ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200278

Common and Discriminative Semantic Pursuit for Multi-Modal Multi-Label Learning

Yi Zhang and **Jundong Shen** and **Zhecheng Zhang** and **Chongjun Wang**[1]

Abstract. Multi-modal multi-label (MMML) learning provides an important framework to learn complex objects with diverse representations and annotations. Most existing multi-modal multi-label learning approaches focus on exploiting shared information of all modalities, but neglect specific information of each modality. Besides, how to effectively utilize relationship among modalities is also a challenging issue. In this paper, we propose a novel MMML learning approach called Common and Discriminative Semantic Pursuit (CoDiSP), which learns low-dimensional common representation with all modalities, and extracts discriminative information of each modality by enforcing orthogonal constraint. Meanwhile, the common representation is used as a new modality and added to the specific modal sequence. Furthermore, CoDiSP learns deep models with adaptive depth and exploits label correlations simultaneously based on the extracted modal sequence. Finally, extensive experiments on several benchmark MMML datasets show superior performance of CoDiSP compared with other state-of-the-art approaches.

1 INTRODUCTION

On the one hand, multi-modal learning has attracted much attention with the rapid development of data collection technology, where data collected from diverse sources can be represented with multiple modal features. In contrast to single modal learning, multi-modal learning mainly exploits the consistent and complementary properties among different modalities and improves the learning performance [26]. The major challenge of multi-modal problem is how to jointly model heterogeneous modalities in a mutually beneficial way.

On the other hand, multi-label learning [33] has been an important and practical research topic, since one instance may have multiple annotations. As a fundamental framework handling objects with multiple annotations, multi-label learning has been widely applied in many real-world applications, such as image annotation [24], document categorization [14], information retrieval [6] and bioinformatics [31]. A challenging issue for multi-label learning is the exploitation of label correlations [5].

In real-world applications, data are often represented with multiple modalities and associated with multiple labels simultaneously. Multi-modal multi-label (MMML) [29] learning provides a learning framework for modeling such complex objects. Existing MMML learning approaches mainly focus on exploiting shared information among all modalities to eliminate noise and redundancy, which map each modality into a shared subspace. However, it is usually done in an independent way, which neglects communication among different modalities. To overcome the above challenges, we present a novel end-to-end multi-modal multi-label neural network framework named Common and Discriminative Semantic Pursuit (CoDiSP), to exploit relationship among different modalities and label correlations with the help of extracted common and specific modal features. The main contributions of this paper are summarized as follows:

- A novel MMML learning approach called Common and Discriminative Semantic Pursuit (CoDiSP) is proposed, which can exploit common and discriminative information, and make full use of inter-relationship among different modalities.
- Without pre-determining the modal order for prediction, our approach is able to sequentially learn the modal dependency and label correlations with the extracted modal sequence.
- Experiments on benchmark datasets demonstrate that CoDiSP performs favorably against state-of-the-art approaches. Besides, we conduct extensive experiments to analyze common semantic, modal dependency and convergence of CoDiSP.

The remainder of this paper is organized as follows. Section 2 briefly reviews some related work of multi-modal multi-label learning. Section 3 presents technical details of the proposed approach. Section 4 reports detailed results of comparative experiments. Finally, Section 5 gives the conclusion.

2 RELATED WORK

In this section, we will briefly review some state-of-the-art approaches in both multi-modal and multi-label learning.

For multi-label learning, each instance is associated with multiple interdependent labels and the goal is to exploit various types of label correlations. In terms of the order of label correlations, these approaches can be divided into three strategies. For first-order approach, Binary Relevance (BR) [1] takes each label independently, which neglects the relationship among labels. For second-order approach, Calibrated Label Ranking (CLR) [4] considers the pairwise relationships between labels. For high-order approach, Classifier Chains (CC) [16] addresses connections among random subsets of labels. So far, many approaches have been developed to improve the performance of multi-label learning by exploring various types of label correlations [11] [35]. For example, a boosting approach [10] exploits label correlations with a hypothesis reuse mechanism. However most of the existing approaches take label correlations as prior knowledge [9], which may not correctly characterize the real relationships among labels and final predictions are not explicitly correlated. To tackle this problem, CAMEL [3] learns the high-order label correlations via sparse reconstruction in the label space.

[1] National Key Laboratory for Novel Software Technology at Nanjing University, Nanjing University, Nanjing 210023, China, email: {njuzhangy, jdshen, zzc}@smail.nju.edu.cn, chjwang@nju.edu.cn

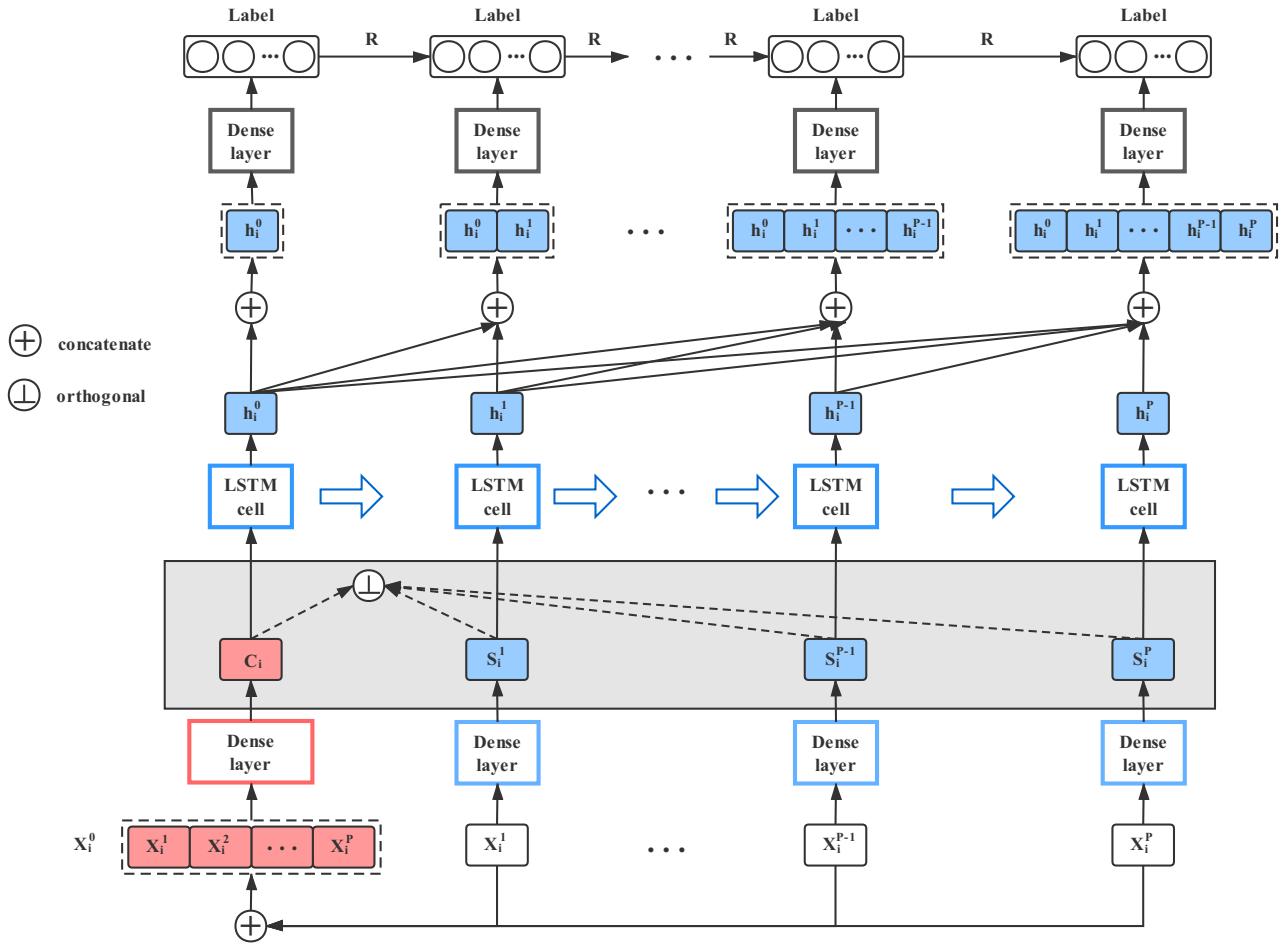


Figure 1. The overall flowchart of the proposed CoDiSP approach. Firstly, for i-th instance $\boldsymbol{X}_i = [\boldsymbol{X}_i^1, \boldsymbol{X}_i^2, \cdots, \boldsymbol{X}_i^P]$ with P modalities, we concatenate all modalities as raw common modality $\boldsymbol{X}_i^0$. Secondly, CoDiSP enforces orthogonal constraint to exploit common information $\boldsymbol{C}_i$ and modal-specific discriminative information $\{\boldsymbol{S}_i^m\}_{m=1}^p$. Meanwhile, $\boldsymbol{C}_i$ is added to the specific modal sequence as a new modal, and then input the new modal sequence $\{\boldsymbol{C}_i, \boldsymbol{S}_i^1, \boldsymbol{S}_i^1, \cdots, \boldsymbol{S}_i^P\}$ to LSTM structure in order. At t-th step, we stack all previous hidden output as $\boldsymbol{H}_i^t = [\boldsymbol{h}_i^0, \boldsymbol{h}_i^1, \cdots, \boldsymbol{h}_i^t]$, and exploit label correlation $\boldsymbol{R}$ based on current modal features. Finally, we make label prediction with stacked output $\boldsymbol{H}_i^t$, label correlation matrix $\boldsymbol{R}$ and prediction at $(t-1)$-th step.

For multi-modal learning, the goal is to improve performance with heterogeneous modalities or reduce the sample complexity with accumulated multi-modal data [27]. Directed Acyclic Graph (DAG) [20] models an adaptive modalities acquisition system, which learns decision rules that adaptively select modalities for each example as necessary to make a confident prediction. Nevertheless, DAG needs to list all permutations of modal sequence in the training phase. Discriminative Modal Pursuit (DMP) [28] approach is proposed to learn a serialized modal extraction decision methods, which can balance the classification performance and modal feature extraction cost.

What's more, there have been some researches for multi-modal multi-label learning. [2] proposes a new classification framework using the multi-label correlation information to address the problem of simultaneously combining multiple feature modalities and maximum margin classification. [8] performs joint label specific feature selection and take the label correlation matrix as prior knowledge for model training. Considering that label heterogeneity and feature heterogeneity often co-exist, [25] proposes a novel graph-based model for Learning with both Label and Feature heterogeneity (L^2F), which imposes the modal consistency by requiring that modal-based classifiers generate similar predictions on the same examples. Multi-Label Co-Training (MLCT) [23] introduces a predictive reliability measure to select samples, and applies label-wise filtering to confidently communicate labels of selected samples among co-training classifiers. To sufficiently consider the complementary information among multiple modals, LSA-MML [30] is proposed to seek a predictive common representation of multiple modals and the corresponding projection model between the common representation and labels. To further fully extract the complementarity and correlation information effectively, SMISFL [21] jointly learns multiple modal-individual transformations and one sharable transformation. [35] aims to reamin concensus on mutli-modal latent spaces by Hilbert-Schmidt independence criterion during the mapping procedure. However, there is no communication among various modalities. Hence, SIMM [22] is proposed to leverage shared subspace exploitation and view-specific information extraction. Nevertheless, previous approaches rarely consider the label correlation. CS3G approach [29] handles types of interactions between multiple labels, while no interaction between features from different modalities in the model training phase. To make each modality interacts and further reduce modal extraction cost, Multi-modal Classifier Chains (MCC) [34] extends Classifier Chains to exploit label correlations with partial modalities.

3 METHODOLOGY

This section mainly gives the detail description of Common and Discriminative Semantic Pursuit (CoDiSP) approach after a preliminary notation explanation.

3.1 Notation

Formally, let $\mathcal{X} = \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_m} \times \cdots \times \mathbb{R}^{d_P}$ be the feature space of P modalities, where $d_m (1 \leq m \leq P)$ is the dimensionality of the m-th modal. Let $\mathcal{Y} = \{y_k\}_{k=1}^{L}$ be the label space with L labels. Given the training dataset with N data samples $\mathcal{D} = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1}^{N}$, where $\boldsymbol{X}_i = [\boldsymbol{X}_i^1, \boldsymbol{X}_i^2, \cdots, \boldsymbol{X}_i^P] \in \mathcal{X}$ is the feature vector and $\boldsymbol{Y}_i \subseteq \mathcal{Y}$ is the label vector of the i-th instance $\boldsymbol{X}_i$. The task of multi-modal multi-label learning is to learn a function $\boldsymbol{F} : \mathcal{X} \to 2^{\mathcal{Y}}$ from $\mathcal{D}$, which can assign a set of proper labels for the unseen instance.

3.2 CoDiSP approach

We introduce Common and Discriminative Semantic Pursuit (CoDiSP) approach in detail. As illustrated in Fig. 1, CoDiSP approach contains 4 major components: feature mapping layer, orthogonal layer, LSTM inference layer and label prediction layer.

3.2.1 Feature mapping layer

Considering the dimensionality of different modalities is heterogeneous which is difficult to input to the network, we are supposed to map the original modal feature vectors to the same dimension.

Common modality mapping Aiming at exploiting common subspace representation of all modalities, we concatenate all modalities in $\boldsymbol{X}_i$ to formalize a new single modal $\boldsymbol{X}_i^0 = [\boldsymbol{X}_i^1, \boldsymbol{X}_i^2, \cdots, \boldsymbol{X}_i^P] \in \mathbb{R}^{d_{all}}$, where $d_{all} = d_1 + d_2 + \cdots + d_P$. And then we add a dense layer to transform the original common modality to d dimension common vector according to Eq. 1.

$$\boldsymbol{C}_i = ReLU(\boldsymbol{X}_i^0 \boldsymbol{U}_0 + \boldsymbol{b}_0) \tag{1}$$

where $\boldsymbol{U}_0 \in \mathbb{R}^{d_{all} \times d}$ is weight vector, $\boldsymbol{b}_0 \in \mathbb{R}^{1 \times d}$ is bias vector.

Specific modality mapping Each modality in the original feature vector $\boldsymbol{X}_i$ can be used to extract specific information. And then we add a dense layer to transform the m-th original specific modality $\boldsymbol{X}_i^m$ with d_m dimension to d dimension $\boldsymbol{S}_i^m$ by Eq. 2.

$$\boldsymbol{S}_i^m = ReLU(\boldsymbol{X}_i^m \boldsymbol{U}_m + \boldsymbol{b}_m), m = 1, \cdots, P \tag{2}$$

where $\boldsymbol{U}_m \in \mathbb{R}^{d_m \times d}$ is weight vector, $\boldsymbol{b}_m \in \mathbb{R}^{1 \times d}$ is bias vector.

As a result, for the i-th instance $\boldsymbol{X}_i$, feature mapping layer is used to map common and specific modal feature vector into the same d dimension modal feature vector $\boldsymbol{C}_i$ and $\{\boldsymbol{S}_i^m\}_{m=1}^{P}$, respectively.

3.2.2 Orthogonal layer

Different modalities often includes two main different information: common information shared among all the modalities, and discriminative information of its own specific modality. Therefore, we will introduce our CoDiSP approach based on the following two parts.

Common semantic pursuit For traditional single modal multi-label learning approach, concatenating all modalities can obtain better performance than those of the best single modal. As a result, we seek a predictive common representation from concatenated modalities. Furthermore, the corresponding prediction model between common representation and multiple labels can be represented by Eq. 3.

$$\boldsymbol{f}(\boldsymbol{C}_i) = \sigma(\boldsymbol{C}_i \boldsymbol{W}_C + \boldsymbol{b}_C) \tag{3}$$

where $\boldsymbol{W}_C \in \mathbb{R}^{d \times L}$ is weight vector, $\boldsymbol{b}_C \in \mathbb{R}^{1 \times L}$ is the bias vector.

And thus, the common loss function implied in the common features prediction model can be represented as:

$$\mathcal{L}_{comm} = -\sum_{i=1}^{N_b} \sum_{k=1}^{L} \left(Y_i^k log Z_i^k + (1 - Y_i^k) log(1 - Z_i^k) \right) \tag{4}$$

where Y_i^k is the ground-truth of $\boldsymbol{X}_i$ on the k-th label. $Y_i^k = 1$ if k-th label is the relevance label, 0 otherwise. $Z_i^k = \boldsymbol{f}(\boldsymbol{C}_i)$ is the prediction with the common modal vector $\boldsymbol{C}_i$, which is extracted from original modal feature vector $\boldsymbol{X}_i$. And N_b is batch size.

Discriminative semantic pursuit Complementary information among different modalities is of great importance. In order to jointly enhance communication between common modality $\boldsymbol{X}_i^0$ and other specific modality $\{\boldsymbol{X}_i^m\}_{m=1}^{P}$, we need to extract specific modal features from $\{\boldsymbol{S}_i^m\}_{m=1}^{P}$, which eliminate the shared information with common information. We penalize the independence between common modal vector $\boldsymbol{C}_i$ and each specific modal vector $\{\boldsymbol{S}_i^m\}_{m=1}^{P}$ with orthogonal loss function:

$$\mathcal{L}_{orth} = \sum_{i=1}^{N_b} \sum_{m=1}^{P} \| \boldsymbol{C}_i^T \boldsymbol{S}_i^m \|_2^2 \tag{5}$$

where $\| \cdot \|_2$ is the L_2-norm. $\mathcal{L}_{orth}$ encourages $\boldsymbol{S}_i^m$ extracted from the original m-th modal vector $\boldsymbol{X}_i^m$ to be as discriminative from $\boldsymbol{C}_i$ as possible.

3.2.3 LSTM inference layer

After preparing common modal features $\boldsymbol{C}_i$ and discriminative modal features $\{\boldsymbol{S}_i^m\}_{m=1}^{P}$, which are all in the same d dimension, we input $\{\boldsymbol{C}_i, \boldsymbol{S}_i^1, \cdots, \boldsymbol{S}_i^P\}$ to the network in order. LSTM inference layer has 3 gates as well as 2 states [7]: input gate, forget gate, output gate, cell state and hidden state. At t-th step, the hidden features of $\boldsymbol{X}_i$ in LSTM structure can be represented as $\boldsymbol{h}_i^t \in \mathbb{R}^{d_h}$. To better exploit relationship among different modalities, we stack all the previous hidden outputs as $\boldsymbol{H}_i^t = [\boldsymbol{h}_i^0, \boldsymbol{h}_i^1, \cdots, \boldsymbol{h}_i^t] \in \mathbb{R}^{(t+1)d_h}$, where d_h is the dimension of the hidden layer. All the parameters in LSTM structure are denoted as $\boldsymbol{\Psi}$.

3.2.4 Label prediction layer

It is well-known that exploiting label correlations is crucially important in multi-label learning and each modality contains its own specific contribution to the multi-label prediction. In this paper, CoDiSP models label correlations with extracted modal information stored in the memory of LSTM layer.

At the t-th step, we add a fully connected structure between hidden layer and label prediction layer, which makes label prediction with stacked hidden outputs $\boldsymbol{H}_i^t$. The final label prediction is composed of the prediction of the current modality and the prediction of modality

Algorithm 1 Training algorithm for CoDiSP approach

Input:
$\mathcal{D} = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1}^{N}$: Training dataset;
N_b: batch size
Output:
$\boldsymbol{F}^P$: classifier trained with extracted modal sequence

```
Concatenate all modalities of i-th instance X_i as X_i^0 = [X_i^1, X_i^2, ..., X_i^P], i = 1, ..., N
repeat
  Randomly select N_b instances from D without replacement
  for i = 1 : N_b do
    Map raw common modal X_i^0 ∈ R^{d_1+...+d_P} to C_i ∈ R^d
    Map each specific raw modality X_i^m ∈ R^{d_m} to S_i^m ∈ R^d with the same dimension, m = 1, ..., P
    for t = 0 : P do
      if t = 0 then
        Input common modal features C_i to LSTM cell
      else
        Input discriminative modal features S_i^t to LSTM cell
      end if
      Stack hidden output H_i^t = [h_i^0, h_i^1, ..., h_i^t]
      Compute label prediction F^t(H_i^t) with Eq. 6
      Compute label loss function L_{pred}^{i,t} with Eq. 7
    end for
  end for
  Compute common loss function L_comm with Eq. 4
  Compute orthogonal loss function L_orth with Eq. 5
  Compute overall loss function L with Eq. 8
  Compute the derivative ∂L/∂Θ
  Update parameters in Θ
until converge
return F^P
```

used in the last step. And then we predict multiple labels at t-th step according to a nonlinear softmax function according to Eq. 6.

$$\boldsymbol{F}^t(\boldsymbol{H}_i^t) = \begin{cases} \sigma(\boldsymbol{H}_i^t \boldsymbol{W}_L^t + \boldsymbol{b}_L^t) & t = 0 \\ \sigma(\boldsymbol{H}_i^t \boldsymbol{W}_L^t + \boldsymbol{F}^{t-1}(\boldsymbol{H}_i^{t-1})\boldsymbol{R} + \boldsymbol{b}_L^t) & t > 0 \end{cases} \tag{6}$$

where $\boldsymbol{W}_L^t \in \mathbb{R}^{((t+1)d_h)\times L}$ denotes the fully connected weights between $\boldsymbol{H}_i^t$ and label prediction layer, $\boldsymbol{b}_L^t \in \mathbb{R}^{1\times L}$ is the bias vector.

- $\boldsymbol{H}_i^t \boldsymbol{W}_L^t$ is similar to BR, which predicts each label independently.
- $\boldsymbol{F}^{t-1}(\boldsymbol{H}_i^{t-1})\boldsymbol{R}^T$ is the prediction of other labels, in which $\boldsymbol{F}^{t-1}(\boldsymbol{H}_i^{t-1}) \in \mathbb{R}^{1\times L}$ denotes label prediction at the $(t-1)$-th step. Meanwhile, we learn label correlations matrix $\boldsymbol{R} \in \mathbb{R}^{L\times L}$, where L is the number of labels. The k-th row and j-th column of $\boldsymbol{R}$ represents the contribution of the k-th label prediction in $(t-1)$-th step to j-th label, which is denoted as R_{kj}.

Furthermore, we design binary cross-entropy loss function for final label prediction at t-th step by Eq. 7.

$$\mathcal{L}_{pred}^{i,t} = -\sum_{k=1}^{L} \left(Y_i^k log \hat{Y}_i^{k,t} + (1 - Y_i^k) log(1 - \hat{Y}_i^{k,t}) \right) \tag{7}$$

where $\hat{Y}_i^{k,t}$ the prediction of $\boldsymbol{X}_i$ on the k-th label at t-th step, predicted by $\boldsymbol{F}^t$ in Eq. 6.

Above all, we combine common loss, orthogonal loss and label loss function together to compute the overall loss function $\mathcal{L}$:

$$\mathcal{L} = (\sum_{i=1}^{N_b} \sum_{t=0}^{P} \mathcal{L}_{pred}^{i,t}) + \alpha \mathcal{L}_{orth} + \beta \mathcal{L}_{comm} \tag{8}$$

where α and β control the trade-off among different loss terms.

$\boldsymbol{\Theta} = [\boldsymbol{U}_0, \boldsymbol{b}_0, \boldsymbol{U}_m, \boldsymbol{b}_m, \boldsymbol{W}_C, \boldsymbol{b}_C, \boldsymbol{\Psi}, \boldsymbol{R}, \boldsymbol{W}_L^t, \boldsymbol{b}_L^t]$ denotes all the parameters need to be updated in CoDiSP, where $m = 1, \cdots, P$, $t = 0, \cdots, P$. Then we adopt popular optimization algorithm Adam [13] to update parameters in $\boldsymbol{\Theta}$ simultaneously. The pseudo code of CoDiSP in the training phase is presented in Algorithm 1.

4 EXPERIMENTS

4.1 Dataset description

We employ 8 benchmark multi-modal multi-label datasets for performance evaluation, the details are summarized in Table 1.

Table 1. Characteristic of the benchmark multi-modal multi-label datasets. $\#N$, $\#P$ and $\#L$ denote the number of instances, modalities and labels in each dataset, respectively. $\#D$ shows the dimensionality of each modality.

Name	$\#N$	$\#P$	$\#L$	$\#D$
Yeast	2417	2	14	[24, 79]
Emotions	593	3	6	[32, 32, 8]
MSRC	591	3	24	[500,1040,576]
ML2000	2000	3	5	[500,1040,576]
Taobao	2079	4	30	[500,48,81,24]
FCVID	4388	5	28	[400,400,400,400,400]
Scene	2407	6	6	[49, 49, 49, 49, 49, 49]
MSRA	15000	7	50	[256,225,64,144,75,128,7]

- *Yeast* [32] [22] has two modalities including the genetic expression (70 attributes) and the phylogenetic profile of a gene attributes (24 attributes).
- *Emotions* [19] [34] is a publicly available multi-label dataset with rhythmic attributes and timbre attributes.
- *MSRC* [17] is used for object class recognition. As for each image, there are 3 types of modalities including: BoW, FV and HOG.
- *ML2000* [32] is an image dataset and the modal features are extracted similarly to *MSRC*.
- *Taobao* [29] is used for shopping items classification. Description images of items are crawled from a shopping website, and four types of features, i.e., BoW, Gabor, HOG, HSVHist, are extracted to construct 4 modalities of data. Corresponding categories path of an item provides the label sets.
- *FCVID* [12] is the Fudan-Columbia Video Dataset [11], a subset of 4388 videos with most frequent category names are tested. Each video may come from more than one category and features can be extracted in diverse ways. 5 types of features, namely HOF, HOG, CNN, Trajectory and SIFT are extracted for each video, then PCA is conducted to reduce the dimension of each modal to 400.
- *Scene* [1] [34] is public multi-label dataset with 6 modalities.
- *MSRA* is a subset of a salient object recognition dataset [15], which contains 15000 instances from 50 categories, including 256 RGB color histogram features, 225 dimension block-wise color moments, 64 HSV color histogram, 144 color correlogram, 75 distribution histogram, 128 wavelet features and 7 face features.

Table 2. Comparison results (mean ± standard deviation) of CoDiSP with compared approaches on benchmark datasets. The best performance for each criterion is bolded. ↑ / ↓ indicates the larger / smaller the better of the criterion.

Approaches	Hamming Loss ↓							
	Yeast	*Emotions*	*MSRC*	*ML2000*	*Taobao*	*FCVID*	*Scene*	*MSRA*
CAMEL(B)	0.193±0.007	0.218±0.014	0.059±0.009	0.089±0.011	0.032±0.001	0.018± 0.001	0.144±0.007	0.046±0.001
CAMEL(C)	0.189±0.007	0.207±0.025	0.106±0.020	0.098±0.011	0.033±0.002	0.020±0.001	0.076±0.006	**0.045±0.001**
DMP	0.199±0.007	0.196±0.013	0.067±0.008	0.103±0.010	0.053±0.002	0.027±0.002	0.106±0.007	0.046±0.001
CS3G	0.255±0.008	0.290±0.013	0.077±0.008	0.119±0.010	0.064±0.003	0.020±0.001	0.217±0.013	0.050±0.001
MCC	0.213±0.009	0.214±0.023	0.068±0.008	0.105±0.012	0.054±0.002	0.027±0.001	0.102±0.010	0.048±0.001
CoDiSP	**0.189±0.009**	**0.173±0.025**	**0.050±0.008**	**0.086±0.011**	**0.027±0.002**	**0.012±0.001**	**0.064±0.009**	0.046±0.001
Approaches	Ranking Loss ↓							
	Yeast	*Emotions*	*MSRC*	*ML2000*	*Taobao*	*FCVID*	*Scene*	*MSRA*
CAMEL(B)	0.162±0.011	0.176±0.023	**0.033±0.009**	**0.063±0.011**	0.151±0.014	0.027±0.006	0.158±0.018	0.159±0.010
CAMEL(C)	0.163±0.012	0.179±0.038	0.153±0.074	0.067±0.010	0.159±0.012	0.031±0.005	0.057±0.011	0.154±0.009
DMP	0.200±0.010	0.150±0.021	0.047±0.010	0.074±0.010	0.238±0.020	0.052±0.007	0.088±0.007	0.204±0.005
CS3G	0.211±0.012	0.225±0.018	0.043±0.010	0.092±0.013	0.171±0.009	**0.027±0.003**	0.289±0.030	0.140±0.007
MCC	0.224±0.016	0.178±0.029	0.054±0.011	0.082±0.011	0.235±0.016	0.052±0.005	0.101±0.011	0.195±0.005
CoDiSP	**0.159±0.010**	**0.136±0.022**	0.034±0.012	0.064±0.013	**0.098±0.013**	0.029±0.004	**0.052±0.007**	**0.120±0.005**
Approaches	Subset Accuracy ↑							
	Yeast	*Emotions*	*MSRC*	*ML2000*	*Taobao*	*FCVID*	*Scene*	*MSRA*
CAMEL(B)	0.173±0.022	0.248±0.034	0.322±0.058	0.655±0.040	0.150±0.020	0.534±0.030	0.300±0.033	0.057±0.010
CAMEL(C)	0.201±0.018	0.272±0.048	0.070±0.073	0.633±0.036	0.238±0.046	0.485±0.019	0.646±0.024	0.066±0.010
DMP	0.151±0.022	0.284±0.058	0.291±0.033	0.617±0.031	0.218±0.017	0.520±0.033	0.531±0.032	0.053±0.005
CS3G	0.048±0.009	0.165±0.041	0.205±0.043	0.564±0.033	0.104±0.017	0.571±0.018	0.327±0.037	0.061±0.009
MCC	0.190±0.034	0.299±0.037	0.345±0.088	0.662±0.032	0.213±0.023	0.522±0.024	0.662±0.038	0.076±0.003
CoDiSP	**0.222±0.030**	**0.403±0.069**	**0.470±0.058**	**0.745±0.034**	**0.527±0.040**	**0.783±0.026**	**0.777±0.030**	**0.148±0.010**
Approaches	Macro F1 ↑							
	Yeast	*Emotions*	*MSRC*	*ML2000*	*Taobao*	*FCVID*	*Scene*	*MSRA*
CAMEL(B)	0.392±0.025	0.590±0.039	0.680±0.043	0.804±0.025	0.034±0.008	0.697±0.022	0.458±0.033	0.069±0.004
CAMEL(C)	0.458±0.020	0.615±0.058	0.208±0.041	0.781±0.023	0.182±0.028	0.659±0.017	0.772±0.021	0.079±0.002
DMP	0.322±0.012	0.628±0.032	0.643±0.045	0.781±0.020	0.231±0.015	0.660±0.023	0.689±0.020	0.054±0.002
CS3G	0.217±0.017	0.461±0.023	0.506±0.040	0.738±0.021	0.125±0.015	0.693±0.014	0.282±0.034	0.041±0.007
MCC	0.340±0.018	0.621±0.039	0.652±0.024	0.780±0.027	0.222±0.026	0.667±0.009	0.718±0.026	0.073±0.004
CoDiSP	**0.472±0.022**	**0.714±0.043**	**0.755±0.047**	**0.828±0.021**	**0.398±0.028**	**0.834±0.020**	**0.830±0.022**	**0.194±0.013**
Approaches	Example F1 ↑							
	Yeast	*Emotions*	*MSRC*	*ML2000*	*Taobao*	*FCVID*	*Scene*	*MSRA*
CAMEL(B)	0.618±0.014	0.534±0.039	0.805±0.032	0.769±0.031	0.054±0.012	0.551±0.031	0.341±0.035	0.232±0.010
CAMEL(C)	0.628±0.013	0.581±0.049	0.618±0.074	0.739±0.029	0.266±0.044	0.503±0.017	0.695±0.027	0.248±0.006
DMP	0.609±0.017	0.605±0.037	0.789±0.028	0.753±0.026	0.336±0.014	0.642±0.026	0.620±0.026	0.216±0.007
CS3G	0.549±0.011	0.538±0.028	0.729±0.024	0.697±0.029	0.323±0.025	0.673±0.022	0.366±0.036	0.273±0.011
MCC	0.585±0.025	0.624±0.043	0.798±0.024	0.787±0.022	0.333±0.026	0.647±0.019	0.713±0.029	0.266±0.005
CoDiSP	**0.629±0.018**	**0.690±0.050**	**0.850±0.026**	**0.827±0.022**	**0.541±0.039**	**0.796±0.026**	**0.818±0.023**	**0.341±0.013**
Approaches	Micro F1 ↑							
	Yeast	*Emotions*	*MSRC*	*ML2000*	*Taobao*	*FCVID*	*Scene*	*MSRA*
CAMEL(B)	0.642±0.012	0.607±0.038	0.814±0.029	0.806±0.024	0.101±0.023	0.695±0.024	0.457±0.037	0.329±0.014
CAMEL(C)	0.658±0.012	0.637±0.048	0.629±0.071	0.783±0.024	0.373±0.053	0.658±0.016	0.763±0.019	0.349±0.010
DMP	0.632±0.014	0.658±0.027	0.796±0.025	0.783±0.021	0.359±0.015	0.658±0.023	0.683±0.019	0.311±0.009
CS3G	0.566±0.011	0.574±0.031	0.744±0.026	0.744±0.021	0.332±0.026	0.720±0.015	0.372±0.036	0.324±0.018
MCC	0.615±0.021	0.653±0.041	0.799±0.023	0.784±0.024	0.354±0.021	0.663±0.014	0.709±0.027	0.359±0.006
CoDiSP	**0.662±0.015**	**0.719±0.044**	**0.851±0.025**	**0.826±0.021**	**0.591±0.041**	**0.828±0.021**	**0.820±0.023**	**0.430±0.012**

Table 3. Comparison results (mean ± standard deviation) of CoDiSP-NC and CoDiSP, where CoDiSP-NC denotes the model using only discriminative information $\{\boldsymbol{S}^m\}_{m=1}^P$ of each modality in the CoDiSP network. The best performance for each criterion is bolded. ↑ / ↓ indicates the larger / smaller the better of the criterion.

Datasets	Approaches	Evaluation Metrics					
		Hamming Loss ↓	Ranking Loss ↓	Subset Accuracy ↑	Macro F1 ↑	Example F1 ↑	Micro F1 ↑
Yeast	CoDiSP-NC	0.195±0.008	0.164±0.009	0.214±0.025	0.461±0.016	0.617±0.015	0.649±0.013
	CoDiSP	**0.189±0.009**	**0.159±0.010**	**0.222±0.030**	**0.472±0.022**	**0.629±0.018**	**0.662±0.015**
Emotions	CoDiSP-NC	0.174±0.021	0.136±0.019	0.393±0.065	0.707±0.038	0.664±0.039	0.710±0.040
	CoDiSP	**0.173±0.025**	**0.136±0.022**	**0.403±0.069**	**0.714±0.043**	**0.690±0.050**	**0.719±0.044**
MSRC	CoDiSP-NC	0.055±0.012	**0.033±0.012**	0.455±0.063	0.735±0.066	0.838±0.039	0.840±0.037
	CoDiSP	**0.050±0.008**	0.034±0.012	**0.470±0.058**	**0.755±0.047**	**0.850±0.026**	**0.851±0.025**
ML2000	CoDiSP-NC	0.091±0.012	0.072±0.013	0.731±0.032	0.816±0.022	0.815±0.024	0.813±0.022
	CoDiSP	**0.086±0.011**	**0.064±0.013**	**0.745±0.034**	**0.828±0.021**	**0.827±0.022**	**0.826±0.021**
Taobao	CoDiSP-NC	0.029±0.002	0.113±0.015	0.471±0.034	0.346±0.028	0.473±0.035	0.540±0.041
	CoDiSP	**0.027±0.002**	**0.098±0.013**	**0.527±0.040**	**0.398±0.028**	**0.541±0.039**	**0.591±0.041**
FCVID	CoDiSP-NC	0.015±0.001	0.052±0.007	0.708±0.024	0.780±0.014	0.718±0.021	0.777±0.015
	CoDiSP	**0.012±0.001**	**0.029±0.004**	**0.783±0.026**	**0.834±0.020**	**0.796±0.026**	**0.828±0.021**
Scene	CoDiSP-NC	0.069±0.008	0.055±0.010	0.757±0.025	0.814±0.019	0.792±0.022	0.803±0.021
	CoDiSP	**0.064±0.009**	**0.052±0.007**	**0.777±0.030**	**0.830±0.022**	**0.818±0.023**	**0.820±0.023**
MSRA	CoDiSP-NC	0.046±0.001	0.132±0.004	0.115±0.007	0.101±0.007	0.259±0.016	0.353±0.015
	CoDiSP	**0.046±0.001**	**0.120±0.005**	**0.148±0.010**	**0.194±0.013**	**0.341±0.013**	**0.430±0.012**

4.2 Compared approaches

Considering CoDiSP is related to multi-modal multi-label learning, the performance of CoDiSP is compared against 5 approaches, including a state-of-the-art multi-label learning approach with two types of feature inputs and 3 multi-modal multi-label approaches.

- CAMEL(B) & CAMEL(C): CAMEL [3] is a novel multi-label learning approach that aims to explicitly account for the correlated predictions of labels while training the desired model simultaneously. CAMEL(B) stands for the best performance obtained from the best single modality. CAMEL(C) stands for concatenating all the modalities as a single modal input.
- DMP [28]: A multi-modal learning approach, which predicts the label information and decide the modalities to be extracted simultaneously. Here we treat each label independently with DMP.
- CS3G [29]: A multi-modal multi-label approach utilizes multi-modal information in a privacy-preserving style to deal with multi-label tasks, which treats each modality unequally and has the ability to extract the most useful modal features for final prediction.
- MCC [34]: A multi-modal multi-label approach that makes great use of modalities, and can make a convince prediction with many instead of all modalities.

4.3 Evaluation metrics

To evaluate the performance of CoDiSP compared with other multi-modal multi-label learning approaches, we adopt 6 widely-used multi-label evaluation metrics, including Hamming Loss, Ranking Loss, Subset Accuracy, Macro F1, Example F1 and Micro F1 [33]. For Hamming Loss and Ranking Loss, smaller value indicates better performance, while larger value of the other 4 evaluation metrics means better performance. And all the employed evaluation metrics vary within the interval $[0, 1]$.

4.4 Experimental results

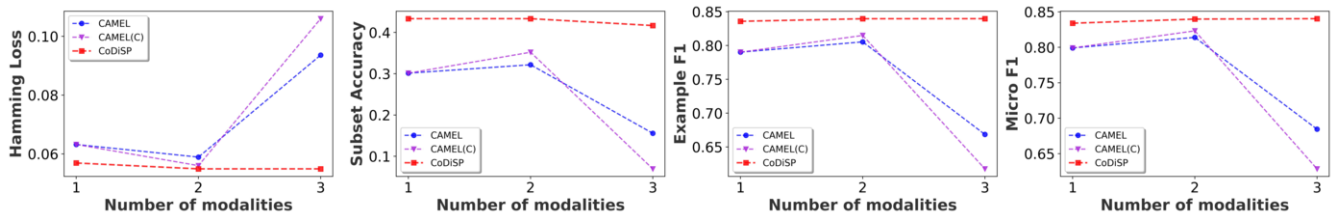
4.4.1 Comparison with state-of-the-arts

For all these approaches, we report the best results of the optimal parameters in terms of classification performance. Meanwhile, we perform 10-fold cross validation (CV) and take the average value of the results in the end. We set trade-off parameter $\alpha = 0.1$, $\beta = 100$ and batch size $N_b = 64$. Furthermore, we drop out 40% modal features at each step to avoid over-fitting [18].

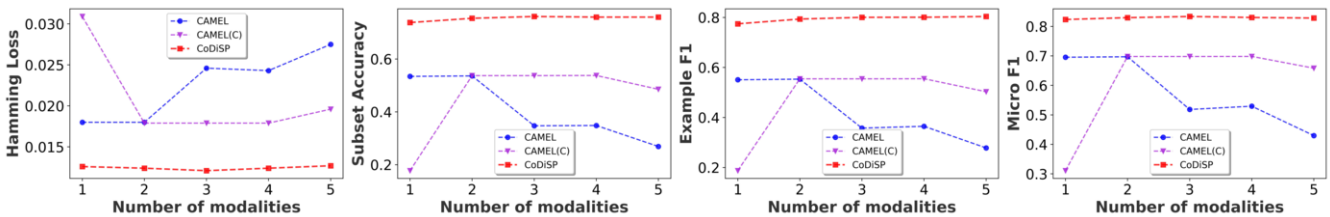
Table 2 demonstrates the comparison results of different approaches on benchmark datasets. Based on the experimental results, several observations are obtained as follows: 1) CoDiSP approach achieves the best performance on all benchmark datasets compared with other state-of-the-art approaches, which demonstrates the highly competitive performance of our proposed approach in multi-modal multi-label learning problem. 2) Concatenating all modalities as a single modality may not always achieve better performance than those of best single modal, which indicates the necessity of taking both concatenated modalities and single modality into consideration.

4.4.2 Common Semantic Analysis

In order to validate the effectiveness of common information extraction, we keep the basic structure of our proposed CoDiSP model and remove the common features $\boldsymbol{C}$ from the modal sequence, which is denoted as CoDiSP-NC. As shown in Table 3, CoDiSP performs better than CoDiSP-NC, which shows the significance of common information exploitation of all the modalities, i.e., it is not good enough to merely fuse semantic of specific modalities. Even though, CoDiSP-NC performs better than other state-of-the-art approaches shown in Table 2, which shows the effectiveness of exploiting discriminative information from each specific modality.

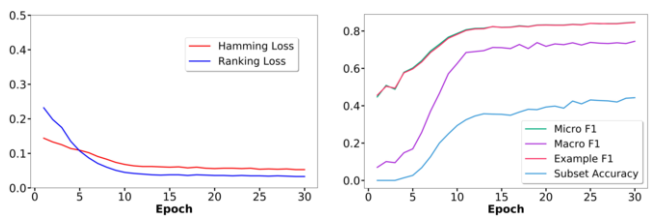


(a) *MSRC* with 3 modalities



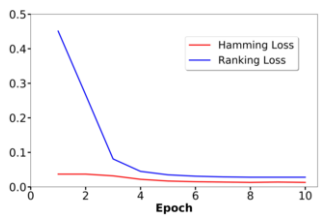
(b) *FCVID* with 5 modalities

Figure 2. Performance of CAMEL, CAMEL(C) and CoDiSP with increase of the modality on the *MSRC* and *FCVID* dataset. With the emergence of new modality, CAMEL adopts the new modality as input, while CAMEL(C) concatenates new modality with previous modalities as input.

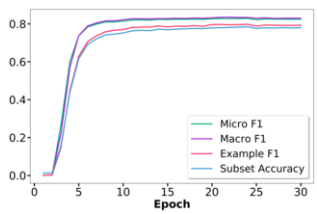


(a) *MSRC*, Hamming&Ranking Loss

(b) *MSRC*, F1&Subset Accuracy



(c) *FCVID*, Hamming&Ranking Loss



(d) *FCVID*,F1&Subset Accuracy

Figure 3. Convergence analysis of CoDiSP approach on the *MSRC* and *FCVID* dataset.

4.4.3 *Modal dependency analysis*

To further evaluate the effectiveness of CoDiSP in sequentially learning modal dependency without pre-determining the modal order, we provide comparison for the state-of-the-art multi-label approach CAMEL in terms of Hamming Loss, Subset Accuracy, Example F1 and Micro F1. Based on the comparison results in Fig. 2, we observe that: 1) CAMEL performs differently with different modalities, so it is significant to exploit discriminative information of each specific modality. 2) With the emergence of new modality, the performance of CAMEL(C) is getting worse on *MSRC* dataset, while the performance of CAMEL(C) is getting better on *FCVID* dataset. In contrast, the curve of CoDiSP tends to be more stable. This phenomenon shows that concatenating modalities is not a wise decision, and extraction of the common features of all modalities with orthogonal constraint is effective.

4.4.4 *Convergence analysis*

We conduct convergence experiments to validate the convergence of CoDiSP, due to the page limit, we only give the convergence results on two datasets: *MSRC* and *FCVID*. As shown in Fig. 3, CoDiSP approach can converge fast within a small number of epochs.

5 CONCLUSION

Rapid development of data collection techniques has spawned the multi-modal multi-label learning, while the modalities extracted from different channels are inconsistent. In this paper, a novel multi-modal multi-label learning approach named Common and Discriminative Semantic Pursuit (CoDiSP) is proposed to exploit consistent and complementary information of multiple modalities, which encodes the commonality and discrimination of different modalities in latent semantic space. And then, common information is added to the specific modal sequence as a new modal. What's more, CoDiSP explores modal relationship and hidden label correlations when inputing into the LSTM network with features of each modality in the new modal sequence. Experiments on 8 benchmark datasets validate the effectiveness of our proposed CoDiSP approach compared with other state-of-the-art approaches. Furthermore, we conduct extensive experiments to analyze common semantic pursuit, modal dependency and convergence. In the future, how to transfer the learned model between different domains is an interesting work.

ACKNOWLEDGEMENTS

This paper is supported by the National Key Research and Development Program of China (Grant No. 2018YFB1403400), the National Natural Science Foundation of China (Grant No. 61876080), the Key Research and Development Program of Jiangsu(Grant No. BE2019105), the Collaborative Innovation Center of Novel Software Technology and Industrialization at Nanjing University.

REFERENCES

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

[2] Zheng Fang and Zhongfei Zhang, 'Simultaneously combining multi-view multi-label learning with maximum margin classification', in *2012 IEEE 12th International Conference on Data Mining*, pp. 864–869. IEEE, (2012).

[3] Lei Feng, Bo An, and Shuo He, 'Collaboration based multi-label learning', in *Thirty-Third AAAI Conference on Artificial Intelligence*, pp. 3550–3557, (2019).

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

[5] Eva Gibaja and Sebastián Ventura, 'A tutorial on multilabel learning', *ACM Computing Surveys (CSUR)*, **47**(3), 52, (2015).

[6] Siddharth Gopal and Yiming Yang, 'Multilabel classification with meta-level features', in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 315–322. ACM, (2010).

[7] Sepp Hochreiter and Jürgen Schmidhuber, 'Long short-term memory', *Neural computation*, **9**(8), 1735–1780, (1997).

[8] Jun Huang, Guorong Li, Qingming Huang, and Xindong Wu, 'Learning label specific features for multi-label classification', in *2015 IEEE International Conference on Data Mining*, pp. 181–190. IEEE, (2015).

[9] Jun Huang, Guorong Li, Qingming Huang, and Xindong Wu, 'Joint feature selection and classification for multilabel learning', *IEEE transactions on cybernetics*, **48**(3), 876–889, (2017).

[10] Sheng-Jun Huang, Yang Yu, and Zhi-Hua Zhou, 'Multi-label hypothesis reuse', in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 525–533. ACM, (2012).

[11] Sheng-Jun Huang and Zhi-Hua Zhou, 'Multi-label learning by exploiting label correlations locally', in *Twenty-sixth AAAI conference on artificial intelligence*, (2012).

[12] Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang, 'Exploiting feature and class relationships in video categorization with regularized deep neural networks', *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **40**(2), 352–364, (2018).

[13] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', in *Proceedings of the 3rd International Conference on Learning Representations*, (2015).

[14] Ximing Li, Jihong Ouyang, and Xiaotang Zhou, 'Supervised topic models for multi-label classification', *Neurocomputing*, **149**, 811–819, (2015).

[15] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum, 'Learning to detect a salient object', *IEEE Transactions on Pattern analysis and machine intelligence*, **33**(2), 353–367, (2010).

[16] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank, 'Classifier chains for multi-label classification', *Machine learning*, **85**(3), 333, (2011).

[17] Florian Schroff, Antonio Criminisi, and Andrew Zisserman, 'Harvesting image databases from the web', *IEEE transactions on pattern analysis and machine intelligence*, **33**(4), 754–766, (2010).

[18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, 'Dropout: a simple way to prevent neural networks from overfitting', *The journal of machine learning research*, **15**(1), 1929–1958, (2014).

[19] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P Vlahavas, 'Multi-label classification of music into emotions.', in *ISMIR*, volume 8, pp. 325–330, (2008).

[20] Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama, 'Efficient learning by directed acyclic graph for resource constrained prediction', in *Advances in Neural Information Processing Systems*, pp. 2152–2160, (2015).

[21] Fei Wu, Xiao-Yuan Jing, Jun Zhou, Yimu Ji, Chao Lan, Qinghua Huang, and Ruchuan Wang, 'Semi-supervised multi-view individual and sharable feature learning for webpage classification', in *The World Wide Web Conference*, pp. 3349–3355. ACM, (2019).

[22] Xuan Wu, Qing-Guo Chen, Yao Hu, Dengbao Wang, Xiaodong Chang, Xiaobo Wang, and Min-Ling Zhang, 'Multi-view multi-label learning with view-specific information extraction', in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3884–3890. AAAI Press, (2019).

[23] Yuying Xing, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang, 'Multi-label co-training', in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 2882–2888. AAAI Press, (2018).

[24] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai, 'Exploit bounding box annotations for multi-label object recognition', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 280–288, (2016).

[25] Pei Yang, Hongxia Yang, Haoda Fu, Dawei Zhou, Jieping Ye, Theodoros Lappas, and Jingrui He, 'Jointly modeling label and feature heterogeneity in medical informatics', *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **10**(4), 39, (2016).

[26] Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang, 'Comprehensive semi-supervised multi-modal learning', in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 4092–4098. AAAI Press, (2019).

[27] Yang Yang, Yi-Feng Wu, De-Chuan Zhan, Zhi-Bin Liu, and Yuan Jiang, 'Complex object classification: A multi-modal multi-instance multi-label deep network with optimal transport', in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2594–2603. ACM, (2018).

[28] Yang Yang, De-Chuan Zhan, Ying Fan, and Yuan Jiang, 'Instance specific discriminative modal pursuit: A serialized approach', in *Asian Conference on Machine Learning*, pp. 65–80, (2017).

[29] Han-Jia Ye, De-Chuan Zhan, Xiaolin Li, Zhen-Chuan Huang, and Yuan Jiang, 'College student scholarships and subsidies granting: A multi-modal multi-label approach', in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 559–568. IEEE, (2016).

[30] Changqing Zhang, Ziwei Yu, Qinghua Hu, Pengfei Zhu, Xinwang Liu, and Xiaobo Wang, 'Latent semantic aware multi-view multi-label classification', in *Thirty-Second AAAI Conference on Artificial Intelligence*, (2018).

[31] 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).

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

[33] Min-Ling Zhang and Zhi-Hua Zhou, 'A review on multi-label learning algorithms', *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, **26**(8), 1819, (2014).

[34] Yi Zhang, Cheng Zeng, Hao Cheng, Chongjun Wang, and Lei Zhang, 'Many could be better than all: A novel instance-oriented algorithm for multi-modal multi-label problem', in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 838–843. IEEE, (2019).

[35] Pengfei Zhu, Qi Hu, Qinghua Hu, Changqing Zhang, and Zhizhao Feng, 'Multi-view label embedding', *Pattern Recognition*, **84**, 126–135, (2018).