Beyond Shared Subspace: A View-Specific Fusion for Multi-View Multi-Label Learning

Gengyu Lyu*, Xiang Deng*, Yanan Wu, Songhe Feng†

Beijing Key Laboratory of Traffic Data Analysis and Mining
School of Computer and Information Technology, Beijing Jiaotong University
{18112030, 20120346, 19112034, shfeng}@bjtu.edu.cn

Abstract

In multi-view multi-label learning (MVML), each instance is described by several heterogeneous feature representations and associated with multiple valid labels simultaneously. Although diverse MVML methods have been proposed over the last decade, most previous studies focus on leveraging the shared subspace across different views to represent the multi-view consensus information, while it is still an open issue whether such shared subspace representation is necessary when formulating the desired MVML model. In this paper, we propose a DeepGCN based View-Specific MVML method (D-VSM) which can bypass seeking for the shared subspace representation, and instead directly encoding the feature representation of each individual view through the deep GCN to couple with the information derived from the other views. Specifically, we first construct all instances under different feature representations into the corresponding feature graphs respectively, and then integrate them into a unified graph by integrating the different feature representations of each instance. Afterwards, the graph attention mechanism is adopted to aggregate and update all nodes on the unified graph to form structural representation for each instance, where both intra-view correlations and inter-view alignments have been jointly encoded to discover the underlying semantic relations. Finally, we derive a label confidence score for each instance by averaging the label confidence of its different feature representations with the multi-label soft margin loss. Extensive experiments have demonstrated that our proposed method significantly outperforms state-of-the-art methods.

Introduction

Multi-View Multi-Label learning (MVML) learns from the training data, where each instance is represented by several heterogeneous feature representations and associated with multiple valid labels simultaneously (Luo et al. 2013; Liu et al. 2015; Zhang et al. 2018; Tan, Yu, and Wang 2019). Recently, such learning paradigm has been widely used in many real-world applications. For example, in film classification (Figure 1), given the film of *The Big Bang Theory*, which is represented by diverse channel information (*audio*, *cover picture*, *text description*) and annotated with multiple labels (*comedy movie*, *America*, *Mark Cendrowski*), MVML provides an effective framework to learn from such complicate data and predicts proper labels for unseen instances.

Figure 1: An exemplar of multi-view multi-label learning.

The main challenge to deal with multi-view data lies in how to integrate the multiple types of heterogeneities in an efficient way. A general practice is to learn a shared subspace representation to excavate and exploit the consensus and complementary information among different views. For example, (Liu et al. 2015) employs matrix factorization to seek a shared low-dimensional representation, which further strengthens the complementarities across different views by considering the different contributions of multiple views' reconstruction. (Zhang et al. 2018) also learns a shared subspace representation under matrix factorization framework, and it simultaneously employs Hilbert-Schmidt independence criterion to further remain the consensus on the shared representation. Although the above methods have achieved competitive performance in many MVML tasks, they suffer from the limitation of shared subspace inevitably, i.e., it is hard for a single shared subspace to fully capture the global structure of multi-view data and comprehensively characterize all the relevant labels, without exploring the distinctive information hidden in individual views.

To tackle the above issue, in this paper, we bypass the shared subspace strategy and propose a DeepGCN based View-Specific MVML method named D-VSM, where each individual view fused with other views' complementarities can directly contribute to the final discriminative model. Specifically, we first construct all instances under different views into different feature graphs respectively, i.e.,

*Gengyu Lyu and Xiang Deng have equal contributions.
†Songhe Feng is the corresponding author.

each view corresponds to a feature graph and each node is described by one feature representation of an instance. Then, the above graphs are integrated into a unified feature graph by connecting the different feature representation nodes within each instance. Afterwards, we employ graph attention mechanism to fuse both intra-view correlations and inter-view alignments into each feature node to form structural representations for each instance. Here, the intra-view correlations reflect the instance relationship under each individual view, while the inter-view alignments reflects the view connections across each instance's views. Finally, we derive a label confidence score for each instance by averaging the label confidence of its different feature representations with the multi-label soft margin loss.

In summary, the contributions of our paper lie in the following aspects:

- We propose a novel MVML method named D-VSM, which unveils new opportunity to surpass the limitations of shared subspace to better compromising of multi-view consensus and complementary information.
- D-VSM not only exploits the consensuses and complementarities across different views, but also focuses more on view-specific information extraction, which significantly improves the performance of the learning model.
- Enormous experimental results as well as comprehensive ablation study have demonstrated the superiority of our proposed D-VSM against state-of-the-art methods.

Related Work

Multi-view multi-label learning (MVML) is related to two branches of studies: multi-label learning (MLL) and multi-view learning (MVL). Due to page limit, we briefly review some related works about the two studies and introduce some recent works about MVML. For more details, please refer to (Zhang and Zhou 2013; Zhao et al. 2017).

Multi-Label Learning (MLL)

Multi-Label Learning focuses on learning from data with multiple labels, and existing MLL methods can be generally grouped into two categories: *Problem Transformation*-based methods and *Algorithm Adaption*-based methods. 1) *Problem Transformation*-based methods usually transfer the MLL problem into some single-label problems, and adapt existing single-label learning algorithms to handle multi-label data, such as BR (Tsoumakas and Katakis 2007), ECC (Read et al. 2011) and RakeLD (Tsoumakas, Katakis, and Vlahavas 2011). 2) *Algorithm Adaption*-based methods usually convert the task of multi-label classification to some well-established learning scenarios, and extend some off-the-shelf algorithms to directly deal with multi-label data. ML-KNN (Zhang and Zhou 2007), MLARAM (Benites and Sapozhnikova 2015) and LIFT (Zhang and Wu 2015) are the representative methods for such category.

Multi-View Learning (MVL)

Multi-view learning learns from examples with heterogeneous features, and its challenge lies in how to integrate the different feature representations in an effective way. Recently, (Nie, Cai, and Li 2017) proposes a parameter-free multi-view model, which learns the local structure among multi-view data to achieve semi-supervised classification. (Li and He 2020) proposes a bipartite graph based multi-view clustering method, where a unified bipartite graph matrix is employed to fuse the consensus information across different views and directly form the final clustering results. Besides, there are also many other MVL methods for different tasks, such as clustering (Bickel and Scheffer 2004), retrieval (Kludas, Bruno, and Marchand 2007) and classification (Luo et al. 2015), etc.

Multi-View Multi-Label Learning (MVML)

Multi-view multi-label learning can be regarded as an integration of MLL and MVL, which aims to learn from training data with diverse representations and rich semantics. To learn from such complicated data, (Xing et al. 2018) proposes a predictive reliability measure, which selects examples that share label information with other views in an co-training manner. (Tan, Yu, and Wang 2019; Zhang, Jia, and Li 2020) focus on learning a shared subspace to fuse the complementarities across different views, and directly obtain the corresponding projection model between the shared subspace and labels. (Zhang et al. 2018) leverages matrix factorization to learn a shared subspace representation, and it simultaneously employs Hilbert-Schmidt independence criterion to further remain the consensuses on the shared representation. Besides the above methods, some recent methods have been proposed to learn from multi-view data with weak labels, such as (Tan et al. 2018; Wu et al. 2019; Li and Chen 2021).

The Proposed Method

Formally speaking, we denote $\mathcal{X} = \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \ldots \times \mathbb{R}^{d_T}$ as the feature space with T views and $\mathcal{Y} = \{c_1, c_2, \ldots, c_q\}$ as the label space with q class labels, where d_t ($1 \leq t \leq T$) is the feature dimension of t-th view. Given the MVML training data $\mathcal{D} = \{(\boldsymbol{X}_i, \boldsymbol{y}_i) | 1 \leq i \leq n\}$ with n instances, where $\boldsymbol{X}_i \in \mathcal{X}$ is represented by T feature vectors $[\boldsymbol{x}_i^{(1)}; \boldsymbol{x}_i^{(2)}; \ldots; \boldsymbol{x}_i^{(T)}]$ and $\boldsymbol{y}_i \in \{0,1\}^{q \times 1}$ is the label vector associated with $\boldsymbol{X}_i$, our proposed D-VSM aims to integrate these diverse representations from different views to construct a robust multi-label classifier $\boldsymbol{f} : \mathcal{X} \mapsto 2^{\mathcal{Y}}$ and further predicts proper labels for unseen instances. Figure 2 illustrates the overview architecture of D-VSM, which consists of three key components: *Multi-View Feature Graph Construction*, *Structural Feature Representation* and *Multi-Label Classification*.

Multi-View Feature Graph Construction

As depicted in Figure 2, we construct all instances under different views into different graphs $\mathbb{G}^{(t)} = (\mathbb{V}^{(t)}, \mathbb{E}^{(t)})$ respectively, where $t \in \{1, 2, \ldots, T\}$. The nodes $\mathbb{V}^{(t)}$ in each graph represent the feature representations under t-th view, while the edges $\mathbb{E}^{(t)}$ encode their similarity. Specifically, in each graph $\mathbb{G}^{(t)}$, we describe each instance node $v_i^{(t)}$ by a d_t-

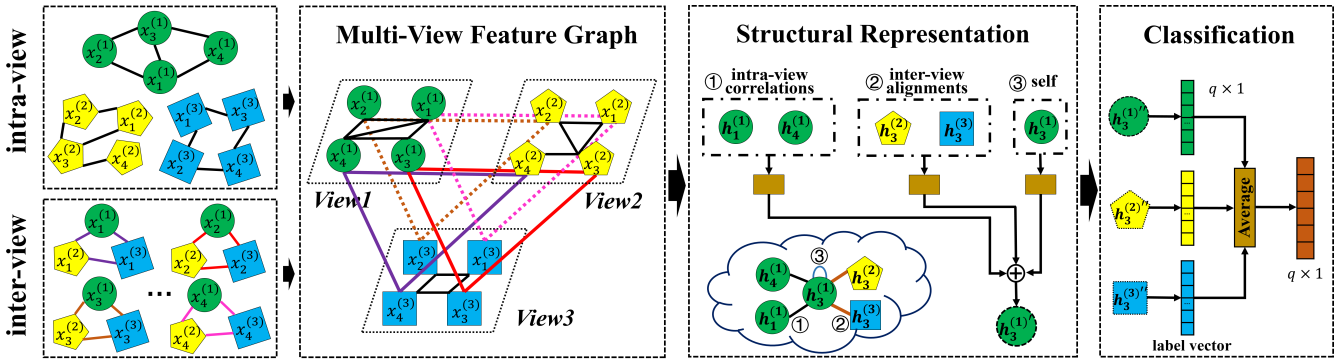

Figure 2: The framework of our proposed D-VSM, which consists of three components: (1) *Multi-View Feature Graph Construction*, where each node is connected with k intra-view neighbors and $V-1$ inter-view aligned node(s); (2) *Structural Feature Representation*, where each feature node consists of three different structural information, i.e., self-portraits, intra-view correlations and inter-view alignments. (3) *Multi-Label Classification*, where the final label confidence of each instance is derived by averaging the label confidences from different views with multi-label soft margin loss.

dimensional vector and then the edges $e^{(t)} \in \mathbb{E}^{(t)}$ between each pair of nodes can be produced following:

$$e_{ij}^{(t)} = \begin{cases} 1, & \text{where } v_j^{(t)} \in \mathcal{N}(v_i^{(t)}) \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where $\mathcal{N}(v_i^{(t)})$ denotes the k-nearest neighbors (measured by Euclidean distance) of $v_i^{(t)}$, and $e_{ij}^{(t)} = 1$ indicates an undirected edge from $v_i^{(t)}$ to $v_j^{(t)}$, $e_{ij}^{(t)} = 0$ otherwise.

After obtaining each individual feature representation graph, we connect the different feature representation nodes within each instance and integrate the above individual feature graphs into a unified multi-view feature representation graph, where the edges between different types of feature nodes (i.e., different views) encodes the view correlations between their connected views.

Structural Feature Representation

Given the original features $\{\boldsymbol{x}_1^{(t)}, \boldsymbol{x}_2^{(t)}, \ldots, \boldsymbol{x}_n^{(t)}\}$ under t-th view, we employ an attention-based DeepGCN architecture (R-GCN) (Schlichtkrull et al. 2018) to compute the hidden representation of each feature node $\{\boldsymbol{x}_1^{(t)'}, \boldsymbol{x}_2^{(t)'}, \ldots, \boldsymbol{x}_n^{(t)'}\}$ by attending its intra-view k-nearest neighbors and inter-view aligned feature representations.

Specifically, we first transform each original input feature vector $\{\boldsymbol{x}_i^{(t)}|_{i=1}^n\}$ into higher-level feature $\boldsymbol{h}_i^{(t)} = \sigma(\boldsymbol{H}^{(t)}\boldsymbol{x}_i^{(t)})$ to obtain sufficient expressive power, where $\boldsymbol{H}^{(t)} \in \mathbb{R}^{d \times d_t}$ is the shared linear transformation matrix and $\sigma(\cdot) = \max(0, \cdot)$ is the element-wise activation function. Then, each feature representation node in the unified multi-view graph can be updated by

$$\boldsymbol{h}_i^{(t)'} \Leftarrow \sigma\left(\boldsymbol{W}_0^{(t)}\boldsymbol{h}_i^{(t)} + \sum_{j=1}^{k}\frac{1}{k}\boldsymbol{W}_1^{(t)}\boldsymbol{h}_j^{(t)} + \sum_{o=1}^{T-1}\frac{1}{T-1}\boldsymbol{W}_2^{(t)}\boldsymbol{h}_i^{(o)}\right), \tag{2}$$

where $\boldsymbol{x}_j^{(t)}$ is the k-nearest neighbors of $\boldsymbol{x}_i^{(t)}$ under t-th view, $\boldsymbol{W}_0^{(t)}, \boldsymbol{W}_1^{(t)}, \boldsymbol{W}_2^{(t)} \in \mathbb{R}^{d \times d}$ encode the weight matrices, k and V denote the number of neighbors and views respectively. According to Eq. (2), we can observe that each feature representation is coupled with three types of structural information, i.e., self-portraits (first term), intra-view correlations (second term) and inter-view alignments (third term). Here, the intra-view correlations integrate the contributions of its k-nearest neighbors under the same view, while the inter-view alignments fuse the complementarity information across different views within the same instance, which jointly strengthes its identification capacity of characterizing instances and further improves the robustness of final model.

Furthermore, in order to avoid the model falling into overfitting, inspired by (Schlichtkrull et al. 2018), we regularize the weights $\boldsymbol{W}_0^{(t)}$, $\boldsymbol{W}_1^{(t)}$ and $\boldsymbol{W}_2^{(t)}$ as linear combinations for basis transformations $Q_c^{(t)} \in \mathbb{R}^{d \times d}$ with coefficients $a_{rc}^{(t)}$, i.e.,

$$\boldsymbol{W}_r^{(t)} = \sum_c a_{rc}^{(t)} Q_c^{(t)}, \text{ where } r, c \in \{0, 1, 2\}. \tag{3}$$

In addition, to further consider the contributions of other instances' representations in different views and strength the identification capacity of the learned structural feature representations, in our experiments, we also exploit the outputs of Eq. (2) as its inputs and repeat such propagation operation to fuse more inter-view complementary information into each feature node, then obtain desired structural feature representations $\boldsymbol{h}_i^{(t)''}$ for subsequent multi-label classification.

Multi-Label Classification

In our proposed D-VSM, we focus on bypassing the limitations of shared subspace and directly employing each individual structural feature representation $\boldsymbol{h}_i^{(t)''}$ to derive label confidence scores $[p_{i1}^{(t)}, p_{i2}^{(t)}, \ldots, p_{iq}^{(t)}]$ for each instance $\boldsymbol{x}_i^{(t)}$. Afterwards, the final label confidence score

Algorithm 1: The Training Process of **D-VSM**

Inputs:

$\mathcal{D}$: multi-view multi-label training set $\{(\boldsymbol{X}_i, \boldsymbol{y}_i)\}$;

I_m: the number of epoches;

Process:

1. Construct V individual feature graphs $\mathbb{G}^{(t)}$ under different views, where each edge is defined by Eq. (1);

2. Integrate the V individual feature graphs into a unified multi-view feature graph by connecting different feature representation nodes within an instance;

3. for epoch $= 1$ **to** I_m

4. **// Forward Propagation**

5. Transform the original feature $\boldsymbol{x}_i^{(t)}$ into higher-level feature by $\boldsymbol{h}_i^{(t)} = \sigma(\boldsymbol{H}^{(t)}\boldsymbol{x}_i^{(t)})$;

6. Update $\boldsymbol{h}_i^{(t)}$ by Eq. (2);

7. Repeat **Step 6** and obtain $\boldsymbol{h}_i^{(t)''}$;

8. Obtain the label confidence vector $\boldsymbol{p}_i^{(t)}$ of $\boldsymbol{X}_i$ under t-th view and Calculate the final label confidence $\boldsymbol{p}_i$ by Eq. (4);

9. **// Backward Propagation**

10. Update the model parameters by minimizing multi-label soft margin loss in Eq. (5);

11. end for

Output:

$\boldsymbol{f}$: the classification model of D-VSM;

$[p_{i1}, p_{i2}, \ldots, p_{iq}]$ of each instance $\boldsymbol{X}_i$ is calculated by averaging the label confidences from different views

$$p_{ij} = \frac{1}{T}\sum_{t=1}^{T} p_{ij}^{(t)}, \text{ where } i \in [n] \text{ and } j \in [q], \tag{4}$$

with the widely-used multi-label soft margin loss, i.e.,

$$\mathcal{L} = \sum_{i=1}^{n}\sum_{j=1}^{q}\left(-y_{ij}\log(S(p_{ij})) + (1-y_{ij})\log(1-S(p_{ij}))\right), \tag{5}$$

where $S(p_{ij}) = \frac{1}{1+\exp(-p_{ij})}$ is the sigmoid function.

Experiments

Experimental Setup

To evaluate the performance of our proposed D-VSM, we implement experiments on six benchmark data sets. *Emotions*[1] have 593 pieces of music described by two views: 8 rhythmic properties and 64 timbre properties. *Scene*[1] comprises of 2407 images, where 294 features from two views separately reflect the luminance and chromaticity of color. *Corel5k* (Duygulu et al. 2002) and *Espgame* (Von Ahn and Dabbish 2004) contain 4999 and 20770 images respectively, all of which are represented by 4 different features: GIST, HSV, HUE, DIFT. *Pascal* (Everingham et al. 2010) and *Mirflickr* (Huiskes and Lew 2008), besides the above four views, add the textual views to describe their tag features. Table 1 summarizes the characteristics of the above data sets.

[1]http://mulan.sourceforge.net/datasets-mlc.html

Data sets	Instances	Views	$D_{min-max}$	Labels
Emotions	593	2	8 - 64	6
Scene	2407	2	98 - 196	6
Corel5k	4999	4	100 - 4096	260
Pascal	9963	5	512 - 4096	20
Iaprtc12	19627	6	100 - 4096	291
Espgame	20770	4	100 - 4096	268
Mirflickr	25000	5	100 - 4096	38

Table 1: Characteristics of our employed data sets. And $D_{min-max}$ is the smallest-largest dimensions of features.

Meanwhile, we employ six state-of-the-art methods from two categories for comparative studies: 1) Multi-label learning methods including **ML-KNN**, **RakeLD** and **LSPC**, which concatenate all view features as the input of the learning model; 2) Multi-view multi-label methods including **LrMMC**, **SIMM** and **FIMAN**, which fuses the complementarities across different views for classification model induction. The configured parameters of the above methods are set according to the suggestions in respective literature.

- **ML-KNN** (Zhang and Zhou 2007): which concatenates the features of all views as the model input, and induces the model via k-NN scheme. [configuration: $k = 10$] ;
- **RakeLD** (Tsoumakas, Katakis, and Vlahavas 2011): which randomly breaks the initial label set into several small label subsets and employs LP strategy to train the classifier. [configuration: $k = q/10$];
- **LrMMC** (Liu et al. 2015): which aims to learn a low-dimensional shared subspace and leverages the matrix completion for MVML classification. [configuration: $\gamma \in \{10^{-4}, 10^{-3}, \ldots, 10^{3}\}$];
- **LSPC** (Szyma¨½ski, Kajdanowicz, and Kersting 2016): which divides the label space according to the label co-occurrence graphs, and then obtains an ensemble of multi-label classifier.;
- **SIMM** (Wu et al. 2019): which simultaneously leverages the shared subspace exploitation and view-specific information extraction, and induces MVML model via minimizing confusion adversarial loss and multi-label loss. [configuration: $\alpha = 1$, $\beta \in \{10^{-4}, 10^{-3}, \ldots, 10^{-1}\}$] ;
- **FIMAN** (Wu et al. 2020): which aims to learn from multi-view data with partial multiple labels, where an aggregate manifold structure is leveraged to adaptively fuse feature representation from different views. [configuration: $k = 10$, $t_d = 0.4$, $t_p = 0.6$ and $\eta = 1$];

In addition, six popular multi-label metrics are employed to evaluate each comparing method, including *Hamming Loss* (H-L), *Ranking Loss* (R-L), *One-Error* (O-E), *Coverage* (Cov), *Average Precision* (A-P) and *Micro-F1* (M-F1), whose detailed definitions can be found in (Zhang and Zhou 2013) or (Sun and Zong 2021). Finally, we conduct experimental comparison between our proposed D-VSM and all comparing methods, where five-fold cross-validation is performed on each data set.

H-L	Emotions	Scene	Corel5k	Pascal	Iaprtc12	Espgame	Mirflickr
D-VSM	**0.179±0.017**	**0.075±0.005**	**0.012±0.000**	**0.048±0.000**	**0.018±0.000**	**0.017±0.000**	**0.005±0.000**
ML-KNN	0.311±0.001	0.153±0.009	0.033±0.000	0.181±0.001	0.055±0.000	0.050±0.000	0.022±0.000
RakeLD	0.250±0.019	0.155±0.012	0.085±0.002	0.248±0.003	0.193±0.002	0.177±0.001	0.172±0.001
LrMMC	0.196±0.011	0.082±0.006	0.013±0.000	0.073±0.000	0.029±0.000	0.028±0.000	0.006±0.000
LSPC	0.251±0.014	0.221±0.008	0.020±0.000	0.219±0.003	0.027±0.000	0.025±0.000	0.013±0.000
SIMM	0.307±0.004	0.179±0.002	0.013±0.000	0.060±0.001	0.019±0.000	**0.017±0.000**	0.006±0.000
FIMAN	0.231±0.013	0.195±0.005	0.018±0.000	0.116±0.002	0.026±0.000	0.028±0.000	-
R-L	Emotions	Scene	Corel5k	Pascal	Iaprtc12	Espgame	Mirflickr
D-VSM	**0.137±0.015**	**0.058±0.007**	**0.084±0.004**	**0.077±0.001**	**0.090±0.004**	**0.133±0.003**	**0.170±0.001**
ML-KNN	0.347±0.017	0.123±0.008	0.143±0.004	0.277±0.006	0.172±0.003	0.189±0.004	0.234±0.003
RakeLD	0.195±0.029	0.133±0.011	0.831±0.005	0.610±0.020	0.568±0.007	0.578±0.007	0.540±0.010
LrMMC	0.233±0.016	0.115±0.011	0.173±0.004	0.336±0.005	0.390±0.002	0.410±0.003	0.251±0.006
LSPC	0.185±0.022	0.233±0.023	0.860±0.005	0.868±0.003	0.996±0.000	0.993±0.000	0.702±0.005
SIMM	0.344±0.047	0.280±0.026	0.160±0.005	0.097±0.006	0.124±0.002	0.164±0.003	0.268±0.006
FIMAN	0.161±0.026	0.107±0.006	0.085±0.000	0.118±0.003	0.111±0.002	0.154±0.002	-
O-E	Emotions	Scene	Corel5k	Pascal	Iaprtc12	Espgame	Mirflickr
D-VSM	**0.214±0.046**	**0.189±0.016**	**0.410±0.018**	**0.274±0.008**	**0.422±0.010**	**0.464±0.010**	**0.860±0.006**
ML-KNN	0.535±0.038	0.331±0.017	0.707±0.018	0.626±0.012	0.714±0.008	0.737±0.003	0.955±0.001
RakeLD	0.325±0.039	0.362±0.012	0.831±0.005	0.864±0.009	0.953±0.003	0.918±0.005	0.939±0.006
LrMMC	0.338±0.032	0.272±0.019	0.776±0.015	0.596±0.005	0.944±0.003	0.992±0.001	0.944±0.003
LSPC	0.295±0.036	0.397±0.028	0.890±0.008	0.926±0.007	0.990±0.001	0.988±0.002	0.939±0.003
SIMM	0.501±0.092	0.603±0.039	0.614±0.012	0.391±0.009	0.528±0.007	0.536±0.004	0.888±0.004
FIMAN	0.258±0.042	0.280±0.018	0.489±0.017	0.313±0.011	0.511±0.002	0.628±0.004	-
Cov	Emotions	Scene	Corel5k	Pascal	Iaprtc12	Espgame	Mirflickr
D-VSM	1.624±0.038	0.372±0.051	**53.21±1.947**	**2.342±0.044**	**75.08±1.833**	**86.75±1.634**	**137.6±1.690**
ML-KNN	2.703±0.182	0.702±0.034	83.87±1.994	6.851±0.178	130.3±1.115	120.3±1.330	181.2±2.577
RakeLD	1.932±0.132	0.754±0.054	195.6±4.611	13.11±0.457	261.9±1.890	237.1±1.167	304.7±5.157
LrMMC	2.198±0.094	0.677±0.057	96.72±1.300	7.900±0.060	196.7±1.226	210.9±1.020	186.9±3.199
LSPC	1.905±0.138	1.252±0.109	257.3±0.499	17.08±0.065	289.9±0.061	266.8±0.078	335.9±2.740
SIMM	**0.457±0.051**	**0.248±0.020**	95.99±3.146	2.772±0.140	106.0±1.892	110.1±1.260	162.6±2.897
FIMAN	1.796±0.189	0.628±0.020	53.94±0.790	3.486±0.081	97.06±1.436	102.8±1.183	-
A-P	Emotions	Scene	Corel5k	Pascal	Iaprtc12	Espgame	Mirflickr
D-VSM	**0.835±0.027**	**0.890±0.011**	**0.475±0.008**	**0.767±0.003**	**0.412±0.002**	**0.364±0.003**	**0.369±0.001**
ML-KNN	0.620±0.014	0.801±0.009	0.303±0.010	0.432±0.010	0.238±0.004	0.225±0.002	0.302±0.005
RakeLD	0.764±0.030	0.760±0.025	0.121±0.008	0.179±0.005	0.056±0.002	0.063±0.003	0.266±0.011
LrMMC	0.763±0.020	0.852±0.012	0.215±0.010	0.422±0.004	0.219±0.003	0.170±0.003	0.053±0.001
LSPC	0.773±0.025	0.647±0.020	0.075±0.004	0.116±0.003	0.021±0.000	0.020±0.003	0.272±0.006
SIMM	0.634±0.043	0.608±0.027	0.292±0.004	0.685±0.010	0.326±0.003	0.308±0.002	0.119±0.003
FIMAN	0.806±0.027	0.827±0.010	0.430±0.007	0.721±0.003	0.348±0.002	0.284±0.002	-
M-F1	Emotions	Scene	Corel5k	Pascal	Iaprtc12	Espgame	Mirflickr
D-VSM	**0.700±0.034**	**0.777±0.015**	**0.399±0.004**	**0.636±0.004**	**0.385±0.003**	**0.332±0.003**	**0.055±0.002**
ML-KNN	0.154±0.003	0.113±0.001	0.030±0.001	0.074±0.001	0.031±0.000	0.027±0.000	0.002±0.000
RakeLD	0.615±0.035	0.635±0.015	0.153±0.007	0.199±0.011	0.091±0.002	0.086±0.001	0.020±0.001
LrMMC	0.685±0.018	0.772±0.017	0.273±0.009	0.283±0.015	0.281±0.003	0.209±0.003	0.032±0.003
LSPC	0.653±0.022	0.544±0.019	0.153±0.004	0.084±0.003	0.004±0.001	0.008±0.001	0.052±0.002
SIMM	0.034±0.051	0.001±0.002	0.038±0.010	0.343±0.011	0.047±0.005	0.047±0.003	0.000±0.000
FIMAN	0.671±0.014	0.616±0.008	0.361±0.009	0.008±0.002	0.289±0.001	0.242±0.002	-

Table 2: Experimental comparisons of D-VSM with other comparing methods on six evaluation metrics, where the best performances on each metric are shown in bold face. "-" indicates that FIMAN needs over 256G of RAM on *Mirflickr* data set.

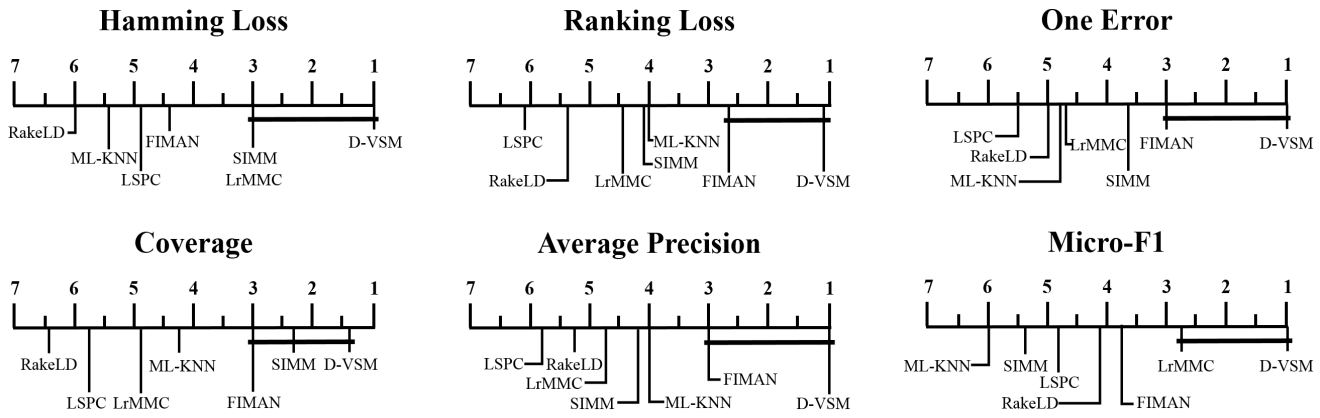

Figure 3: Experimental Comparisons of our proposed D-VSM against other comparing algorithms with the Bonferroni-Dunn test. Algorithms not connected with D-VSM are significantly inferior to D-SVM (CD = 3.046 at 0.05 significance level).

Emotions	Hamming Loss	Ranking Loss	One Error	Coverage	Average Precision	Micro-F1
D-VSM-v0k0	0.194±0.016	0.148±0.007	0.237±0.033	1.708±0.132	0.817±0.015	0.661±0.046
D-VSM-v0k1	0.186±0.020	0.139±0.012	0.234±0.018	1.644±0.106	0.824±0.016	0.687±0.025
D-VSM-v1k0	0.181±0.013	0.140±0.017	0.229±0.037	1.651±0.134	0.826±0.022	0.698±0.026
D-VSM	**0.179±0.017**	**0.137±0.015**	**0.214±0.046**	**1.624±0.038**	**0.835±0.027**	**0.700±0.034**
Corel5k	Hamming Loss	Ranking Loss	One Error	Coverage	Average Precision	Micro-F1
D-VSM-v0k0	0.012±0.000	0.096±0.004	0.486±0.038	60.54±1.689	0.420±0.016	0.220±0.021
D-VSM-v0k1	0.012±0.000	0.089±0.004	0.440±0.013	56.49±3.220	0.456±0.006	0.322±0.020
D-VSM-v1k0	0.012±0.000	0.103±0.007	0.450±0.014	61.65±3.541	0.440±0.012	0.357±0.010
D-VSM	**0.012±0.000**	**0.084±0.002**	**0.410±0.018**	**53.21±1.947**	**0.475±0.008**	**0.399±0.004**
Mirflickr	Hamming Loss	Ranking Loss	One Error	Coverage	Average Precision	Micro-F1
D-VSM-v0k0	0.007±0.001	0.206±0.006	0.933±0.016	158.3±2.215	0.328±0.003	0.045±0.003
D-VSM-v0k1	0.005±0.000	0.185±0.003	0.885±0.008	146.2±1.735	0.346±0.001	0.052±0.002
D-VSM-v1k0	0.006±0.000	0.193±0.005	0.908±0.010	150.5±1.766	0.330±0.001	0.049±0.002
D-VSM	**0.005±0.000**	**0.170±0.001**	**0.860±0.006**	**137.6±1.690**	**0.369±0.001**	**0.055±0.002**

Table 3: The experimental results of our proposed D-VSM and its three degenerated methods over all employed evaluation metrics on *Emotions* and *Corel5k* data sets, where D-VSM-v0k1, D-VSM-v1k0 and D-VSM-v0k0 do not consider the inter-view alignments, intra-view correlations and both of them, respectively.

Evaluation Metric	τ_F	critical value
Hamming Loss	7.629	
Ranking Loss	8.806	
One Error	5.710	2.365
Coverage	15.550	Methods: 7, Data sets: 7
Average Precision	7.108	
Micro-F1	9.954	

Table 4: Friedman statics τ_F in terms of each evaluation metric (at 0.05 significance level).

Experimental Results

Table 2 illustrates the experimental comparisons between our proposed D-VSM and other six comparing methods on all evaluation metrics, where the mean metrics results and standard deviations are recorded respectively. Out of 252 (7 data sets × 6 methods × 6 metrics) statistical comparisons, the following observations can be made:

- Among all comparing methods, D-VSM is superior to **ML-KNN**, **RakeLD**, **LrMMC**, **LSPC** and **FIMAN** in all cases, and it also outperforms **SIMM** in 95.2% cases.
- D-VSM achieves the best performance on all metrics except for *Coverage*. And on *Coverage* metric, it is also superior to other comparing methods over 95% cases.
- The improvements of D-VSM against other methods are quite significant, especially it ranks first in almost all comparisons and is well ahead of the second.

In order to comprehensively evaluate the superiority of the proposed D-VSM, *Friedman test* (Demšar 2006) is utilized as the statistical test to analyze the relative performance

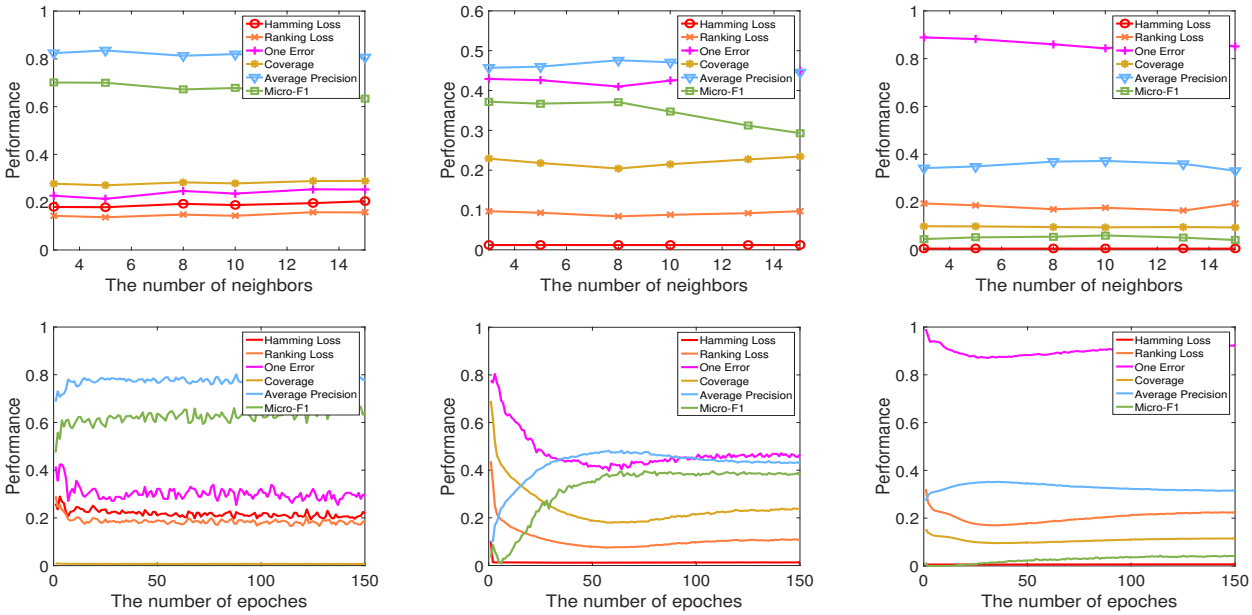

Figure 4: The parameter analysis [top] and convergence [bottom] analysis of D-VSM on *Emotions* [left], *Corel5k* [mid] and *Mirflickr* [right] data sets respectively, where the *Coverage* results are normalized by the number of class labels (q) so as to make all metric results be characterized in a unified figure.

among the comparing algorithms. According to Table 4, the null hypothesis of distinguishable performance among the comparing algorithms is rejected at 0.05 significance level. Therefore, we further employ the post-hoc Bonferroni-Dunn test (Demšar 2006) to show the relative performance among the comparing algorithms. Figure 3 illustrates the CD diagrams on each evaluation metric, where the average rank of each comparing algorithm is marked along the axis. According to Figure 3, it is observed that D-VSM ranks 1st on all evaluation metrics and it performs significant superiority against most comparing methods.

Further Analysis

Ablation Study

In order to evaluate the effect of the employed intra-view correlations and inter-view alignments, we conduct the Ablation Study between D-VSM and its three degenerated algorithms D-VSM-v0k1, D-VSM-v0k1 and D-VSM-v0k0, where each degenerated algorithm ignores the inter-view alignments, intra-view correlations and both of them, respectively. Table 3 records the experimental results on *Emotions*, *Corel5k* and *Mirflickr* data sets. According to Table 3, D-VSM-v0k1 outperforms D-VSM-v1k0 in most cases, which indicates that intra-view correlations may have greater contributions than inter-view alignments to the robustness of model. Meanwhile, D-VSM significantly outperforms its three degenerated algorithms, which also strongly demonstrates the superiority of employing both of two relationships simultaneously when learning from MVML data.

Sensitivity Analysis

We study the sensitivity analysis of D-VSM with respect to its employed parameter k: the number of intra-view neighbors. Figure 4 show the performance of D-VSM as k increases from 3 to 15 on *Emotions* [left], *Corel5k* [mid] and *Mirflickr* [right] data sets. As illustrated in Figure 4, the performance of D-VSM improves slightly and become stable shortly as k increases. In our experiments, we set $k = 5$.

Convergence Analysis

We conduct the convergence analysis of D-VSM on both *Emotions* [left], *Corel5k* [mid] and *Mirflickr* [right] data sets, where experimental results are illustrated in Figure 4. According to Figure 4, we can easily observe that the performance of D-VSM gradually improves and soon reaches stability as the number of epoches increases. Therefore, the convergence of D-VSM is empirically demonstrated.

Conclusion

In this paper, we proposed a DeepGCN based View-Specific MVML method named D-VSM, which fuses the complementarities across different views into each individual view, and directly employs these individual views to induce the final model. Compared with previous methods, D-VSM surpasses the limitations of shared subspace, and improves the model performance by exploiting both the complementary information across different views and the view-specific information within individual view simultaneously. Enormous experimental results have verified that our proposed D-VSM has significant superiority against state-of-the-art methods when learning from MVML data.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 61872032, No. 62072027, No. 62076021), the Beijing Natural Science Foundation (No. 4202058, No. 4202057, No. 4202060), the Fundamental Research Funds for the Central universities (No. 2019JBM020, No. 2020YJS026), and in part by the National Key Research and Development Project (No. 2018AAA0100300).

References

Benites, F.; and Sapozhnikova, E. 2015. Haram: a hierarchical aram neural network for large-scale text classification. In *IEEE International Conference on Data Mining*, 847–854.

Bickel, S.; and Scheffer, T. 2004. Multi-view clustering. In *International Conference on Data Mining*, volume 4, 19–26.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan): 1–30.

Duygulu, P.; Barnard, K.; Freitas, J.; and Forsyth, D. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, 97–112.

Everingham, M.; Gool, L.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes Challenge. *International Journal of Computer Vision*, 88(2): 303–338.

Huiskes, M.; and Lew, M. 2008. The MIR flickr retrieval evaluation. In *ACM International Conference on Multimedia Information Retrieval*, 39–43.

Kludas, J.; Bruno, E.; and Marchand, S. 2007. Information fusion in multimedia information retrieval. In *International Workshop on Adaptive Multimedia Retrieval*, 147–159.

Li, L.; and He, H. 2020. Bipartite graph based multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, Early Access: 1–15.

Li, X.; and Chen, S. 2021. A Concise yet Effective Model for Non-Aligned Incomplete Multi-view and Missing Multi-label Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Early Access: 1–15.

Liu, M.; Luo, Y.; Tao, D.; Xu, C.; and Wen, Y. 2015. Low-rank multi-view learning in matrix completion for multi-label image classification. In *AAAI Conference on Artificial Intelligence*, 2778–2784.

Luo, Y.; Liu, T.; Tao, D.; and Xu, C. 2015. Multiview matrix completion for multilabel image classification. *IEEE Transactions on Image Processing*, 24(8): 2355–2368.

Luo, Y.; Tao, D.; Xu, C.; Xu, C.; Liu, H.; and Wen, Y. 2013. Multiview vector-valued manifold regularization for multilabel image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 24(5): 709–722.

Nie, F.; Cai, G.; and Li, X. 2017. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI Conference on Artificial Intelligence*, 2408–2414.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2011. Classifier chains for multi-label classification. *Machine Learning*, 85(3): 333–359.

Schlichtkrull, M.; Kipf, T.; Bloem, P.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, 593–607.

Sun, S.; and Zong, D. 2021. LCBM: A Multi-view Probabilistic Model for Multi-label Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2682–2696.

Szyma¨½ski, P.; Kajdanowicz, T.; and Kersting, K. 2016. How Is a Data-Driven Approach Better than Random Choice in Label Space Division for Multi-Label Classification? *Entropy*, 18(8): 282.

Tan, Q.; Yu, G.; Domeniconi, C.; and Zhang, Z. 2018. Incomplete multi-view weak-label learning. In *International Joint Conference on Artificial Intelligence*, 2703–2709.

Tan, Q.; Yu, G.; and Wang, J. 2019. Individuality-and commonality-based multiview multilabel learning. *IEEE Transactions on Cybernetics*, 51(3): 1716–1727.

Tsoumakas, G.; and Katakis, I. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3): 1–13.

Tsoumakas, G.; Katakis, I.; and Vlahavas, I. 2011. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7): 1079–1089.

Von Ahn, L.; and Dabbish, L. 2004. Labeling images with a computer game. In *SIGCHI Conference on Human Factors in Computing Systems*, 319–326.

Wu, J.; Wu, X.; Chen, Q.; and Zhang, M. 2020. Feature-induced manifold disambiguation for multi-view partial multi-label learning. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 557–565.

Wu, X.; Chen, Q.; Hu, Y.; Wang, D.; Wang, X.; and Zhang, M. 2019. Multi-View Multi-Label Learning with View-Specific Information Extraction. In *Intenational Joint Conference of Artificial Intelligence*, 3884–3890.

Xing, Y.; Yu, G.; Domeniconi, C.; Wang, J.; and Zhang, Z. 2018. Multi-label co-training. In *International Joint Conference on Artificial Intelligence*, 2882–2888.

Zhang, C.; Yu, Z.; Hu, Q.; Zhu, P.; and Wang, X. 2018. Latent semantic aware multi-view multi-label classification. In *AAAI Conference on Artificial Intelligence*, 4414–4421.

Zhang, F.; Jia, X.; and Li, W. 2020. Tensor-based multi-view label enhancement for multi-label learning. In *Intenational Joint Conference of Artificial Intelligence*, 2369–2375.

Zhang, M.; and Wu, L. 2015. Lift: Multi-Label Learning with Label-Specific Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1): 107–120.

Zhang, M.; and Zhou, Z. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7): 2038–2048.

Zhang, M.; and Zhou, Z. 2013. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8): 1819–1837.

Zhao, J.; Xie, X.; Xu, X.; and Sun, S. 2017. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38: 43–54.