Multi-View Multi-Label Learning with View-Specific Information Extraction

Xuan Wu[1,2] , **Qing-Guo Chen**[2] , **Yao Hu**[2] , **Dengbao Wang**[4] , **Xiaodong Chang**[2] , **Xiaobo Wang**[2] and **Min-Ling Zhang**[1,3*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China

[2]Alibaba Group, Hangzhou, China

[3]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China

[4]College of Computer and Information Science, Southwest University, Chongqing 400715, China

wuxuan@seu.edu.cn, {qingguo.cqg, yaoohu}@alibaba-inc.com, dbwang@email.swu.edu.cn, {tommy.cxd, yongshu.wxb}@alibaba-inc.com, zhangml@seu.edu.cn* (corresponding author)

Abstract

Multi-view multi-label learning serves an important framework to learn from objects with diverse representations and rich semantics. Existing multi-view multi-label learning techniques focus on exploiting shared subspace for fusing multi-view representations, where helpful view-specific information for discriminative modeling is usually ignored. In this paper, a novel multi-view multi-label learning approach named SIMM is proposed which leverages shared subspace exploitation and view-specific information extraction. For shared subspace exploitation, SIMM jointly minimizes confusion adversarial loss and multi-label loss to utilize shared information from all views. For view-specific information extraction, SIMM enforces an orthogonal constraint w.r.t. the shared subspace to utilize view-specific discriminative information. Extensive experiments on real-world data sets clearly show the favorable performance of SIMM against other state-of-the-art multi-view multi-label learning approaches.

1 Introduction

In many real-world applications, objects are with diverse representations and rich semantics simultaneously. For example, in image analysis, a natural scene image can often be represented by its visual features such HSV color histogram, globe feature (Gist) and scale invariant feature transform (SIFT), while it may be annotated with labels {*sky*, *water*, *cloud*}. In video annotation, the representations of a film e.g. *Avengers: Infinity War* are usually from multiple channels of information such as text description, audio, cover picture, frame extraction, meanwhile it can be annotated with *action movie* (type), *America* (country) and *Anthony Russot and Joseph Russo* (directors). The main challenge of such problems is how to integrate the multiple types of heterogeneities in an efficient and general way. Multi-view multi-label learning serves an important framework to solve the above problem.

Formally speaking, let $\mathcal{X} = \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \cdots \times \mathbb{R}^{d_k}$ be the feature space of k views representations, where $d_v (1 \leq v \leq k)$ is the feature dimension of v-th view. Let $\mathcal{Y} = \{y_j\}_{j=1}^{q}$ be the label space with q class labels. Given the training set $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{Y}_i) | 1 \leq i \leq m\}$, where $\boldsymbol{x}_i \in \mathcal{X}$ is the feature vector $[\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \cdots, \boldsymbol{x}_i^k]$ and $\boldsymbol{Y}_i \subseteq \mathcal{Y}$ is the set of relevant labels associated with $\boldsymbol{x}_i$. The task of multi-view multi-label learning is to learn a predictive model $h : \mathcal{X} \rightarrow 2^{\mathcal{Y}}$ from $\mathcal{D}$ which can assign a set of proper labels for the unseen instance.

There have been some attempts for multi-view multi-label problem. In [Xing *et al.*, 2018], the authors introduce a predictive reliability measure to select samples which are used to share label information with other views in the co-training manner. [Zhu *et al.*, 2016; Liu *et al.*, 2015; Luo *et al.*, 2013] focus on the issue of multi-view image classification and map each view into a shared space to eliminate noise and redundancy while maintaining sparse, low-rank and manifold structure of image data respectively. [Zhu *et al.*, 2018; Zhang *et al.*, 2018] aims to further remain consensus on multi-view latent spaces by using Hilbert–Schmidt independence criterion during the mapping procedure.

Nonetheless, the above methods suffer from the following two different issues. Firstly, when trying to find shared information among all views to eliminate noise and redundancy, a common practice is to map each view into a shared subspace. However, this procedure is always done in an independent way, i.e. there is no communication among various views. It is hard to ensure that common semantic information is fully tapped. Secondly, the underlying fact that each view contains their own specific contribution to the multi-label prediction, is ignored. For example, a picture of pink rose may be tagged with two labels {*pink*, *flower*}, while its representation can be described by HSV and Gist. We can easily find the correspondence between them, i.e. *pink* is described by HSV view feature, while *flower* is described by Gist view feature. However, existing methods only try to find the shared information between HSV (color) and Gist (texture), while it is more reasonable to consider extracting their own specific information.

To deal with the above two issues, we present a novel multi-view multi-label neural network framework, which we

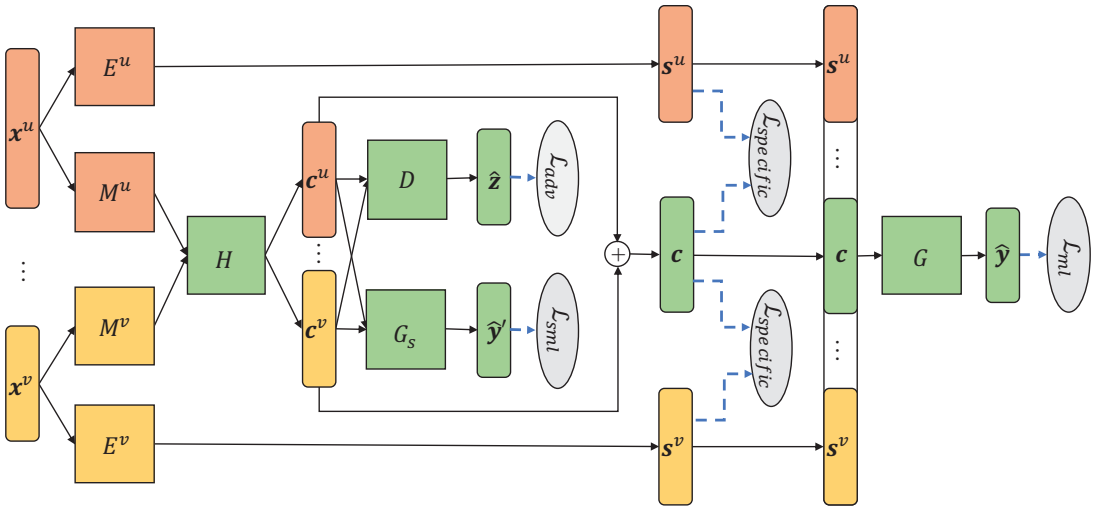


Figure 1: The general flowchart of the proposed SIMM method. Firstly, SIMM jointly minimizes confusion adversarial loss $\mathcal{L}_{adv}$ and multi-label loss $\mathcal{L}_{sml}$ to utilize shared information from all views. Secondly, SIMM enforces an orthogonal constraint $\mathcal{L}_{sml}$ to utilize view-specific discriminative information. Finally, shared and specific information are synergized to characterize semantics.

call SIMM (view-Specific Information extraction for Multi-view Multi-label learning), to leverage shared subspace exploitation and view-specific information extraction. Firstly, SIMM jointly minimizes confusion adversarial loss and multi-label loss to utilize shared information from all views. Secondly, SIMM enforces an orthogonal constraint w.r.t. the shared subspace to utilize view-specific discriminative information. Finally, the shared and view-specific information are synergized to learn the semantics. Extensive experiments on real-world multi-view multi-label data sets clearly show that SIMM achieves highly competitive performance.

The rest of this paper is organized as follows. Section 2 presents technical details of the proposed approach. Section 3 discusses existing works related to SIMM. Section 4 reports detailed results of comparative experiments. Finally, Section 5 concludes.

2 The Proposed Method

The goal of multi-view multi-label learning is to fully integrate various representations of a single object and assign the proper rich semantics to it. As mentioned above, information from different views usually contains the shared and specific parts. Thus, two key steps of SIMM are shared subspace exploitation and view-specific information extraction.

SIMM implements those two key points by using a neural network framework. Considering the network, the overall loss function naturally turns out to be the following form:

$$\mathcal{L} = \mathcal{L}_{ml} + \alpha\mathcal{L}_{shared} + \beta\mathcal{L}_{specific} \tag{1}$$

where α and β are trade-off factors that control the interaction of the loss terms. Final multi-label classification loss $\mathcal{L}_{ml}$ controls the model to predict the output labels. Let y_i^j denote the ground-truth label of $\boldsymbol{x}_i$ on the j-th label. $y_i^j = 1$ if j-th label is the relevance label, $y_i^j = 0$ otherwise. Let $\hat{y}_i^j$ denote the prediction output. The $\mathcal{L}_{ml}$ can be calculated as follows:

$$\mathcal{L}_{ml} = -\sum_{i=1}^{m}\sum_{j=1}^{q} y_i^j \log \hat{y}_i^j + (1-y_i^j)\log(1-\hat{y}_i^j) \tag{2}$$

2.1 Shared Subspace Exploitation

In traditional multi-view learning, the shared subspace exploitation is often implemented in an independent way. Thus, the communication among all views is neglected. Inspired by [Liu *et al.*, 2017], SIMM implements the mapping in a *confusion* way by minimizing the adversarial loss $\mathcal{L}_{adv}$. In particular, SIMM aims to confuse the discriminator that the coming shared subspace representation belongs to which view.

Let $\boldsymbol{c}^v$ be the l dimension shared subspace representation of original v-th view feature $\boldsymbol{x}^v$. It is extracted by a shared subspace extraction layer H, i.e. $\boldsymbol{c}^v = H(M^v(\boldsymbol{x}^v))$, where $M^v(\cdot)$ is used to map each v dimension original feature vector into the same l dimension. Let $\boldsymbol{z}_i$ be the k dimension view label vector of $\boldsymbol{c}_i^v$, where only z_i^v equals to 1 while others equal to 0. This indicates that $\boldsymbol{c}_i^v$ comes from the v-th view. A training set $\mathcal{D}_{adv} = \{(\boldsymbol{c}_i^v, \boldsymbol{z}_i)|1 \le v \le k, 1 \le i \le m\}$ can be constructed for discriminator D. Let $\hat{\boldsymbol{z}}$ be the prediction output by D, i.e. $\hat{\boldsymbol{z}}_i = D(\boldsymbol{c}_i^v)$, $\mathcal{L}_{adv}$ can be formed as:

$$\mathcal{L}_{adv} = \mathcal{F}(-\sum_{i=1}^{m}\sum_{v=1}^{k} z_i^v \log \hat{z}_i^v) \tag{3}$$

where $\mathcal{F}(\cdot)$ should be a monotonically decreasing function[1]. In this way, we confuse the discriminator so as not to recognize the true view of the coming shared subspace representation, which indicates that there is no special information in it. In other words, the incorporated representation $\boldsymbol{c}^v$ only contains the shared information of $\boldsymbol{x}^v$.

However, it might be problematic if only $\mathcal{D}_{adv}$ is used as noise can also confuse the discriminator easily. Therefore, we use $\mathcal{L}_{sml}$ (Shared subspace Multi-Label Loss) to guarantee that $\boldsymbol{c}^v$ contains certain semantics. In particular, a training set $\mathcal{D}_{sml} = \{(\boldsymbol{c}_i^v, \boldsymbol{y}_i)|1 \le v \le k, 1 \le i \le m\}$ can be constructed for shared subspace representation prediction layer

[1] In this paper, $\mathcal{F}(x) = e^{-x}$

Table 1: Characteristic of the real-world multi-view multi-label data sets.

Data Set	$\lvert\mathcal{D}\rvert$	$V(\mathcal{D})$	$VDim(\mathcal{D})$	$CL(\mathcal{D})$	$LCard(\mathcal{D})$	$LDen(\mathcal{D})$	$DL(\mathcal{D})$	$PDL(\mathcal{D})$	Domain
Emotions	593	2	8 / 64	6	1.869	0.311	27	0.046	*music*
Yeast	2,417	2	24 / 79	14	4.237	0.303	198	0.082	*biology*
Corel5k	4,999	4	100 / 512 / 1,000 / 4,096	260	3.397	0.013	2,992	0.599	*image*
EspGame	20,770	4	100 / 512 / 1,000 / 4,096	268	4.686	0.018	18,158	0.874	*image*
Pascal	9,963	5	100 / 512 / 1,000 / 4,096 / 804	20	1.465	0.073	271	0.027	*image*
Mirflickr	25,000	5	100 / 512 / 1,000 / 4,096 / 457	38	4.716	0.142	4,464	0.179	*image*
Youku3w	29,617	4	64 / 128 / 2,048 / 2,048	114	1.188	0.010	1,044	0.035	*video*
Youku15w	148,089	4	64 / 128 / 2,048 / 2,048	114	1.188	0.010	2,074	0.014	*video*

G_s. Let $\hat{\boldsymbol{y}}'$ be the prediction output by G_s, i.e. $\hat{\boldsymbol{y}}_i'^v = G_s(\boldsymbol{c}_i^v)$. $\mathcal{L}_{sml}$ can be formed as:

$$\mathcal{L}_{sml} = -\sum_{i=1}^{m}\sum_{v=1}^{k}\sum_{j=1}^{q} y_i^j \log \hat{y}_i'^{j,v} + (1-y_i^j)\log(1-\hat{y}_i'^{j,v}) \quad (4)$$

The shared subspace exploitation by SIMM is no longer conducted in an independent way as H and D meet the information from all views with G_s maintaining semantics. Combining the above losses together, $\mathcal{L}_{shared}$ can be written as:

$$\mathcal{L}_{shared} = \mathcal{L}_{adv} + \mathcal{L}_{sml} \quad (5)$$

2.2 View-Specific Information Extraction

In SIMM, helpful view-specific information is further considered. Actually, it is difficult to define what is the specific information of a particular view. However, we can extract it from original information by eliminating shared information. This can be implemented by enforcing an orthogonal constraint. Let $\boldsymbol{s}^v$ be the l dimension feature vector extracted by specific information extraction layer E^v, i.e. $\boldsymbol{s}^v = E^v(\boldsymbol{x}^v)$. Let $\boldsymbol{c}$ be the l dimension feature vector containing shared information among all views. $\boldsymbol{c}$ is performed by element-wise addition on each individual $\boldsymbol{c}^v$, i.e. $\boldsymbol{c} = \sum_{v=1}^{k} \boldsymbol{c}^v$. The $\mathcal{L}_{specific}$ encourages orthogonality between $\boldsymbol{s}^v$ and $\boldsymbol{c}$:

$$\mathcal{L}_{specific} = \left\| \boldsymbol{s}^{v\mathsf{T}} \boldsymbol{c} \right\|_2^2 \quad (6)$$

where $\|\cdot\|_2^2$ is the squared L_2-norm. $\mathcal{L}_{specific}$ encourages $\boldsymbol{s}^v$ extracted from the original v-th view feature vector $\boldsymbol{x}^v$ to be as different from $\boldsymbol{c}$ as possible.

2.3 The Whole Framework

The whole framework of SIMM is shown in Figure 1. The goal of training is to minimize the whole loss function (1) with respect to parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_E, \boldsymbol{\theta}_M, \boldsymbol{\theta}_H, \boldsymbol{\theta}_D, \boldsymbol{\theta}_{G_s}, \boldsymbol{\theta}_G\}$. Each unit is a neural network framework and we use popular optimization algorithm Adam [Kingma and Ba, 2015] to do the back propagation to update $\boldsymbol{\Theta}$ simultaneously in the training phase. In the testing phase, given an unseen instance $\boldsymbol{x}^*$, its prediction output $\hat{\boldsymbol{y}}^*$ can be obtained as:

$$\hat{\boldsymbol{y}}^* = G([E^1(\boldsymbol{x}^{*1}), \cdots, E^v(\boldsymbol{x}^{*v}), \cdots, E^k(\boldsymbol{x}^{*k}), \boldsymbol{c}]) \quad (7)$$

where $\boldsymbol{c} = \sum_{v=1}^{k} H(M^v(\boldsymbol{x}^{*v}))$

3 Related Work

In the past years, multi-label learning has been widely used in many kinds of applications, such as text categorization[Ueda and Saito, 2003], bioinformatics [Zhang and Zhou, 2006], web mining[Tang *et al.*, 2009], etc. Following [Zhang and Zhou, 2014], existing multi-label methods can be categorized into two groups, i.e, problem transformation methods and algorithm adaptation methods. Problem transformation methods aim to tackle multi-label learning problem by transforming it into other well-established learning scenarios, such as Binary Relevance [Boutell *et al.*, 2004] transforms the task of multi-label learning into the task of binary classification while Calibrated Label Ranking [Fürnkranz *et al.*, 2008] transforms the task of multi-label learning into the task of label ranking. Algorithm adaptation methods tackle multi-label learning problem by adapting popular learning techniques to deal with multi-label data directly, such as ML-kNN [Zhang and Zhou, 2007] adapts lazy learning technique while ML-DT [Clare and King, 2001] adapts decision tree technique.

In multi-view learning, the information in some views can help handle the weakness of other views. Furthermore, multi-view learning can be embedded into multi-label learning tasks naturally to further improve the classification performance. The most important part in multi-view multi-label learning is how to utilize and communicate heterogeneous information among all views under the multi-label framework. [Xing *et al.*, 2018] focuses on selecting reliable samples from one view and passing them to other views. However, the communication among views is only on the label level. [Zhu *et al.*, 2016] introduces a block-row regularizer to reduce the noise and redundancy. The method in [Liu *et al.*, 2015] seeks a low-dimensional common representation of all views and constructs the classifier based on matrix completion. [Luo *et al.*, 2013] integrates multiple features under multi-view vector-valued manifold regularization. [Zhu *et al.*, 2018; Zhang *et al.*, 2018] try to remain latent semantic when studying the low-dimensional common representation.

However, for most existing methods the mapping from original views to the shared subspace is conducted in an independent way where there is only limited communication among each view. Furthermore, the specific characteristics of individual view is ignored. SIMM serves as the first attempt towards jointly enhancing communication during shared subspace exploitation and remaining view-specific information.

Table 2: Predictive performance of each comparing algorithm (mean ± std. deviation) on the multi-view multi-label data sets.

Comparing Methods	Hamming Loss ↓							
	Emotions	Yeast	Corel5k	EspGame	Pascal	Mirflickr	Youku3w	Youku15w
SIMM	**0.178±0.020**	**0.191±0.010**	**0.011±0.000**	**0.017±0.000**	**0.044±0.001**	**0.079±0.001**	**0.008±0.000**	**0.008±0.000**
Benchmark	0.189±0.018	0.194±0.011	**0.011±0.000**	**0.017±0.000**	0.049±0.002	0.083±0.001	0.009±0.000	**0.008±0.000**
COMMON	0.189±0.006	0.200±0.007	0.013±0.000	**0.017±0.000**	0.066±0.001	0.104±0.002	0.010±0.000	0.010±0.000
ML-KNN (C)	0.193±0.012	0.195±0.009	0.012±0.000	**0.017±0.000**	0.064±0.002	0.110±0.001	0.009±0.001	0.009±0.000
ML-KNN (B)	0.200±0.016	0.208±0.008	0.012±0.000	**0.017±0.000**	0.055±0.001	0.109±0.001	**0.008±0.000**	**0.008±0.000**
LSAMML	0.284±0.019	0.298±0.005	0.013±0.000	**0.017±0.000**	0.064±0.001	0.107±0.001	0.009±0.000	0.009±0.000
F2L21F	0.225±0.024	0.315±0.012	0.017±0.000	0.025±0.000	0.091±0.004	0.129±0.003	0.015±0.000	0.015±0.000
Comparing Methods	Average Precision ↑							
	Emotions	Yeast	Corel5k	EspGame	Pascal	Mirflickr	Youku3w	Youku15w
SIMM	**0.831±0.036**	0.773±0.011	0.549±0.011	**0.396±0.005**	**0.793±0.008**	**0.774±0.006**	**0.687±0.008**	**0.728±0.002**
Benchmark	0.822±0.029	**0.774±0.012**	**0.559±0.008**	0.395±0.005	0.763±0.009	0.754±0.005	0.608±0.005	0.682±0.003
COMMON	0.825±0.029	0.757±0.014	0.382±0.017	0.305±0.005	0.580±0.027	0.677±0.005	0.520±0.037	0.639±0.004
ML-KNN (C)	0.799±0.032	0.764±0.012	0.441±0.010	0.288±0.004	0.571±0.009	0.607±0.004	0.410±0.007	0.460±0.003
ML-KNN (B)	0.795±0.020	0.753±0.009	0.416±0.009	0.270±0.004	0.658±0.011	0.608±0.004	0.634±0.006	0.672±0.003
LSAMML	0.779±0.040	0.611±0.013	0.475±0.014	0.346±0.005	0.690±0.012	0.674±0.004	0.637±0.006	0.656±0.003
F2L21F	0.798±0.030	0.607±0.016	0.314±0.013	0.316±0.007	0.644±0.019	0.625±0.014	0.630±0.006	0.661±0.002
Comparing Methods	One Error ↓							
	Emotions	Yeast	Corel5k	EspGame	Pascal	Mirflickr	Youku3w	Youku15w
SIMM	**0.224±0.057**	0.221±0.020	**0.349±0.014**	**0.452±0.011**	**0.251±0.013**	**0.173±0.009**	**0.423±0.008**	**0.374±0.003**
Benchmark	0.241±0.042	**0.215±0.020**	**0.349±0.007**	0.463±0.011	0.288±0.014	0.195±0.010	0.516±0.008	0.430±0.005
COMMON	0.233±0.046	0.241±0.021	0.519±0.036	0.544±0.012	0.528±0.027	0.236±0.013	0.608±0.036	0.477±0.005
ML-KNN (C)	0.270±0.058	0.228±0.025	0.477±0.025	0.601±0.010	0.499±0.016	0.355±0.007	0.679±0.008	0.630±0.004
ML-KNN (B)	0.282±0.029	0.233±0.022	0.506±0.017	0.652±0.009	0.367±0.014	0.360±0.009	0.456±0.008	0.413±0.004
LSAMML	0.316±0.064	0.363±0.030	0.418±0.019	0.496±0.011	0.358±0.014	0.236±0.007	0.469±0.008	0.454±0.005
F2L21F	0.277±0.058	0.361±0.039	0.598±0.016	0.543±0.011	0.401±0.019	0.301±0.013	0.470±0.007	0.446±0.004
Comparing Methods	Ranking Loss ↓							
	Emotions	Yeast	Corel5k	EspGame	Pascal	Mirflickr	Youku3w	Youku15w
SIMM	**0.134±0.031**	**0.160±0.007**	**0.053±0.004**	**0.109±0.003**	**0.061±0.003**	**0.065±0.002**	**0.020±0.001**	**0.015±0.000**
Benchmark	0.140±0.026	0.161±0.008	0.054±0.004	**0.109±0.002**	0.074±0.004	0.072±0.002	0.031±0.001	0.020±0.001
COMMON	0.139±0.024	0.169±0.012	0.117±0.012	0.158±0.002	0.139±0.015	0.100±0.002	0.045±0.009	0.025±0.001
ML-KNN (C)	0.167±0.027	0.168±0.008	0.105±0.006	0.170±0.002	0.179±0.007	0.136±0.002	0.161±0.005	0.137±0.002
ML-KNN (B)	0.171±0.020	0.173±0.008	0.112±0.007	0.170±0.002	0.165±0.007	0.133±0.003	0.063±0.003	0.054±0.001
LSAMML	0.185±0.040	0.345±0.012	0.133±0.006	0.208±0.005	0.139±0.009	0.174±0.003	0.041±0.002	0.031±0.001
F2L21F	0.165±0.024	0.350±0.013	0.275±0.013	0.250±0.013	0.178±0.018	0.206±0.016	0.050±0.002	0.033±0.001
Comparing Methods	Coverage ↓							
	Emotions	Yeast	Corel5k	EspGame	Pascal	Mirflickr	Youku3w	Youku15w
SIMM	**0.272±0.027**	**0.442±0.011**	**0.133±0.008**	**0.285±0.005**	**0.099±0.003**	**0.230±0.004**	**0.024±0.002**	**0.019±0.000**
Benchmark	0.277±0.028	0.444±0.012	0.135±0.006	**0.285±0.004**	0.115±0.005	0.243±0.004	0.035±0.001	0.024±0.001
COMMON	0.276±0.025	0.445±0.012	0.281±0.027	0.390±0.005	0.186±0.017	0.298±0.004	0.051±0.010	0.030±0.001
ML-KNN (C)	0.299±0.021	0.450±0.012	0.249±0.013	0.419±0.005	0.233±0.008	0.346±0.004	0.178±0.006	0.152±0.002
ML-KNN (B)	0.303±0.021	0.455±0.008	0.262±0.014	0.415±0.004	0.222±0.006	0.340±0.006	0.074±0.003	0.063±0.001
LSAMML	0.315±0.030	0.623±0.011	0.327±0.013	0.523±0.009	0.202±0.010	0.428±0.006	0.049±0.002	0.037±0.001
F2L21F	0.301±0.024	0.627±0.011	0.559±0.020	0.578±0.020	0.240±0.022	0.459±0.022	0.060±0.002	0.040±0.001
Comparing Methods	Micro-F1 ↑							
	Emotions	Yeast	Corel5k	EspGame	Pascal	Mirflickr	Youku3w	Youku15w
SIMM	**0.694±0.043**	**0.655±0.019**	0.419±0.014	0.236±0.009	**0.632±0.011**	**0.629±0.005**	0.471±0.010	**0.529±0.005**
Benchmark	0.681±0.036	0.649±0.019	**0.444±0.012**	0.255±0.007	0.609±0.012	0.615±0.006	0.380±0.010	0.495±0.006
COMMON	0.665±0.016	0.616±0.014	0.024±0.009	0.022±0.003	0.235±0.025	0.323±0.019	0.101±0.030	0.180±0.009
ML-KNN (C)	0.668±0.026	0.639±0.016	0.259±0.011	0.084±0.006	0.327±0.019	0.367±0.008	0.242±0.008	0.290±0.005
ML-KNN (B)	0.652±0.030	0.608±0.013	0.226±0.013	0.069±0.003	0.447±0.019	0.379±0.015	**0.472±0.006**	0.524±0.004
LSAMML	0.185±0.067	0.035±0.008	0.146±0.014	0.072±0.005	0.259±0.013	0.268±0.005	0.239±0.008	0.236±0.005
F2L21F	0.651±0.038	0.465±0.020	0.278±0.014	**0.291±0.006**	0.471±0.017	0.493±0.009	0.448±0.006	0.471±0.002

4 Experiment

4.1 Experimental Setting

Data Sets

A total of eight multi-view multi-label data sets are employed for performance evaluation including six benchmark data sets[2] and two real-world video annotation data sets.

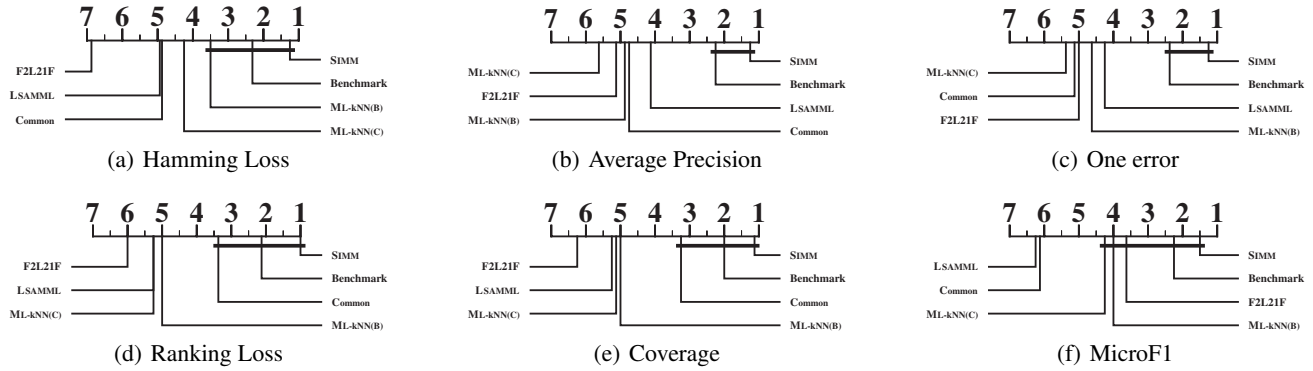
`Emotions` has two feature views: 8 rhythmic attributes and 64 timbre attributes. `Yeast` also has two views including the concatenation of the genetic expression (79 attributes) and the phylogenetic profile of a gene (24 attributes). For `Corel5k` [Duygulu *et al.*, 2002] and `EspGame` [Von Ahn and Dabbish, 2004], we choose four representative feature views: DenseHue, Gist, DenseSift, HSV, whose dimensions are 100, 512, 1000 and 4096 respectively. For `Pascal` [Everingham *et al.*, 2010] and `Mirflickr` [Huiskes and Lew, 2008], Tag view is further utilized as the textual view besides the above four views [Guillaumin *et al.*, 2010], whose dimensions are 804 and 457.

`Youku` data set consists 148,089 videos in 114 classes collected from the Video Application YOUKU. Each object has four views: textual description in the video title, the audio information, cover picture and video frame extraction. The dimensions are 100, 128, 2048, 2048 respectively. For textual view, we extract the weights of each word in the title.

[2]Publicly available at http://mulan.sourceforge.net and http://lear.inrialpes.fr/people/guillaumin/data.php

Table 3: Friedman statistics F_F in terms of each evaluation metric and the critical value at 0.05 significance level (# comparing algorithms k = 7, # data sets N = 8).

Evaluation metric	F_F	critical value
Hamming Loss	20.2930	2.3240
Average Precision	9.1649	
One error	7.8626	
Ranking Loss	21.1260	
Coverage	24.2040	
Micro-F1	15.0070	



(a) Hamming Loss (b) Average Precision (c) One error

(d) Ranking Loss (e) Coverage (f) MicroF1

Figure 2: Comparison of SIMM (control algorithm) against other comparing algorithms with the Bonferroni-Dunn test. Algorithms not connected with SIMM in the CD diagram are considered to have significantly different performance from the control algorithm (CD=3.1853 at 0.05 significance level).

Using the pre-trained word embedding, title embedding vector can be obtained by weighting the word embedding. For audio view, we follow [Gemmeke *et al.*, 2017] and reduce dimension by using Principle Component Analysis (PCA). For image view, we use Inception-v3 proposed in [Szegedy *et al.*, 2016]. For video view, we extract video frames every two seconds and use the same procedure as image view to obtain frame-level features. Then, these frame-level features are aggregated into video-level by mean-pooling and L_2-normalization. Label information is obtained by manual labeling, from which we choose the most frequent 114 ones.

Given the data set $\mathcal{D}$, we use $|\mathcal{D}|$, $V(\mathcal{D})$, $VDim(\mathcal{D})$, $CL(\mathcal{D})$ to represent its *number of samples*, *number of views*, *dimension of each view*, *class labels*. Moreover, several other properties of multi-label data sets are denoted as [Zhang and Zhou, 2014]:

Label cardinality: $LCard(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{q}\mathbb{I}(y_{ij} = +1)$ which counts the average number of relevant labels per example;

Label density: $LDen(D) = \frac{LCard(\mathcal{D})}{CL(\mathcal{D})}$ which normalizes label cardinality by the number of class labels;

Distinct label sets: $DL(\mathcal{D}) = |\{\boldsymbol{y}|\exists \boldsymbol{x} : (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}\}|$ which counts the number of distinct label vectors (relevant label set) existing in $\mathcal{D}$;

Proportion of distinct label sets: $PDL(\mathcal{D}) = \frac{DL(\mathcal{D})}{|\mathcal{D}|}$ which normalizes $DL(\mathcal{D})$ by the number of examples.

Table 1 summarizes detailed characteristics of the experimental data sets.

Comparing Algorithms

The performance of SIMM is compared against six algorithms, including two baselines related to SIMM, a multi-label algorithm ML-KNN with two types of feature inputs and two multi-view multi-label algorithms.

Benchmark: A multi-layer perceptron where the number of parameters is no less than SIMM. The input is the concatenation of all view features.

COMMON: The prediction results are only obtained by shared information. i.e. $\hat{\boldsymbol{y}}^{*\prime} = \frac{1}{k}\sum_{v=1}^{k}\hat{\boldsymbol{y}}^{*\prime v}$

ML-KNN(C) & ML-KNN(B) [Zhang and Zhou, 2007]: A lazy multi-label algorithm with two types feature inputs. (C) stands for the concatenation of all views. (B) stands for the best performance obtained from the best single view. [configuration: $k = 10$]

F2L21F [Zhu *et al.*, 2016]: A multi-view multi-label algorithm using block-row regularizer to reduce the overlap, noise, and redundancy in multiple views. [configuration: $\lambda_1 = 10, \lambda_2 = 10$]

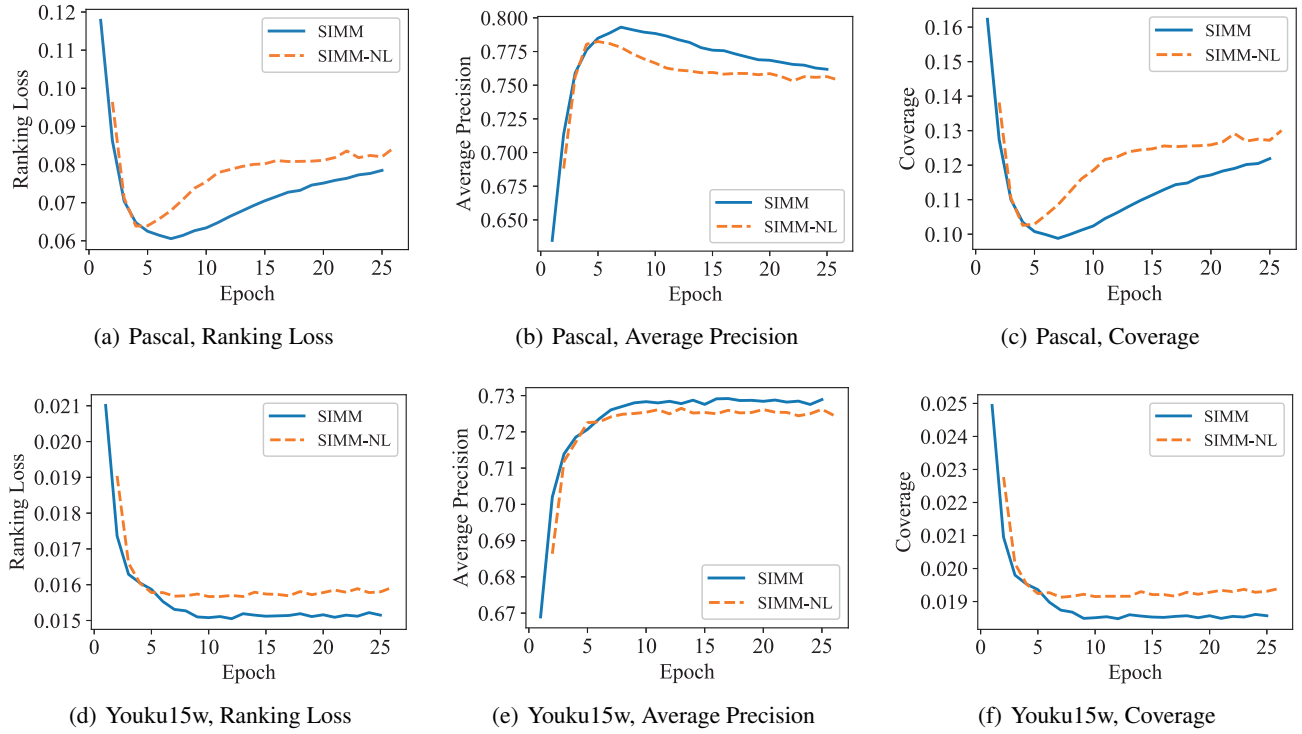
LSAMML [Zhang *et al.*, 2018]: A multi-view multi-label algorithm using Hilbert-Schmidt Independence Criterion to enhance the dependence between different pairs of views. [configuration: $r \in \{2, 3, 4, 5\}$, β and $\gamma \in \{0.01, 0.1, 1, 10, 100\}$]

For **SIMM**, in order to make the model more elegant and lightweight, we set each module to be only a fully connected layer without hidden layer. l is fixed to 64. In light of comparison to COMMON, α is fixed to 1. β is selected from $\{0.1, 0.01, 0.001, 0.0001\}$.

Experimental Protocol

In this paper, six widely-used multi-label metrics are employed for performance evaluation, including *Hamming Loss*, *Average Precision*, *One Error*, *Ranking Loss*, *Coverage* and *Micro-F1*, which consider the performance of multi-label predictor from various aspects. Concrete metric definitions can be found in [Zhang and Zhou, 2014], and *Coverage* metric is normalized by the number of class labels (i.e. q). For *Average Precision* and *Micro-F1*, the larger the values the better the performance. For the other four metrics, the smaller the values the better the performance.

For each data set, ten-fold cross-validation is performed where the mean metrics results and standard deviations are recorded for all comparing approaches.



(a) Pascal, Ranking Loss (b) Pascal, Average Precision (c) Pascal, Coverage

(d) Youku15w, Ranking Loss (e) Youku15w, Average Precision (f) Youku15w, Coverage

Figure 3: Comparison of SIMM against no constraints version on Pascal and Youku15w.

4.2 Experimental Results

Detailed experimental results are reported in Table 2, where the best performance on each data set is shown in boldface. To analyze the relative performance among the comparing approaches in a systematic manner, *Friedman test* is employed as the statistical test for performance comparison.

Table 3 reports Friedman statistics F_F and the corresponding critical values in terms of each evaluation metric. It is obvious that the null hypothesis of equal performance is rejected at 0.05 significance level. Accordingly, post-hoc Boferroni-Dunn test is performed to compare the relative performance among the comparing approaches. The CD diagrams are presented in Figure 2, where the average rank of each approach is marked along the axis (the smaller the better).

Based on the reported experimental results, the following observations can be made: (a) Among the 48 configurations (8 data sets $\times$ 6 evaluation metrics), SIMM ranks 1st and 2nd in 87.5% and 10.4% cases respectively. (b) It is remarkable that SIMM achieves best performance in all cases on *Hamming Loss*, *Ranking Loss* and *Coverage*. (c) It is noticeable that SIMM outperforms all the comparing algorithms on `Emotions`, `Pascal`, `Mirflickr` and `Youku15w` on all metrics while outperforms all the comparing algorithms on `Espgame` and `Youku3w` on all metrics except on *Micro-F1*.

4.3 Further Analysis

In order to examine the effectiveness of $\mathcal{L}_{shared}$ and $\mathcal{L}_{specific}$, we remain the basic structure of our model and remove these two constraints by setting α and β to zero. Figure 3 shows the comparison results on `Pascal` and `Youku15w` data sets. The curves show the change of the metric results with the number of epoch. It is clearly shown that the model performed worse without $\mathcal{L}_{shared}$ and $\mathcal{L}_{specific}$ (SIMM-NL). In other words, it is not enough to fuse the information from different views directly. SIMM is a good help to separate the shared and specific information.

5 Conclusion

In this paper, a novel neural network based approach is proposed to solve the multi-view multi-label problem. Specifically, we enhance the communication among views while remain individual specific characteristics through two steps, i.e. shared subspace exploitation and view-specific information extraction. Effectiveness of the proposed approach is validated via comprehensive experiments on real-world multi-view multi-label data sets. In the future, it is interesting to extend the structure of each neural network module. Meanwhile, note that $\boldsymbol{s}^v$ and $\boldsymbol{c}$ are fixed with the same dimension due to the orthogonality constraint and therefore a more general way of specific information extraction can be explored.

References

[Boutell *et al.*, 2004] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[Clare and King, 2001] Amanda Clare and Ross D King. Knowledge discovery in multi-label phenotype data. In *Lecture Notes in Computer Science 2168*, pages 42–53. Springer, Berlin, 2001.

[Duygulu *et al.*, 2002] Pinar Duygulu, Kobus Barnard, Joao FG de Freitas, and David A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Lecture Notes in Computer Science 2353*, pages 97–112. Springer, Berlin, 2002.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[Fürnkranz *et al.*, 2008] Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. Multilabel classification via calibrated label ranking. *Machine learning*, 73(2):133–153, 2008.

[Gemmeke *et al.*, 2017] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780, New Orleans, LA, 2017.

[Guillaumin *et al.*, 2010] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 902–909, San Francisco, CA, 2010.

[Huiskes and Lew, 2008] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, Vancouver, Canada, 2008.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, 2015.

[Liu *et al.*, 2015] Meng Liu, Yong Luo, Dacheng Tao, Chao Xu, and Yonggang Wen. Low-rank multi-view learning in matrix completion for multi-label image classification. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2778–2784, Austin, TX, 2015.

[Liu *et al.*, 2017] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1–10, 2017.

[Luo *et al.*, 2013] Yong Luo, Dacheng Tao, Chang , Chao Xu, Hong Liu, and Yonggang Wen. Multiview vector-valued manifold regularization for multilabel image classification. *IEEE transactions on neural networks and learning systems*, 24(5):709–722, 2013.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[Tang *et al.*, 2009] Lei Tang, Suju Rajan, and Vijay K Narayanan. Large scale multi-label classification via metalabeler. In *Proceedings of the 18th international conference on World wide web*, pages 211–220, Madrid, Spain, 2009.

[Ueda and Saito, 2003] Naonori Ueda and Kazumi Saito. Parametric mixture models for multi-labeled text. In *Advances in Neural Information Processing Systems 15*, pages 737–744. MIT Press, Cambridge, MA, 2003.

[Von Ahn and Dabbish, 2004] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 319–326, Vienna, Austria, 2004.

[Xing *et al.*, 2018] Yuying Xing, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang. Multi-label cotraining. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2882–2888, Stockholm, Sweden, 2018.

[Zhang and Zhou, 2006] Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.

[Zhang and Zhou, 2007] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[Zhang and Zhou, 2014] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.

[Zhang *et al.*, 2018] Changqing Zhang, Ziwei Yu, Qinghua Hu, Pengfei Zhu, Xinwang Liu, and Xiaobo Wang. Latent semantic aware multi-view multi-label classification. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 4414–4421, New Orleans, LA, 2018.

[Zhu *et al.*, 2016] Xiaofeng Zhu, Xuelong Li, and Shichao Zhang. Block-row sparse multiview multilabel learning for image classification. *IEEE transactions on cybernetics*, 46(2):450–461, 2016.

[Zhu *et al.*, 2018] Pengfei Zhu, Qi Hu, Qinghua Hu, Changqing Zhang, and Zhizhao Feng. Multi-view label embedding. *Pattern Recognition*, 84:126–135, 2018.