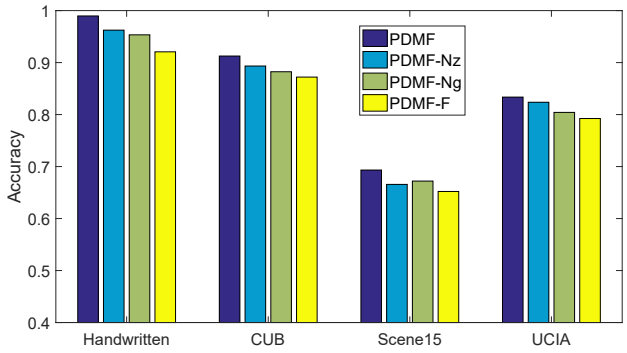
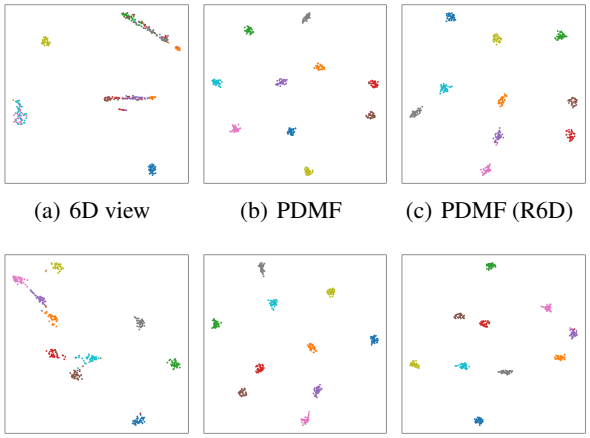


Figure 5: Ablation studies on all datasets.



(a) 6D view (b) PDMF (c) PDMF (R6D)

(d) PDMF-F (e) PDMF-F (R6D) (f) PDMF-P

Figure 6: t-SNE visualization on the Handwritten dataset. PDMF-P and PDMF-F denote the individual Pre-training and Fine-turning stage, respectively. We Remove the 6D view of the Handwritten dataset to construct a varietal dataset (R6D).

resentation learning baselines, we feed the comprehensive representation into the k-Nearest Neighbor (kNN) classifier (k = 9) to calculate Accuracy. Each dataset is randomly divided into training set (80%), validation set (10%) and test set (10%). All the hyperparameters of PDMF and baselines are selected based on the validation set. The averaged performance is reported by running each test case five times. We use stochastic gradient decent and apply Adam for training. The learning rate is set as $1e^{-5}$. We elaborate the implementation details, hyper-parameters setting and network architectures in the supplement materials.

Results Tab. 2 shows the classification performance of PDMF and baseline methods. First, CEN is superior to DIMC on all datasets. This is intuitive since DIMC directly averages the view-specific representations, while CEN models the important sparse connections in the fusion process. Second, DIMC and CEN perform not well on the Handwritten dataset. The reason might this dataset contains 6 dimension morphological features view. It is hard to align this view to the multi-view comprehensive representation. Third, AE^2-Nets performs well on this datasets since it realizes alignment by outer-AE-networks, which learns mapping from comprehensive representation (complete information) to view-specific data (partial information). Finally, PDMF consistently outperforms AE^2-Nets on all datasets, which indicates that the sparse constraints could generate better representation by explicitly leveraging consistency and complementarity of multi-view data. We use t-test with significance level 0.05 to test the significance of performance difference. Results show that PDMF significantly outperforms all the baseline methods.

Analysis *Ablation Studies*: To prove the importance of the pre-training stage, we propose three variations of PDMF, PDMF-Nz, PDMF-Ng and PDMF-F, which remove the first, second and all terms of Eq. (13), respectively. The exper-

iment results are shown in Fig. 5. We can obtain the following points from the experimental results: (1) The performances of variations drop on all datasets, which shows the effectiveness of the pre-training stage; (2) The performance of PDMF-F drops dramatically on the Handwritten dataset. This complies with the experiment results of DIMC and CEN.

t-SNE Visualization: We use t-SNE to visualize the 6D morphological features view and the comprehensive representations. As shown in in Fig.6: (1) the 6D morphological features view has the poor classes separability; (2) as shown in Figs. 6(b)-6(e), we find the performance of PDMF-F immensely promotes when the 6D morphological features view of the Handwritten dataset is removed (expressed as R6D in Fig.6) and the performance of the PDMF is affected slightly. This verifies the pre-training is conductive to alignment in the fine-turning stage, especially when the views are highly diverse; (3) Fig. 6(f) shows the auxiliary comprehensive representation of PDMF-P already reveals compact clusters. This indicates the auxiliary comprehensive representation provides high-quality guidance for the fine-turning stage.

Conclusion

In this paper, we proposed a new MARL method, PDMF, to explicitly capture the consistence and complementarity of multi-view data. In the pre-training stage, PDMF models the relations between the dimensions of the auxiliary comprehensive representation and all views by solving a simple task. In the fine-tuning stage, PDMF learns the mappings from the original data to the comprehensive representation with the help of the auxiliary comprehensive representation and relation matrices. Experimental results on a synthetic toy dataset and 4 real-world datasets confirmed the effectiveness of PDMF.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant Nos. 62133012, 61936006, 62103314, 62203354, 61876144, 61876145, 62073255), the Key Research and Development Program of Shaanxi (Program No. 2020ZDLGY04-07), and Innovation Capability Support Program of Shaanxi (Program No. 2021TD-05), the Open Project of Anhui Provincial Key Laboratory of Multi-modal Cognitive Computation, Anhui University, No. MMC202105.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

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

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Chen, F.; Zhang, D.; Han, M.; Chen, X.; Shi, J.; Xu, S.; and Xu, B. 2022. Vlp: A survey on vision-language pre-training. *arXiv preprint arXiv:2202.09061*.

Fei-Fei, L.; and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 524–531. IEEE.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1): 723–773.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2021. Trusted Multi-View Classification. In *International Conference on Learning Representations*.

Hazirbas, C.; Ma, L.; Domokos, C.; and Cremers, D. 2016. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, 213–228. Springer.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Li, Z.; Xu, P.; Chang, X.; Yang, L.; Zhang, Y.; Yao, L.; and Chen, X. 2023. When Object Detection Meets Knowledge Distillation: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lin, Y.; Gou, Y.; Liu, Z.; Li, B.; Lv, J.; and Peng, X. 2021. COMPLETER: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11174–11183.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *ICML*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Srivastava, N.; and Salakhutdinov, R. R. 2012. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25.

Vendrov, I.; Kiros, R.; Fidler, S.; and Urtasun, R. 2016. Order-embeddings of images and language. In *International Conference on Learning Representations*.

Wan, Z.; Zhang, C.; Geng, Y.; Fu, H.; Peng, X.; Zhu, P.; and Hu, Q. 2021. Cross-view equivariant auto-encoder. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

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

Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; and Huang, J. 2020. Deep multimodal fusion by channel exchanging. *Advances in Neural Information Processing Systems*, 33.

Wen, J.; Zhang, Z.; Fei, L.; Zhang, B.; Xu, Y.; Zhang, Z.; and Li, J. 2022. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Wen, J.; Zhang, Z.; Zhang, Z.; Fei, L.; and Wang, M. 2020a. Generalized incomplete multiview clustering with flexible locality structure diffusion. *IEEE transactions on cybernetics*, 51(1): 101–114.

Wen, J.; Zhang, Z.; Zhang, Z.; Wu, Z.; Fei, L.; Xu, Y.; and Zhang, B. 2020b. DIMC-net: Deep Incomplete Multi-view Clustering Network. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3753–3761.

Xu, C.; Guan, Z.; Zhao, W.; Wu, H.; Niu, Y.; and Ling, B. 2019. Adversarial Incomplete Multi-view Clustering. In *IJCAI*, 3933–3939.

Xu, C.; Zhao, W.; Zhao, J.; Guan, Z.; Song, X.; and Li, J. 2022. Uncertainty-aware multi-view deep learning for internet of things applications. *IEEE Transactions on Industrial Informatics*.

Xu, J.; Li, W.; Liu, X.; Zhang, D.; Liu, J.; and Han, J. 2020. Deep embedded complementary and interactive information for multi-view classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 6494–6501.

Zeng, J.; Tong, Y.; Huang, Y.; Yan, Q.; Sun, W.; Chen, J.; and Wang, Y. 2019. Deep surface normal estimation with hierarchical rgb-d fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6153–6162.

Zhang, C.; Geng, Y.; Han, Z.; Liu, Y.; Fu, H.; and Hu, Q. 2022. Autoencoder in Autoencoder Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Zhang, C.; Wang, S.; Liu, J.; Zhou, S.; Zhang, P.; Liu, X.; Zhu, E.; and Zhang, C. 2021. Multi-view clustering via deep matrix factorization and partition alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4156–4164.