Rethinking Modal-oriented Label Correlations for Multi-modal Multi-label Learning

Yi Zhang, Jundong Shen, Zhecheng Zhang, Lei Zhang, Chongjun Wang*
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{njuzhangyi, jdshen, zzc}@smail.nju.edu.cn, {zhangl, chjwang}@nju.edu.cn

***Abstract*—Multi-modal multi-label learning provides a fundamental framework for complex objects, which can be represented with multiple modalities and annotated with multiple labels simultaneously. Different modalities can usually provide complementary information, which may lead to improved performance. What's more, exploiting label correlations is crucially important to multi-label learning. However, most existing multi-label learning approaches do not sufficiently consider the complementary information among different modalities. In this paper, we propose a novel end-to-end deep learning framework named Rethinking Modal-oriented Label Correlations (RMLC), which sequentially polish the label prediction with each individual modality. In order to explicitly account for the correlated prediction of multiple labels, RMLC leverages an efficient sequential modal-based exploration to rethink label correlations. The final prediction of each label involves the collaboration between modal-specific prediction and the prediction of other labels based on cross-modal interaction. Comprehensive experiments on benchmark datasets validate the effectiveness and competitiveness of the proposed RMLC approach.**

***Index Terms*—multi-modal, multi-label, label correlations, modal-specific, cross-modal**

I. Introduction

With the fast development of data collection techniques, objects are often characterized by features from different data channels, i.e., multi-modal feature representations. For example, a news webpage can be represented with two heterogeneous modalities: text and image; an image can be described using different features, such as texture descriptors, shape descriptors, color descriptors, surrounding texts, and so on [1].

Although multi-modal (or multi-view) learning approaches have been developed and paid more attention, most previous studies assume that each object is annotated with a single label. Nevertheless, in real-world applications, each object may have multiple semantic meanings. For instance, a webpage may be tagged with multiple labels: economics, sports and culture.

As a result, Multi-Modal Multi-Label (MMML) learning serves an important framework to solve complex objects with multiple modalities and multiple labels. For example, in video annotation, a film can be represented from multiple channels including text, audio, picture and frame, meanwhile it can be annotated with *superhero movie* (type), *Marvel Studios* (producer), *America* (country) and *Anthony Russot and Joseph Russo* (directors). The major challenge of MMML learning lies in how to jointly model the multiple types of heterogeneities in a mutually beneficial way. For one thing, the representations of various modalities are quite different from each other and it is a challenge to fuse the multiple modalities directly with large discrepancy. Previous approaches do not explicitly account for the distinctive information hidden in each specific modality. For another, how to exploit label correlations in an effectiveness way is also a challenging issue. Since each specific modality captures a specific property of data, it is impossible for one modality to comprehensively characterize all the relevant labels.

Corresponding author

In this paper, aiming at simultaneously exploit the modal correlations and label correlations in a mutually benefit way, we proposed a novel MMML learning approach named Rethinking Modal-oriented Label Correlations (RMLC). On the one hand, RMLC models the cross-modal interaction (i.e., rethinking process) in a LSTM network [2], which captures the complementary patterns among multiple modalities. On the other hand, we input each specific modality to the RMLC network step by step, in which we make label prediction and exploit label correlations simultaneously. For each label, the final prediction consists of two aspects: (1) its own prediction based on each modal-specific information (2) the prediction of other labels based on cross-modal interaction.

The main contributions are summarized as follows.

- We propose a novel end-to-end deep network structure named Rethinking Modal-oriented Label Correlations (RMLC) for multi-modal multi-label learning, inspired from LSTM.
- RMLC network is better at extracting modal-specific information and long range cross-modal information, which stores modal information with memory cells. Meanwhile, RMLC sequentially exploits label correlations with the collaboration of both modal-specific information and cross-modal interaction.
- Extensive experiments on 5 benchmark multi-modal multi-label datasets verify the effectiveness of RMLC compared with several state-of-the-art approaches.

The rest of the paper is organized as follows. Section II briefly reviews some related works. Section III presents technical details of the proposed approach. Section IV reports experimental results, followed by the conclusion in Section V.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

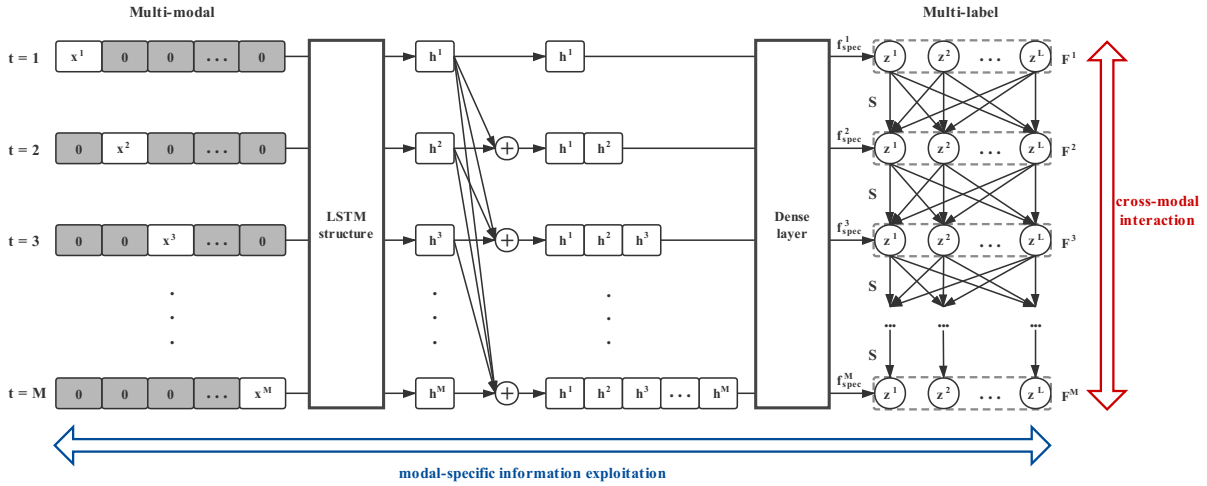


Fig. 1. The overall flowchart of RMLC approach, which is composed of a LSTM structure and a dense layer. $\oplus$ denotes the concatenation the vectors. At t-th step, RMLC inputs modified feature vector to the LSTM structure. After obtaining output vector $\boldsymbol{h}^t$, RMLC stacks all previous output vectors, i.e., $[\boldsymbol{h}^1, \boldsymbol{h}^2, \cdots, \boldsymbol{h}^t]$. Based on the stacked output features of LSTM, RMLC learns label prediction from previous t modalities, which is denoted as $\boldsymbol{f}_{spec}^t$. The LSTM layer is used for rethink process, which goes through M modalities sequentially and each step shares the same label correlation matrix $\boldsymbol{S}$. What's more, the label prediction at each step is propagated to the next step in the label prediction layer. RMLC uses the label prediction form previous step to polishes the modal-specific label prediction, i.e., at t-th step, RMLC makes label prediction with collaboration between modal-specific label prediction $\boldsymbol{f}_{spec}^t$ and previous label predictions $\boldsymbol{F}^t$ of other labels.

II. Related Work

Our work is related to two branches of studies: multi-label learning and multi-modal learning. In this section, we briefly review some state-of-the-art approaches in the two fields.

Multi-label learning [3] [4] deals with objects annotated with multiple interdependent labels. Binary Relevance (BR) [5] is the most straightforward solution that decompose the multi-label problem into a set of independent binary classification tasks, but it neglects correlations among labels. To tackle label dependence, Classifier Chains (CC) [6] was proposed as a high-order approach to consider correlations between labels. However, the performance of CC is seriously affected by the training order of labels. So far, many approaches have been developed to improve the performance of multi-label learning by exploring various types of label correlations. For example, [7] exploits label correlations locally, [8] learns label-specific data representation for each label, [9] exploits global and local label correlations. However, the label correlations are simply obtained by common similarity measures, which may not be able to reflect complex relationships among labels. To address the above limitation, CAMEL [10] is proposed to learn label correlations via sparse reconstruction in the label space, which is capable of reflecting the collaborative relationships among labels regarding the final predictions.

Multi-modal learning aims to jointly utilize different information collected with diverse data collection techniques, e.g., [11] applied deep network to learn features over multiple modal data. Most approaches are mainly derived from the Canonical Correlation Analysis (CCA) methods, such as deep neural networks based CCA [12]. Meanwhile, Multi-modal multi-label learning has been widely studied, e.g., [9] exploits the consensus among different modalities, where multi-view latent spaces are correlated by HilbertSchmidt Independence Criterion (HSIC). It has been recognized that exploring individuality and commonality of heterogeneous features can further boost the performance of multi-modal data mining. SMISFL [13] jointly learns multiple modal-individual transformations and one sharable transformation; SIMM [14] leverages shared subspace exploitation and modal-specific information extraction; ICM2L [13] adopts an ensemble strategy to explicitly explore the individuality and commonality information of multi-label multiple view data in a unified model. Nevertheless, the above approaches rarely consider the label correlations. MMP [15] handles the consistencies among different views by requiring them to generate the same annotation result, and captures the correlations among different labels by imposing the similarity constraints. CS3G approach [16] handles types of interactions between multiple labels, while no interaction between features from different modalities. [17] 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 make each modality interacts and further reduce modal extraction cost, MCC [18] extends Classifier Chains to exploit label correlation with partial modalities.

III. Proposed Methodology

In this section, we first summarize some formal notations used throughout this paper. And then we present detailedly the proposed method named Rethinking Modal-oriented Label Correlations (RMLC), along with its implementation. Fig. 1 shows the overall architecture of RMLC.

A. Preliminaries

In the multi-modal multi-label learning, an instance is characterized with multiple modalities and annotated with multiple labels. Formally, let $\mathcal{X} = \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}^{d_M}$, where $d_m (1 \leq m \leq M)$ is the dimensionality of the m-th modality.

Suppose $\mathcal{D} = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1}^N$ represents a dataset with N samples. For the i-th instance, $\boldsymbol{X}_i = [\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \cdots, \boldsymbol{x}_i^M] \in \mathcal{X}$ is the feature vector with M modalities, where $\boldsymbol{x}_i^m \in \mathbb{R}^{d_m}$. $\boldsymbol{Y}_i = [y_i^1, y_i^2, \cdots, y_i^L] \in \mathcal{Y}$ is the corresponding label vector with L labels, where $y_i^k \in \{-1, 1\}$. $y_i^k = 1 (k = 1, \cdots, L)$ means the k-th label of $\boldsymbol{X}_i$ is relevant; $y_i^k = -1$ otherwise.

The task of multi-modal multi-label learning is to learn a predictive model: $h : \mathcal{X} \to 2^{\mathcal{Y}}$ from $\mathcal{D}$, which can assign a set of proper labels for the unseen instance.

RMLC sequentially polishing the label prediction in a rethink manner and exploits label correlations based on the following two types of information: modal-specific information and cross-modal interaction. What's more, the final prediction of each label is composed of its own prediction and the prediction of other labels.

Algorithm 1 The pseudo code of RMLC approach

Input:
$\mathcal{D} = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1}^N$: Training dataset;
N_b: batch size
λ : trade-off parameter
Output:
$\boldsymbol{F}^M$: multi-modal classifier trained with M-modalties

repeat
  Initialize label correlation matrix $\boldsymbol{S}$
  Randomly select N_b instances from $\mathcal{D}$
  for $t = 1 : M$ **do**
    for $i = 1 : N_b$ **do**
      Modify $\boldsymbol{X}_i$ to $\hat{\boldsymbol{X}}_i^t$ with Eq. 1
      Input $\hat{\boldsymbol{X}}_i^t$ to the LSTM structure
      Stack hidden output $\boldsymbol{H}_i^t = [\boldsymbol{h}_i^0, \boldsymbol{h}_i^1, \cdots, \boldsymbol{h}_i^t]$
      Compute label prediction $\boldsymbol{F}^t(\boldsymbol{H}_i^t)$ with Eq. 4
      Calculate loss function $\mathcal{L}_i^t$ with Eq. 5
    end for
    Calculate overall loss function $\mathcal{L}^t$ with Eq. 6
    Weight propagation: Obtain the derivative $\partial\mathcal{L}^t/\partial\boldsymbol{\Phi}$, $\partial\mathcal{L}^t/\partial\boldsymbol{U}^t$, $\partial\mathcal{L}^t/\partial\boldsymbol{b}^t$ and $\partial\mathcal{L}^t/\partial\boldsymbol{S}$
    Update parameters: $\boldsymbol{\Phi}$, $\boldsymbol{U}^t$, $\boldsymbol{b}^t$ and $\boldsymbol{S}$
  end for
until converge
return $\boldsymbol{F}^M$

B. Modal-specific information exploitation

It is well-known that each modality contains its own specific contribution to the multi-label prediction. Take image annotation as an example, a picture of pink rose may be represented with two modalities: HSV and Gist, while tagged with two labels *pink*, *flower* simultaneously [14]. Intuitively, we can infer *pink* from HSV (color) modality and *flower* from Gist (texture) modality. Existing approaches mainly try to find the shared information between different modalities, while it is more reasonable to consider extracting their own specific information. At the t-th step, RMLC exploits specific information of the t-th individual modality $\boldsymbol{x}_i^t$. It is notable that the dimensionality of various modalities is heterogeneous, which is difficult to input to the LSTM structure.

First of all, we adapt $\boldsymbol{X}_i$ as $\hat{\boldsymbol{X}}_i^t = [\hat{\boldsymbol{x}}_i^1, \hat{\boldsymbol{x}}_i^2, \cdots, \hat{\boldsymbol{x}}_i^M]$ according to Eq. 1.

$$\hat{\boldsymbol{x}}_i^m = \begin{cases} \boldsymbol{x}_i^m & m = t \\ \\ \boldsymbol{0}^m & m \neq t \end{cases} \tag{1}$$

where $\boldsymbol{0}^m \in \mathbb{R}^{d_m}$. For example, if $t = 4$, $\hat{\boldsymbol{X}}_i^4 = [\boldsymbol{0}^1, \boldsymbol{0}^2, \boldsymbol{0}^3, \boldsymbol{x}_i^4, \cdots, \boldsymbol{0}^M]$.

Secondly, we input the modified $\hat{\boldsymbol{X}}_i^t$ to LSTM structure, which includes input gate, forget gate, cell state, output gate, and output vector.

$$\begin{aligned} \boldsymbol{i}_i^t &= \sigma(\boldsymbol{W}_i[\hat{\boldsymbol{X}}_i^t, \boldsymbol{h}_i^{t-1}, \boldsymbol{c}_i^{t-1}] + \boldsymbol{b}_i) \\ \boldsymbol{f}_i^t &= \sigma(\boldsymbol{W}_f[\hat{\boldsymbol{X}}_i^t, \boldsymbol{h}_i^{t-1}, \boldsymbol{c}_i^{t-1}] + \boldsymbol{b}_f) \\ \boldsymbol{c}_i^t &= \boldsymbol{f}_i^t \boldsymbol{c}_i^{t-1} + \boldsymbol{i}_i^t tanh(\boldsymbol{W}_c[\hat{\boldsymbol{X}}_i^{t-1}, \boldsymbol{h}_i^{t-1}] + \boldsymbol{b}_c) \\ \boldsymbol{o}_i^t &= \sigma(\boldsymbol{W}_o[\hat{\boldsymbol{X}}_i^t, \boldsymbol{h}_i^{t-1}, \boldsymbol{c}_i^t] + \boldsymbol{b}_o) \\ \boldsymbol{h}_i^t &= \boldsymbol{o}_i^t tanh(\boldsymbol{c}_i^t) \end{aligned} \tag{2}$$

where $\boldsymbol{W}_i, \boldsymbol{W}_f, \boldsymbol{W}_c, \boldsymbol{W}_o \in \mathbb{R}^h$ represents the weight matrices form the cell to gate vectors, $\boldsymbol{b}_i, \boldsymbol{b}_f, \boldsymbol{b}_c, \boldsymbol{b}_o \in \mathbb{R}^h$ denotes bias vectors. For simplicity, we denote all the parameters in the LSTM structure as $\boldsymbol{\Phi}$.

In order to combine the specific information extracted from previous t modalities $\{\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \cdots, \boldsymbol{x}_i^t\}$, we stack all the output features, which is denoted as $\boldsymbol{H}_i^t = [\boldsymbol{h}_i^1, \boldsymbol{h}_i^2, \cdots, \boldsymbol{h}_i^t]$ and $\boldsymbol{H}_i^t \in \mathbb{R}^{1 \times (t \cdot h)}$.

Last but not the least, we add a dense layer between stacked output features and label prediction layer. The dense layer learns multi-label embedding to transform the output features $\boldsymbol{H}_i^t$ to label vector, which can be predicted by Eq. 3.

$$\begin{aligned} \boldsymbol{f}_{spec}^t(\boldsymbol{H}_i^t) &= \boldsymbol{U}^t \boldsymbol{H}_i^t + \boldsymbol{b}^t \\ &= \boldsymbol{U}^t [\boldsymbol{h}_i^1, \boldsymbol{h}_i^2, \cdots, \boldsymbol{h}_i^t] + \boldsymbol{b}^t \end{aligned} \tag{3}$$

where $\boldsymbol{U}^t \in \mathbb{R}^{(t \cdot h) \times L}$ is the weight vector, $\boldsymbol{b}^t$ is the bias vector, and $\boldsymbol{f}_{spec}^t(\cdot) \in \mathbb{R}^{1 \times L}$.

C. Cross-modal interaction

The exploitation of cross-modal interaction can be considered as a sequential case. To characterize the collaborative relationship among multiple labels regarding the final prediction, RMLC also learns label correlation matrix $\boldsymbol{S} = [s_{jk}]_{L \times L}$, where s_{jk} reflects the contribution of the j-th label with all previous $t-1$ modalities to the k-th label.

Each step in LSTM structure represents a rethink step, which classifies labels and exploit label correlations simultaneously. We denote the predicted label vector as $\boldsymbol{Z}_i^t = [z_i^{1,t}, z_i^{2,t}, \cdots, z_i^{L,t}]$ and we initialize $\boldsymbol{Z}_i^0 = \boldsymbol{0}$ for all instances, where $\boldsymbol{Z}_i^t = F_i^t(\boldsymbol{H}_i^t)$ according to Eq. 4. With the help of

TABLE I

EXAMPLE OF LABEL PREDICTOR $\boldsymbol{f}_{spec}^t(\cdot)$ AND $\boldsymbol{F}^t(\cdot)$, WHERE $\boldsymbol{f}_{spec}^t(\cdot)$ DENOTES LABEL PREDICTION WITH THE MODAL-SPECIFIC INFORMATION EXPLOITATION, $\boldsymbol{F}^t(\cdot)$ DENOTES FINAL LABEL PREDICTION WHICH INVOLVES COLLABORATION BETWEEN ITS OWN PREDICTION AND THE PREDICTION OF OTHER LABELS FORM THE LAST STEP. $\boldsymbol{S}$ DENOTES LABEL CORRELATION MATRIX.

step	label predictor with modal-specific information exploitation	label predictor with both modal-specific and cross-modal interaction
$t=1$	$\boldsymbol{f}_{spec}^1(\boldsymbol{H}_i^1)=\boldsymbol{U}^1\boldsymbol{h}_i^1+\boldsymbol{b}^1$	$\boldsymbol{F}^1(\boldsymbol{H}_i^1)=\sigma(\boldsymbol{U}^1\boldsymbol{h}_i^1+\boldsymbol{b}^1))$
$t=2$	$\boldsymbol{f}_{spec}^2(\boldsymbol{H}_i^2)=\boldsymbol{U}^2[\boldsymbol{h}_i^1,\boldsymbol{h}_i^2]+\boldsymbol{b}^2$	$\boldsymbol{F}^2(\boldsymbol{H}_i^2)=\sigma\big(\boldsymbol{U}^2[\boldsymbol{h}_i^1,\underline{\boldsymbol{h}_i^2]+\boldsymbol{b}^2}+\boldsymbol{F}^1(h_i^1)\boldsymbol{S}\ \big)$
$t=3$	$\boldsymbol{f}_{spec}^3(\boldsymbol{H}_i^3)=\boldsymbol{U}^3[\boldsymbol{h}_i^1,\boldsymbol{h}_i^2,\boldsymbol{h}_i^3]+\boldsymbol{b}^3$	$\boldsymbol{F}^3(\boldsymbol{H}_i^3)=\sigma\big(\underline{\boldsymbol{U}^3[\boldsymbol{h}_i^1,\boldsymbol{h}_i^2,\boldsymbol{h}_i^3]+\boldsymbol{b}^3}+\boldsymbol{F}^2([h_i^1,h_i^2])\boldsymbol{S}\ \big)$
$\cdots$	$\cdots$	$\cdots$
$t=M$	$\boldsymbol{f}_{spec}^M(\boldsymbol{H}_i^M)=\boldsymbol{U}^M[\boldsymbol{h}_i^1,\boldsymbol{h}_i^2,\cdots,\boldsymbol{h}_i^M]+\boldsymbol{b}^M$	$\boldsymbol{F}^M\big(\boldsymbol{H}_i^M\big)=\sigma(\underline{\boldsymbol{U}^M[\boldsymbol{h}_i^1,\boldsymbol{h}_i^2,\cdots,\boldsymbol{h}_i^M]+\boldsymbol{b}^t}+\boldsymbol{F}^{M-1}([h_i^1,\boldsymbol{h}_i^2,\cdots,h_i^{M-1}])\boldsymbol{S}\ \big)$
Output	$\boldsymbol{f}_{spec}^M(\cdot)$	$\boldsymbol{F}^M(\cdot)$

TABLE II

CHARACTERISTIC OF THE BENCHMARK MULTI-MODAL MULTI-LABEL DATASETS. N, M AND L DENOTE THE NUMBER OF INSTANCES, MODALITIES AND LABELS IN EACH DATASET, RESPECTIVELY. d_m SHOWS THE DIMENSIONALITY OF EACH MODALITY.

dataset	# of size N	# of modalities M	# of labels L	# the dimensionality for each modality d_m
ML2000	2000	3	5	[500,1040,576]
MSRC	591	3	24	[500,1040,576]
Taobao	2079	4	30	[500,48,81,24]
FCVID	4388	5	28	[400,400,400,400,400]
MSRA	15000	7	50	[256,225,64,144,75,128,7]

memory in the LSTM structure, each $\boldsymbol{Z}_i^t$ is passed down to $(t+1)$-th step.

Combine both modal-specific prediction and cross-modal interactions together, we predict labels according to a nonlinear function $\boldsymbol{F}^t(\cdot)$.

$$\boldsymbol{F}^t(\boldsymbol{H}_i^t)=\sigma\big(\boldsymbol{f}_{spec}(\boldsymbol{H}_i^t)+\boldsymbol{Z}_i^{t-1}\boldsymbol{S}\big) \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function, $\boldsymbol{Z}_i^{t-1}=\boldsymbol{F}^{t-1}(\boldsymbol{H}_i^{t-1})$ denotes the predicted label vector at $t-1$-th step. The modal-specific term $\boldsymbol{f}_{spec}^t(\cdot)$ concentrates on extracting individual information of each modality to predict labels independently, which is similar to BR. The memory term $\boldsymbol{Z}_i^{t-1}\boldsymbol{S}$ transforms the previous prediction to the current label vector space, which can also be considered as the exploitation for label correlations with cross-modal interactions. Table I shows example of label predictor $\boldsymbol{f}_{spec}^t(\cdot)$ and $\boldsymbol{F}^t(\cdot)$ at each step.

D. Loss function

As a general training procedure, it focuses on reducing the errors made in the current status of the network. Thus we design weight binary cross-entropy loss function for label prediction at the t-th step according to:

$$\mathcal{L}_i^t=-\sum_{k=1}^{L}\big(y_i^k log z_i^{k,t}+(1-y_i^k)log(1-z_i^{k,t})\big) \tag{5}$$

where $\boldsymbol{Z}_i^t=[z_i^{1,t},z_i^{2,t},\cdots,z_i^{L,t}]$ is predicted by $\boldsymbol{F}^t(\boldsymbol{H}_i^t)$ and $z_i^{k,t}$ is the prediction of the k-th label at the t-th step.

Above all, we combine label loss function $\mathcal{L}_i^t$ and regularization term $\|\boldsymbol{U}^t\|_2^2$ to calculate the overall loss function:

$$\mathcal{L}^t=\sum_{i}^{N_b}\mathcal{L}_i^t+\lambda\|\boldsymbol{U}^t\|_2^2 \tag{6}$$

where $\|\cdot\|_2$ represents L_2 norm and λ is the trade-off between the label loss function and the regularization term.

In the training procedure, the derivatives are taken with the help of back propagation technique, and the pseudo code of RMLC is summarized in Algorithm 1. At t-th step, we adopt the popular optimization algorithm Adam [19] to update all the parameters in $\boldsymbol{\Phi}$, $\boldsymbol{U}^t$, $\boldsymbol{b}^t$ and $\boldsymbol{S}$ simultaneously.

IV. EXPERIMENTS

In this section, we conduct extensive experiments on various datasets to validate the effectiveness of RMLC.

A. Dataset description

For comprehensive performance evaluation, 5 benchmark multi-modal multi-label datasets are collected as follows. Table II summarizes the detailed characteristics of these datasets, which are organized in ascending order of M (the number of modalities).

- *ML2000* [20] is an image dataset with 2000 images from 5 categories (desert, mountains, sea, sunset and trees), and we extract 3 types of modalities for each image including: BoW, FV and HOG.
- *MSRC* [21] is used for object class recognition. Similar to *ML2000*, we extract 3 modalities of features: BoW, FV and HOG for each image. Each image may be affiliated to two or more labels. Given different modalities of description for an image, the task is to predict its all possible labels.
- *Taobao* [16] is used for shopping items classification which has 2079 instances and 30 labels. Description images of items are crawled from a shopping website, and four types of features, e.g., BoW, Gabor, HOG,

TABLE III
COMPARISON RESULTS (MEAN ± STANDARD DEVIATION) OF RMLC WITH COMPARED APPROACHES ON BENCHMARK DATASETS. 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 ↑
ML2000	CAMEL(B)	0.089±0.011	0.063±0.011	0.655±0.040	0.804±0.025	0.769±0.031	0.806±0.024
	CAMEL(C)	0.098±0.011	0.067±0.010	0.633±0.036	0.781±0.023	0.739±0.029	0.783±0.024
	DMP	0.103±0.010	0.074±0.010	0.617±0.031	0.781±0.020	0.753±0.026	0.783±0.021
	CS3G	0.119±0.010	0.092±0.013	0.564±0.033	0.738±0.021	0.697±0.029	0.744±0.021
	MCC	0.105±0.012	0.082±0.011	0.662±0.032	0.780±0.027	0.787±0.022	0.784±0.024
	RMLC	**0.080±0.007**	**0.058±0.009**	**0.753±0.024**	**0.835±0.016**	**0.830±0.021**	**0.834±0.016**
MSRC	CAMEL(B)	0.059±0.009	0.033±0.009	0.322±0.058	0.680±0.043	0.805±0.032	0.814±0.029
	CAMEL(C)	0.106±0.020	0.153±0.074	0.070±0.073	0.208±0.041	0.618±0.074	0.629±0.071
	DMP	0.067±0.008	0.047±0.010	0.291±0.033	0.643±0.045	0.789±0.028	0.796±0.025
	CS3G	0.077±0.008	0.043±0.010	0.205±0.043	0.506±0.040	0.729±0.024	0.744±0.026
	MCC	0.068±0.008	0.054±0.011	0.345±0.088	0.652±0.024	0.798±0.024	0.799±0.023
	RMLC	**0.048±0.010**	**0.031±0.010**	**0.472±0.048**	**0.761±0.059**	**0.854±0.035**	**0.856±0.032**
Taobao	CAMEL(B)	0.032±0.001	0.151±0.014	0.150±0.020	0.034±0.008	0.054±0.012	0.101±0.023
	CAMEL(C)	0.033±0.002	0.159±0.012	0.238±0.046	0.182±0.028	0.266±0.044	0.373±0.053
	DMP	0.053±0.002	0.238±0.020	0.218±0.017	0.231±0.015	0.336±0.014	0.359±0.015
	CS3G	0.064±0.003	0.171±0.009	0.104±0.017	0.125±0.015	0.323±0.025	0.332±0.026
	MCC	0.054±0.002	0.235±0.016	0.213±0.023	0.222±0.026	0.333±0.026	0.354±0.021
	RMLC	**0.028±0.002**	**0.115±0.020**	**0.476±0.035**	**0.352±0.027**	**0.477±0.038**	**0.555±0.039**
FCVID	CAMEL(B)	0.018± 0.001	0.027±0.006	0.534±0.030	0.697±0.022	0.551±0.031	0.695±0.024
	CAMEL(C)	0.020±0.001	0.031±0.005	0.485±0.019	0.659±0.017	0.503±0.017	0.658±0.016
	DMP	0.027±0.002	0.052±0.007	0.520±0.033	0.660±0.023	0.642±0.026	0.658±0.023
	CS3G	0.020±0.001	0.027±0.003	0.571±0.018	0.693±0.014	0.673±0.022	0.720±0.015
	MCC	0.027±0.001	0.052±0.005	0.522±0.024	0.667±0.009	0.647±0.019	0.663±0.014
	RMLC	**0.013±0.001**	**0.023±0.003**	**0.757±0.023**	**0.829±0.015**	**0.797±0.023**	**0.821±0.017**
MSRA	CAMEL(B)	0.046±0.001	0.159±0.010	0.057±0.010	0.069±0.004	0.232±0.010	0.329±0.014
	CAMEL(C)	**0.045±0.001**	0.154±0.009	0.066±0.010	0.079±0.002	0.248±0.006	0.349±0.010
	DMP	0.046±0.001	0.204±0.005	0.053±0.005	0.054±0.002	0.216±0.007	0.311±0.009
	CS3G	0.050±0.001	0.140±0.007	0.061±0.009	0.041±0.007	0.273±0.011	0.324±0.018
	MCC	0.048±0.001	0.195±0.005	0.076±0.003	0.073±0.004	0.266±0.005	0.359±0.006
	RMLC	0.048±0.001	**0.132±0.005**	**0.136±0.004**	**0.186±0.007**	**0.334±0.011**	**0.421±0.009**

HSVHist, are extracted to construct 4 modalities of data. Corresponding categories path of an item provides the label sets.

- *FCVID* [22] 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. Given different modalities of description for a video, the task is to predict its possible categories.
- *MSRA* is a subset of a salient object recognition dataset [23], 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.

B. Evaluation metrics

For performance evaluation, we use 6 widely-adopted evaluation metrics, including *Hamming Loss*, *Ranking Loss*, *Subset Accuracy*, *Macro F1*, *Example F1* and *Micro F1* [3], which consider the performance of multi-label predictor from various aspects. All the employed approaches vary within the interval $[0, 1]$. For the first two evaluation metrics, the smaller values indicate the better performance, while for the last four evaluation metrics, the larger indicate the better performance.

C. Compared approaches

The performance of RMLC is compared against 5 multi-modal multi-label learning approaches, listed as follows:

- CAMEL(B) & CAMEL(C): CAMEL [10] is a state-of-the-art multi-label approach, which learn the label correlations via sparse reconstruction in the label space. CAMEL(B) stands for the best performance obtained from the best single modality. CAMEL(C) stands for concatenating all modalities as a single modal input.
- DMP [24]: A multi-modal approach which automatically extracts instance-specifically discriminative modal sequence for reducing the cost of feature extraction. Here, we treat each label independently, i.e., for each label, DMP trains classifiers using different modalities.
- CS3G [16]: A multi-modal multi-label approach utilizing multi-modal information in a privacy-preserving style to

TABLE IV

COMPARISON RESULTS (MEAN ± STANDARD DEVIATION) OF RMLC_NS AND RMLC, WHERE RMLC_NS DENOTES THE MODEL WITHOUT CONSIDERING LABEL CORRELATION MATRIX $\boldsymbol{S}$. 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 ↑
ML2000	RMLC_NS	0.083±0.008	0.059±0.009	0.742±0.028	0.828±0.017	0.821±0.019	0.826±0.017
	RMLC	**0.080±0.007**	**0.058±0.009**	**0.753±0.024**	**0.835±0.016**	**0.830±0.021**	**0.834±0.016**
MSRC	RMLC_NS	0.059±0.006	0.040±0.012	0.359±0.062	0.693±0.058	0.815±0.022	0.819±0.022
	RMLC	**0.048±0.010**	**0.031±0.010**	**0.472±0.048**	**0.761±0.059**	**0.854±0.035**	**0.856±0.032**
Taobao	RMLC_NS	0.030±0.002	0.137±0.017	0.394±0.039	0.284±0.027	0.383±0.038	0.486±0.041
	RMLC	**0.028±0.002**	**0.115±0.020**	**0.476±0.035**	**0.352±0.027**	**0.477±0.038**	**0.555±0.039**
FCVID	RMLC_NS	0.013±0.001	0.025±0.004	0.702±0.021	0.804±0.016	0.703±0.019	0.801±0.017
	RMLC	**0.013±0.001**	**0.023±0.003**	**0.757±0.023**	**0.829±0.015**	**0.797±0.023**	**0.821±0.017**
MSRA	RMLC_NS	**0.045±0.000**	**0.115±0.005**	0.119±0.007	0.125±0.009	0.280±0.008	0.393±0.009
	RMLC	0.048±0.001	0.132±0.005	**0.136±0.004**	**0.186±0.007**	**0.334±0.011**	**0.421±0.009**

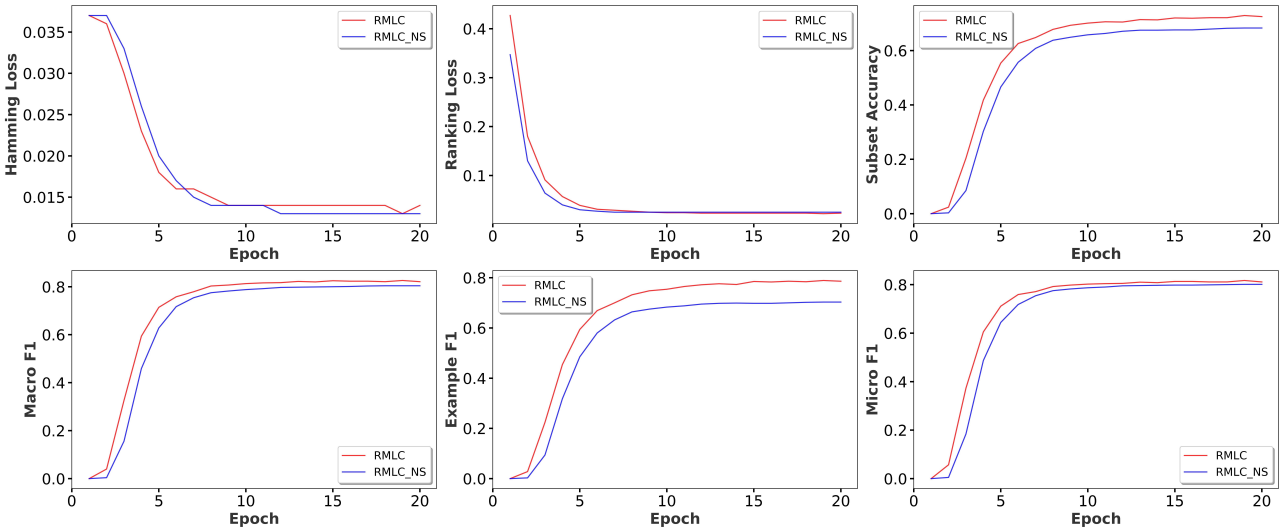


Fig. 2. Convergence analysis of RMLC and RMLC_NS approach on the FCVID dataset.

deal with multi-label tasks. CS3G treats each modality unequally and has the ability of extracting the most useful modal features for final recommendation as well.

- MCC [18]: A novel multi-modal multi-label approach considering not only interrelation among different modalities, but also relationship among different labels. And MCC makes convince prediction with partial modalities.

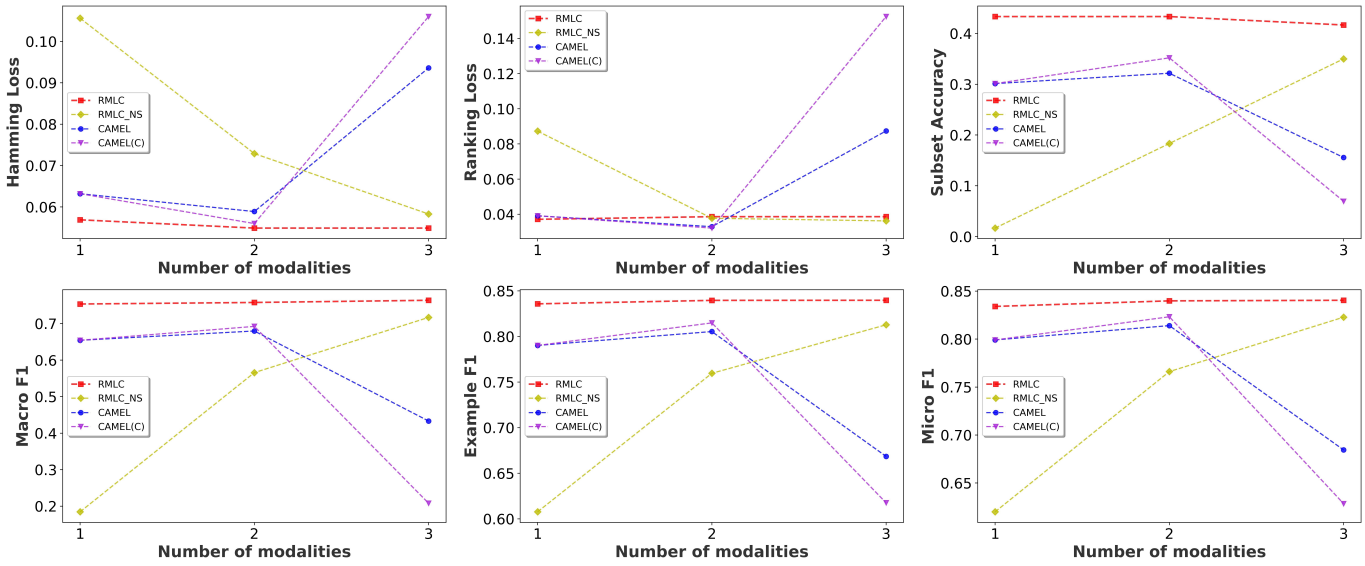
D. Experimental results

For each dataset, we perform 10-fold cross-validation and take the mean metrics results, standard deviations for all compared approaches. Detailed experimental results are reported in Table III, which obviously shows RMLC outperforms the other 5 compared approaches on all evaluation metrics. We set trade-off parameter $\lambda = 0.1$ and batch size $N_b = 64$.

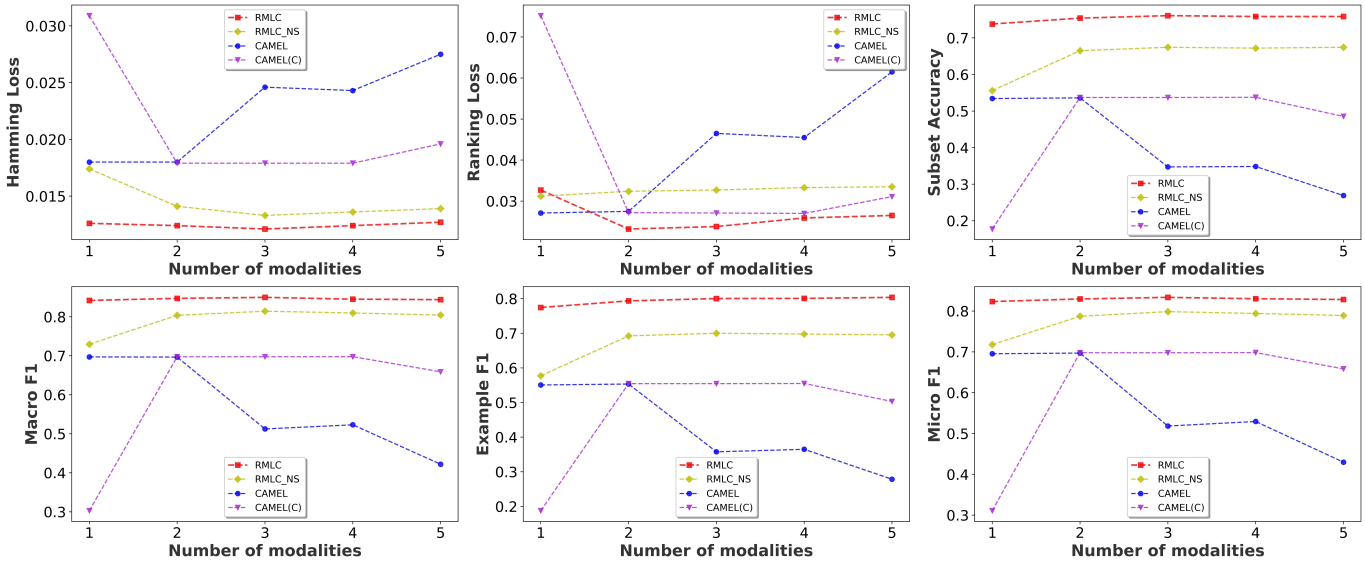
Based on the experimental results in Table III, we obtain the following observations:

- From the results of CAMLE(C) approach, it is obvious shown that roughly concatenating all modalities as a single modality may not always be the wise choice, which indicates that it is necessary to extract information of each individual modality separately.
- RMLC approach achieves the best performance compared with several state-of-the-art approaches on the benchmark datasets, which shows priority of our proposed RMLC approach in multi-modal multi-label learning problem. The main reason is that RMLC takes advantage of multiple modal-specific information, as well as exploiting label correlations to polish multi-label prediction with the help of cross-modal interaction.

1) Influence of label correlations: RMLC improves multi-label prediction in a rethink manner, which makes use of previous modal information to better characterize the relationship among different labels. In order to validate the effectiveness of label correlations across different modalities, we keep the basic structure of RMLC and only adopt $\sigma(\boldsymbol{f}_{spec}^{M})$ as the final prediction, which is denoted as RMLC_NS. As shown in Table IV, RMLC performs better than RMLC_NS, which



(a) *MSRC* with 3 modalities



(b) *FCVID* with 5 modalities

Fig. 3. Performance of RMLC, RMLC_NS, CAMEL and CAMEL(C) with increase of the modality on the *MSRC* and *FCVID* dataset. At t-th step, CAMEL adopts t-th modality as input, while CAMEL(C) concatenates t previous modalities as input.

demonstrates the effectiveness of exploiting label correlations. In other words, it is not enough to merely fuse modal-specific information. RMLC is of great effectiveness to integrate the learned label correlations into the desired label prediction.

Nevertheless, RMLC_NS performs better than other state-of-the-art approaches shown in Table III, which validates the effectiveness of extracting specific information of each individual modalities sequentially.

2) Convergence: We conduct experiments to investigate convergence of RMLC. Due to page limit, we only report the experimental results on the FCVID datasets on all evaluation metrics. The curves in Fig. 2 illustrate the change of metrics results with the number of epoch. It is obvious that both RMLC and RMLC_NS can converge quickly with a few number of epochs, while RMLC performs better than RMLC_NS.

3) Influence of modal-specific information exploitation: To further examine the effectiveness of modal-specific information exploitation, we output label prediction of RMLC at each step. In addition, we conduct experiment on each modality with state-of-the-art multi-label approach CAMEL. In Fig. 3, the curves show the change of the metric results of RMLC, RMLC_NS, CAMEL and CAMEL(C) with previous t modalities. The experimental results in Fig. 3 can not only validate the favorable performance of RMLC in exploiting each individual modality, but also further confirm the contribution of previous t modalities.

V. Conclusion

In this paper, a novel end-to-end network based approach is proposed to solve the multi-modal multi-label problem. Specifically, we enhance the communication among different modalities while remaining modal-specific characteristics. What's more, we exploit label correlations with the collaboration of modal-specific and cross-modal interaction. Experiments on several benchmark multi-modal multi-label datasets demonstrate the superiority of our proposed RMLC over related competitive approaches. Different modalities have different degree of consistency and complementarity, thus it will be an interesting work to exploit label correlations into reinforcement learning environment in the future.

Acknowledgment

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] Q. Tan, G. Yu, J. Wang, C. Domeniconi, and X. Zhang, "Individuality- and commonality-based multiview multilabel learning," *IEEE transactions on cybernetics*, 2019.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 26, no. 8, p. 1819, 2014.

[4] E. Gibaja and S. Ventura, "A tutorial on multilabel learning," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, p. 52, 2015.

[5] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[6] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, p. 333, 2011.

[7] S.-J. Huang and Z.-H. Zhou, "Multi-label learning by exploiting label correlations locally," in *Twenty-sixth AAAI conference on artificial intelligence*, 2012.

[8] J. Huang, G. Li, Q. Huang, and X. Wu, "Learning label-specific features and class-dependent labels for multi-label classification," *IEEE transactions on knowledge and data engineering*, vol. 28, no. 12, pp. 3309–3323, 2016.

[9] Y. Zhu, J. T. Kwok, and Z.-H. Zhou, "Multi-label learning with global and local label correlation," *IEEE Transactions on Knowledge & Data Engineering*, no. 6, pp. 1081–1094, 2018.

[10] L. Feng, B. An, and S. He, "Collaboration based multi-label learning," in *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019, pp. 3550–3557.

[11] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *International Conference on Machine Learning*, 2014, pp. 595–603.

[12] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International Conference on Machine Learning*, 2015, pp. 1083–1092.

[13] F. Wu, X.-Y. Jing, J. Zhou, Y. Ji, C. Lan, Q. Huang, and R. Wang, "Semi-supervised multi-view individual and sharable feature learning for webpage classification," in *The World Wide Web Conference*. ACM, 2019, pp. 3349–3355.

[14] X. Wu, Q.-G. Chen, Y. Hu, D. Wang, X. Chang, X. Wang, and M.-L. Zhang, "Multi-view multi-label learning with view-specific information extraction," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 3884–3890.

[15] Z. He, C. Chen, J. Bu, P. Li, and D. Cai, "Multi-view based multi-label propagation for image annotation." vol. 168, no. C, pp. 853–860, 2015.

[16] H.-J. Ye, D.-C. Zhan, X. Li, Z.-C. Huang, and Y. Jiang, "College student scholarships and subsidies granting: A multi-modal multi-label approach," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 2016, pp. 559–568.

[17] Y. Xing, G. Yu, C. Domeniconi, J. Wang, and Z. Zhang, "Multi-label co-training," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 2882–2888.

[18] Y. Zhang, C. Zeng, H. Cheng, C. Wang, and L. 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)*. IEEE, 2019, pp. 838–843.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[20] M. L. Zhang and Z. H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[21] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 4, pp. 754–766, 2010.

[22] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 40, no. 2, pp. 352–364, 2018.

[23] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 33, no. 2, pp. 353–367, 2010.

[24] Y. Yang, D.-C. Zhan, Y. Fan, and Y. Jiang, "Instance specific discriminative modal pursuit: A serialized approach," in *Asian Conference on Machine Learning*, 2017, pp. 65–80.