M3LA: A Novel Approach Based on Encoder-Decoder with Attention Framework for Multi-modal Multi-label Learning

1st Yinlong Zhu
Computer Science and Technology
Nanjing University
Nanjing, China
zhuyinlong@smail.nju.edu.cn

2nd Yi Zhang
Computer Science and Technology
Nanjing University
Nanjing, China
njuzhangy@smail.nju.edu.cn

***Abstract*—With the exponential growth of digital multimedia resources, in the real-world, most of the data are represented as a multi-modal form and usually with multiple semantic labels. Nowadays, Multi-modal Multi-label learning has become a very hot topic. However, previous methods either have not considered the relation between modalities and labels or the correlation among labels. In this paper, we considered the following three questions: (1) How to model the correlation among labels? (2) Is there a correlation between modality and label? (3) Whether the modal input order affects the prediction of individual instance, and how to find the most appropriate modal input sequence for each instance? To solve above problems, we proposed a novel method for Multi-modal Multi-label learning(MMML), which based on Encoder-Decoder with attention framwork named MMML-Attention(M3LA). The M3LA takes into account all of these issues. Specifically, benefit from the Encoder-Decoder with attention structure, on the one hand, M3LA can model the relation between modalities and labels. On the other hand, we introduce a correlation matrix to learn the correlation among labels, which can be obtained as parameter through the training process. It should be mentioned that label prediction occurs at every step of the decoder, and the prediction of the label is constantly corrected and then the most accurate prediction is obtained. To validate the effectiveness of the proposed method, we expermiented on widely used several benchmark datasets and compared with state-of-art approaches.**

***Index Terms*—multi-label, multi-modal, classfication, machine learning, deep learning**

I. Introduction

In traditional supervised learning, one object is represented by a single modality and associated with only one label. However, in real world, a complex object is often composed of multiple modalities and has multiple semantic labels. e.g. a news report contains multimodal information such as text, images and even audio, it can belong to different tags, such as entertainment, sports, politics etc. Multi-modal Multi-label(MMML) [1] [2] [3] learning provides a framework for such complex objects. A multi-modal multi-label problem has following settings: 1) the set of labels is predefined, meaningful, and human-interpretable. 2)the number of labels is limited in scope and not greater than the number of attributes. 3) Each training example has at least two or more modalities, and is associated with several labels of the label set. 4)labels may be correelated. Solving a problem with multi-label multi-modal data involves many challenages, e.g. how to deal with the correlation among labes and how to get the most appropriate input order of modalities?

The Encoder-Decoder structure is also called the Sequence to Sequence structure which is a variant of the RNN [4] [5] [6]. It has a wide range of applications, such as neural machine translation, text summarization, reading comprehension, speech recognition etc. But there is a issue that the encoder needs to compress all information of sourece sentence into a fixed-length vector. To solve this issue, Dzmitry et al. introduced the attention structure to the Encoder-Decoder model [5]. Inspired by human attention mechanisms, attention structure was first proposed in the field of image recognition [7] which greatly improves the accuracy of image recognition. Hereafter, researchers have used it as a general structure for various fields due to its excellent performance. In this paper, We innovatively use the attention based encoder-decoder model for multi-modal multi-label learning.

For Multi-modal Multi-label learning, during the past years, researchers have proposed a series of methods. However, there are some problems with these methods that either without considering the correlation among labels or the relation between modalities and labels. In addition, these methods input the modality sequentially when predicting the label, rather than selectively inputting the appropriate modality.

In this work, aiming at solving above problems, we proposed a novel Multi-modal Multi-label learning method based on Encoder-Decoder with attention framework(M3LA). We consider that, for each instance, different input order of modal will have an impact on the final result. In other words, to predict an unseen instance as efficient and accurate as possible, we need to figure out a modal extraction for the concerned instance individually [8]. Benefit from the Encoder-Decoder with attention structure, M3LA can model the relation between modality and labels. It has a modality prediction layer that can always choose the most appropriate modality at each step when predicts the label based on result of previous

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

step. Meanwhile, the correlation among labels should also be considered. Based on this reason, M3LA introduces the correlation matrix W_{co} to model this correlation. The correlation matrix is obtained as parameters of the neural network through training. Unlike these classic multi-label methods, it does not require any prior knowledge and it is constantly updated. After training phase, the correlation matrix W_{co} stored information about correlation among labels. Due to the label correlation matrix has been introduced to M3LA and it always choose the most appropriate modal to input the model at each step, in our model, the prediction of the label is constantly corrected, finally, the most accurate prediction is obtained.

In general, the main contributions of this paper are summarized in following four points:

1) We innovatively combined the Multi-modal Multi-label learning with Encoder-Decoder with attention structure, and proposed a novel approach named M3LA which can model the relation between the modality and labels by choosing the most appropriate modal each time according to the modality prediction layer.
2) We also considered the label correlation by introducing the correlation matrix W_{co} which can be obtained as a parameter through training, unlike previous classic multi-label methods, it does not require any prior knowledge and it is constantly updated.
3) The prediction process in M3LA is also innovative. It is continuously corrected using historical information and newly entered information.
4) We conducted a exhaustive experiment on several benchmark datasets comparing to state-of-art methods, and comprehensively evaluation on the performance, obtaining consistently superior performance stably.

Section II gives the related work, and our approach is detailed in Section III. Then Section IV reports our exhaustive experiments and results. Finally, Section V gives the conclusion.

II. RELATED WORK

The exploitation of multi-modal multi-label learning has attracted much attention recently. Our method focuses on the popular Seq2Seq with attention framework to model the multi-modal multi-label classfication problems, which innovatively considers the relation between modality and labels. Meanwhile, we also overcome the problem of correlation between labels by introducing the correlation matrix. Therefore, our work is related to the multi-modal multi-label learning. In this Section, we are going to give a brief review about multi-modal multi-label learning.

Many classic algorithms have been proposed by researchers in the past few decades, a classic multi-label method named binary relevance(BR) [9] generates one binary dataset for each label in which positive patterns are those predicting the label, and the rest are considered to be negative patterns. But different from our method, BR assumes labels are independent which did not consider the correlation between the labels. Another classic method called label combination(LC) [10] can model label correlations in the training data which generates a new class for each possible combination of labels. However, this appoach only takes into account the distinct labelsets in the training set, so it cannot predict unseen lablesets that may also lead to a tendency to overfit the training data [11]. Classifier Chains(CC) [10] is a chain of binary classfication method. It considers correlation among labels in a random manner. In addition, due to the order of chain itself can influence the performance, the authors proposed a Ensemble of Classifier Chains(ECC) which performs strongly. However, Like the Multi-label decision Tree(ML-DT) [12] which scale of binary classifiers constrcted is quadratic because it builds classifiers for any pair of class labes, the time complexity is unacceptable, especially for high-dimensional data. In our model, the contribution of introducing the correlation matrix lies in the way to obtain it. The correlation matrix is obtained as parameters of the neural network through training. Unlike these classic multi-label methods, it does not require any prior knowledge and it is constantly updated.

Thereafter, with the development of deep learning, some researchers have proposed multi-modal and multi-label learning models based on deep learning. Nguyen proposed multi-modal multi-label Latent Dirichlet Allocation(M3LDA), it provides a promising way to understand the relation between input patterns and output semantics [13]. Zhang proposed a extends algorithm based on Classifier Chians named Multi-modal Classifier Chains(MCC) [1] which can make each modality interactions. Discriminative Modal Pursuit(DMP) is a end-to-end serialized adaptive decision approach based on neural network that aims to reduce the modal extraction cost. Meanwhile, DMP can balance the classification performance and modal feature extraction cost [8]. However, they didn't consider the correlation of labels. A Deep multi-modal CNN network named MMCNN-MIML proposed by Song [14], it considered the label correlations by grouping labels in its later layers. But it didn't consider the input order of the modalities. We believe that different modal input order for different samples will affect the final prediction result. Multi-modal Multi-label Multi-instance Deep Network(M3DN) proposed by Yang [3]. It learns the label prediction and exploits label correation simultaneously based on the Optimal Transport. Ye proposed CS3G algorithm which handles types of interactions between multiple label, but there is no interaction between features from different modalities [3]. Although deep learning-based methods can effectively improve performance, previous methods don not work well for multi-label multi-modal learning.

III. PROPOSED METHOD

This section mainly gives the detailed description of the M3LA appoach after a preliminary notation explanation.

A. Notation

Suppose $\mathcal{X} \in \mathbb{R}^d$ denotes the feature of instances space, and $d = d_1 + d_2 + ... + dm$, where m is the number of modalities, and d_k represents the dimensional of the k-th modalities. $\mathcal{Y} =$

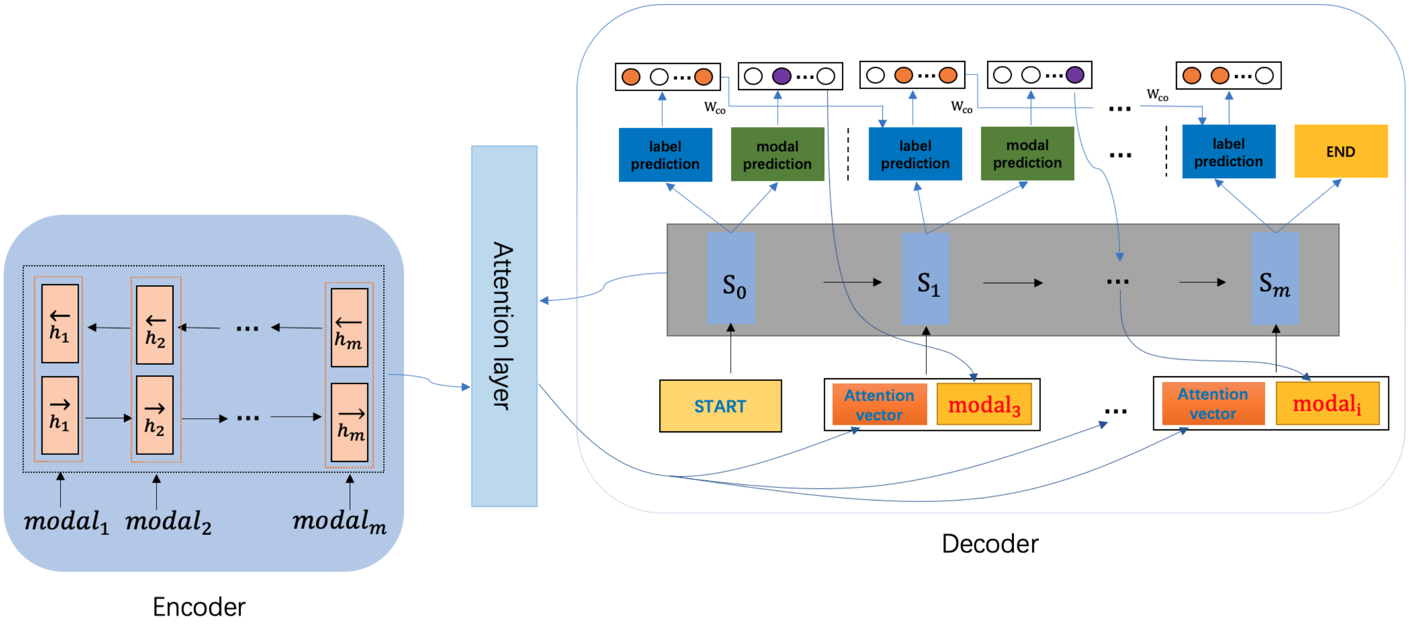

Fig. 1. The structure of our propsed approach M3LA based on Encoder-Decoder with attention framework

$\{0,1\}^q$ denotes the label space with q possible class labels. For the label of instance i, $y_i \in \mathcal{Y}$, and $y_i \in \mathbb{R}^q$, the $y_{ij} = 1(0 \le j \le q)$ represents the instance i belong the j-th class, $y_{ij} = 0$ otherwise. The goal of Multi-modal Multi-label learning is to learn a function $H : \mathcal{X} \rightarrow \mathcal{Y}$ from the training set $\mathcal{D} = \{(x_i, y_i)|1 \le i \le N\}$, where N represents the number of the training set instances. for each Multi-modal Multi-label example (x_i, y_i), $x_i \in \mathcal{X}$ is a d-dimensional feature vector, in order to facilitate the discussion, we represent the example x_i as $\{x_i^1, x_i^2, ..., x_i^m\}$, where x_i^k is k-th modality of x_i with d_k dimension. Thus, there are m disjoint modal feateures for each examples. $y_i \in \mathcal{Y}$ is a q dimension label vector associated with x_i. Let $\mathcal{T}$ denote the test set, for any unseen instance $x \in \mathcal{T}$, the Multi-modal Multi-label classifier $H(.)$ predicts $H(x) \in \mathcal{Y}$ as proper lables for x.

B. M3LA approach

Firstly, we briefly describle the Encoder-Decoder with attention framework, and upon which we build a novel architecture for Multi-modal Multi-label Learning. And the overall framework is shown in figure 1

In the Encoder-Decoder with attention framework, The main task of the ecoder is to encode the information of the modalities, it reads the input modalities sequence of instance, i.e a sequence of vectors $\{x_i^1, x_i^2, ..., x_i^m\}$, obtain the hidden state of each modality. Considering the different dimensions of each modality, we modify the input modality $\tilde{x}_i^k = [0, 0, ..., x_i^k, ..., 0]$, Formally, at step t,

$$\overrightarrow{h}_t = \overrightarrow{F_{en}}(\tilde{x}_i^t, \overrightarrow{h}_{t-1}) \quad (1)$$

$$\overleftarrow{h}_t = \overleftarrow{F_{en}}(\tilde{x}_i^t, \overleftarrow{h}_{t-1}) \quad (2)$$

$$h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \quad (3)$$

which $h_t \in \mathbb{R}^{2h}$ is the combination of forward hidden state and backwrad hidden state at step t, h is the hidden state dimension which can be set manually. $\overrightarrow{F_{en}}$, $\overleftarrow{F_{en}}$ are some nolinear functions, in this paper, we use the GRU [6] as $\overrightarrow{F_{en}}$, $\overleftarrow{F_{en}}$. The s_0 is the initial hidden state of decoder.

$$s_0 = g(h_1, ..., h_m) * W_{init} + b_{init} \quad (4)$$

Where $g(h_1, ..., h_m) = h_m$, and $W_{init} \in \mathbb{R}^{2h \times h}$ is the parameter which initializes the initial hidden state of decoder and $b_{init} \in \mathbb{R}^h$ is the bias vector.

The attention mechanism is an important part of our model, because it always focuses on the information that the model needs to. The detailed attention layer is shown in figure 2 . And the attention vector can be computed as following:

$$a_i = \sum_{j=1}^{m} \alpha_{ij} h_j \quad (5)$$

the weight α_{ij} of each state h_j is computed by

$$\alpha_{ij} = \frac{exp(r_{ij})}{\sum_{k=1}^{m} exp(r_{ik})} \quad (6)$$

where

$$r_{ij} = a(s_{i-1}, h_j) \quad (7)$$

s_{i-1} is represented as hidden state of decoder at step $i-1$, and $a(.)$ is a score function which scores how well the encoder state h_j and decoder state s_{i-1} match.

The decoder, in our proposed approach, consists of two important components: modality prediction and label prediction.

Modality Prediction For a specific instance, we think that different input order of modal may have an impact on the final result. therefore, we need consider both the relation between the modal and labels and the input sequence of modalities when predict labels of instance. In the M3LA, the decoder has a modality prediction layer which can always choose the most appropriate modality to predict label at each step. In other words, modality prediction layer can model the relation between modality and labels. The main task of modality prediction layer is to extract an appropriate sequence of modalities $[o^1, o^2, ..., o^m]$ for label prediction. In details, at step t, for example, firstly, the decoder computes the attention vector according to (5), and then calculates the hidden state s_t of decoder, finally connect with a fully connected layer to get the prediction of modality. Formally,

$$s_t = F_{de}(x^t, s_{t-1}, a_t) \tag{8}$$

$$o^{t+1} = \arg\max \ \sigma(W_{modal} * s_t) + b_{modal} \tag{9}$$

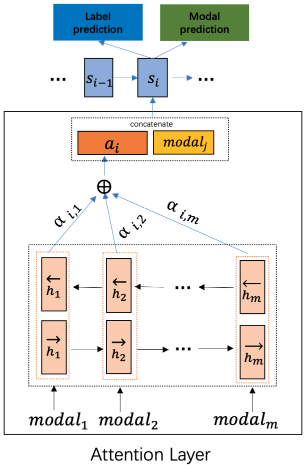

Fig. 2. Attention layer in M3LA

where σ can be any activation function, in this work, we use the softmax as the activation function. $W_{modal} \in \mathbb{R}^{m \times h}$ is the parameter of modality prediction fully connected layer, and $b_{modal} \in \mathbb{R}^m$ is the bias vector of modality prediction layer. F_{de} is same as $\overrightarrow{F_{en}}$ which added attention vector. o^{t+1} is the index of modality input at step $t+1$.

So, x_i^{t+1} denotes the input modality of instance i at step $t+1$, which is represented as following:

$$x_i^{t+1} = [0, 0, ..., x_i^{o^{t+1}}, ..., 0] \tag{10}$$

It is notable that the o^{t+1} may have been chosen at previous step, in order to ensure that each modality is selected, our selection way is to select the one with the highest probability among the modalities that are not selected.

Label prediction Similar to modality prediction, the decoder has a fully connected Label prediction layer. To consider the correlation among the labels, we introduce a correlation matrix $W_{co} \in \mathbb{R}^{q \times q}$ which is a parameter obtained through the training step. After training phase, the correlation matrix W_{co} stored all information about the relation among labels, e.g. $W_{co}[ij] > 0$ indicates that label i has a positive effect on label j, and $W_{co}[ij] < 0$ indicates that label i has a negative effect on label j. At step t, the final label prediction is the product of the prediction of the previous step and the correlation matrix W_{co} plus the result of the current step. Formally,

$$\hat{y}_t = \begin{cases} \sigma(W_{label} * s_t + b_{label}), & t = 1 \\ \sigma(W_{label} * s_t + \hat{y}_{t-1} * W_{co} + b_{label}), & t > 1 \end{cases} \tag{11}$$

Same as above, σ is a activation function which we use the softmax function in this work, $W_{label} \in \mathbb{R}^{q \times h}$ is a parameter of the label prediction layer, and b_{label} is the bias vector, s_t can be obtained according to (8), $\hat{y}_{t-1} \in \mathbb{R}^q$ is the prediction at step $t-1$.

The process of predicting label in our model is constantly corrected, and it thinks like human, corrects the predicted value of the label after selecting the input modality at every step. The final label prediction is

$$\hat{y} = \hat{y}_m \tag{12}$$

It is worth noting that the model returns $\hat{y}$ which is a real-valued respренting probability for each label , in order to decide the proper label for each instance, the real-valued output $\hat{y}$ on each label should be calibrated. Instead of setting a fixed threshold, a method set the threshold t by following:

$$t = \underset{t \in \{0,00,0,001,...,1.00\}}{\arg\min} |LCard(\mathcal{D}) - LCard(H_t(\mathcal{T}))| \tag{13}$$

Label Cardinality(LCard) is a standard measure of ”multi-labelled-ness” introduced in [10], it is simply the average number of labels relevant to each instance, it can be computed by

$$LCard(\mathcal{D}) = \frac{\sum_{i=1}^{|\mathcal{D}|} |y_i|}{|\mathcal{D}|} \tag{14}$$

and $H_t(\mathcal{T})$ is the prediction for test set $\mathcal{T}$ under threshold t.

we consider that, the better performance can be obtained if we set different threshold for each label, therefore, we introduce a measure based on LCard by following:

$$L(\mathcal{D}, j) = \frac{\sum_{i=1}^{|\mathcal{D}|} |y_i[j]|}{|\mathcal{D}|} \tag{15}$$

where $y_i[j]$ is the j-th label of y_i, which equal to 0 or 1. According to (13), (14) and (15), the threshold for each label is,

TABLE I
DETAILS OF DATASETS. N, L, M RESPRENT THE NUMBER OF INSTANCE, LABELS, MODALITIES FOR EACH DATASETS RESPECTIVELY, AND D DENOTE THE DIMENSION OF EACH MODALITY.

DataSets Name	N	L	M	D
FCVID	4388	28	5	400,400,400,400,400
MSRA	15000	50	7	256,225,64,144,75,128,7
ML2000	2000	5	3	500,1040,576
TAOBAO	2079	30	4	500,48,81,24

$$t_i = \underset{t_i \in \{0.00, 0.01, ..., 1.00\}}{\arg\min} \lambda_1 |LCard(\mathcal{D}) - LCard(H_{t_i}(\mathcal{T}))| + \lambda_2 |L(\mathcal{D}, i) - L(H_{t_i}(\mathcal{T}), i)| \tag{16}$$

Where the λ_1, λ_2 are trade-off parameters that indicate which part we prefer focus on. Then,

$$\hat{y}[i] = \begin{cases} 0, & \hat{y}[i] < t_i \\ 1, & \hat{y}[i] \geq t_i \end{cases} \tag{17}$$

Algorithm 1 The persudo code of M3LA

```
Input: Training dataset $\mathcal{D}$
  hyperparameters $\lambda_1, \lambda_2, \lambda_3$
  the number of modalities $m$
  the number of epoches $N_{epo}$
Output: classifier $H(.)$
  for $i = 0 \to N_{epo}$ do
    for $j = 0 \to m$ do
      get $h_j$ by equations (1), (2), (3)
    end for
    initial hidden state $s_0$ of decoder by (4)
    for $t = 0 \to m$ do
      compute attention vector $a_t$ by (5)
      calculate decoder hidden state $s_t$ by 8
      get label prediction $\hat{y}^t$ at step $t$ by (11) and next input modality $x_i^{t+1}$ by 9, (10)
      compute weight $W_{loss}^{(t)}$ for each label by (18)
    end for
    compute Loss by (20)
    compute the derivative $\frac{\partial Loss}{\Phi}$
    update all parameters in $\Phi$
  end for
```

Loss Function Considering the cost-sensitive of each label in Multi-label classification problem [15], we set the importance of each label differently by measuring the difference between the label predicted correctly and incorrectly. At step t, $\hat{y}^{(t)}[k]_0$, $\hat{y}^{(t)}[k]_1$ represent the predict vector $\hat{y}^{(t)}$ when the k-th label is set to 0 and 1 respectively. In addition, for $t = 1$, the weight of each label is set to 1. Formally,

$$W_{loss}^{(t)}[k] = \begin{cases} 1, & t = 1 \\ |K(y, \hat{y}^{(t-1)}[k]_0) - K(y, \hat{y}^{(t-1)}[k]_1)|, & t > 1 \end{cases} \tag{18}$$

where $k = 1, ..., q$, y is the real label of instance, and function $K(.)$ is a cost function that measures the difference between the label predicted correctly and incorrectly.

Like most Multi-label algorithm, we use the Cross Entropy Loss as the loss function.

$$loss = -[y \log p(\hat{y}) + (1 - y) \log p(1 - \hat{y})] \tag{19}$$

Finally, our loss function is showed as (20)

$$Loss = \sum_{t=1}^{m} \sum_{j=1}^{q} -W_{loss}^{(t)}[j](y[j] \log p(\hat{y}^{(t)}[j]) + (1 - y[j]) \log(1 - p(\hat{y}^{(t)}[j])) + \lambda_3 ||\Phi||_2^2 \tag{20}$$

Where Φ represents all parameters in our model, and λ_3 is the regularization factor.

The specific process of the M3LA is summarized in persudo.

IV. EXPERIMENTS

In this section, we validate the effectiveness of proposed M3LA approach by experimenting on 4 widely used benchmark datasets for multi-modal multi-label learning and compare with state-of-art approaches.

A. Datasets and Configurations

M3LA approach innovatively introduces the Encoder-Decoder with attention structure into multi-modal multi-label learing that can not only model the relation between labesl and modalities, but also the correlation among labels. We experiment on 4 public real-world datasets, i.e. FCVID [16], MSRA [17], ML2000 [18], TAOBAO [2], and we multi-modally process these datasets by following [2]. In detail, FCVID is the Fudan-Columbia Video Datasets, which consists of 4388 videos and 28 categories, the 5 operations we conducted are HOF, HOG, CNN, SIFT and Trajectory, then we use PCA to reduce the dimension of each modality to 400. MSRA is a object recognition database which contains 15000 instances and 50 categories, its 7 modalities including 256 RGB color histogram, 225 dimension block-wise color moments, 64 HSV color histogram, 144 color correlogram, 75 distribution histogram, 128 wavelet features and 7 face features. ML2000 is an image datasets for natural scene classfication which consists of 2000 images and 5 categories, we extract BoW, FV and HOG features from each images. TAOBAO is a shopping items classfication Dataset from China E-commerce platform which contains 2079 instances and 30 categories, we also conduct 4 modalities from BoW, Gabor, HOG, HSVHist. All information of datasets are shown in Table I.

TABLE II

Result(mean±std) of M3LA, and compared approach on 4 benchmark datasets. Including 10 commonly used evaluation criteria on multi-label learning, we bolded the best resutl of each criteria, and ↑ / ↓ indicate the lower / higher value of indicator, the better performance.

methods	macro AUC ↑				macro F1 ↑			
	FCVID	MSRA	ML2000	TAOBAO	FCVID	MSRA	ML2000	TAOBAO
BR	0.793±0.009	0.627±0.006	0.793±0.017	0.682±0.020	0.674± 0.015	0.124±0.003	0.647±0.019	0.148±0.007
CC	0.726±0.011	0.573±0.005	0.789±0.014	0.540±0.012	0.565± 0.020	0.118±0.006	0.644±0.019	0.088±0.023
DMP	0.982±0.002	0.879±0.001	0.943±0.010	**0.882±0.023**	0.740±0.018	0.096±0.009	0.782±0.019	0.169±0.058
CS3G	**0.990±0.005**	0.754±0.007	0.924±0.009	0.754±0.028	0.690±0.015	0.071±0.001	0.742±0.022	0.147±0.017
MCC	0.922±0.004	0.766±0.003	0.898±0.012	0.826±0.013	0.646±0.016	0.073±0.004	0.781±0.022	0.228±0.028
M3LA	0.972±0.007	**0.885±0.004**	**0.944±0.010**	0.875±0.025	**0.765±0.016**	**0.157±0.009**	**0.811±0.018**	**0.273±0.052**

methods	micro AUC ↑				micro F1 ↑			
	FCVID	MSRA	ML2000	TAOBAO	FCVID	MSRA	ML2000	TAOBAO
BR	0.792±0.012	0.642±0.005	0.794±0.017	0.746±0.007	0.688± 0.019	0.148±0.002	0.651±0.019	0.201±0.009
CC	0.722±0.012	0.626±0.010	0.789±0.015	0.581±0.022	0.601± 0.022	0.204±0.009	0.646±0.019	0.192±0.043
DMP	0.975±0.002	0.882±0.002	**0.952±0.003**	0.871±0.012	0.759±0.013	0.395±0.021	0.788±0.019	0.381±0.060
CS3G	0.978±0.005	0.872±0.002	0.925±0.009	0.849±0.010	0.723±0.015	0.418±0.004	0.744±0.021	0.336±0.019
MCC	**0.981±0.003**	0.882±0.002	0.932±0.010	0.862±0.010	0.667±0.015	0.405±0.001	0.767±0.023	0.356±0.009
M3LA	0.973±0.003	**0.892±0.003**	0.939±0.004	**0.889±0.012**	**0.791±0.012**	**0.432±0.010**	**0.812±0.018**	**0.476±0.049**

methods	example AUC ↑				example F1 ↑			
	FCVID	MSRA	ML2000	TAOBAO	FCVID	MSRA	ML2000	TAOBAO
BR	0.790±0.012	0.639±0.007	0.807±0.014	0.745±0.006	0.565± 0.023	0.160±0.004	0.667±0.019	0.247±0.013
CC	0.720±0.012	0.619±0.008	0.801±0.014	0.583±0.021	0.438± 0.025	0.182±0.007	0.659±0.023	0.188±0.039
DMP	0.972±0.003	0.873±0.004	0.923±0.014	0.865±0.011	0.682±0.016	0.298±0.023	0.775±0.020	0.283±0.063
CS3G	**0.974±0.006**	0.863±0.004	0.907±0.012	0.839±0.010	0.679±0.016	0.368±0.003	0.705±0.024	0.343±0.023
MCC	0.968±0.016	0.853±0.001	0.887±0.014	0.829±0.010	0.646±0.016	**0.376±0.002**	0.765±0.022	0.346±0.024
M3LA	0.966±0.003	**0.878±0.001**	**0.932±0.012**	**0.866±0.024**	**0.729±0.016**	0.352±0.014	**0.818±0.021**	**0.425±0.047**

methods	RankingLoss ↓				HammingLossm ↓			
	FCVID	MSRA	ML2000	TAOBAO	FCVID	MSRA	ML2000	TAOBAO
BR	0.416±0.024	0.625±0.010	0.337±0.021	0.465±0.011	0.019± 0.001	0.445±0.006	0.212±0.013	0.197±0.012
CC	0.557±0.025	0.678±0.015	0.344±0.023	0.802±0.041	0.021± 0.001	0.158±0.002	0.216±0.012	0.059±0.003
DMP	0.027±0.003	0.126±0.004	0.076±0.014	0.134±0.011	0.016±0.001	0.048±0.001	0.101±0.009	**0.032±0.001**
CS3G	0.025±0.006	0.136±0.004	0.092±0.012	0.160±0.010	0.020±0.001	0.064±0.001	0.119±0.010	0.073±0.002
MCC	0.051±0.006	0.195±0.005	0.082±0.011	0.230±0.024	0.026±0.001	0.047±0.001	0.105±0.012	0.063±0.002
M3LA	**0.023±0.003**	**0.121±0.005**	**0.067±0.012**	**0.133±0.024**	**0.014±0.001**	**0.044±0.001**	**0.092±0.010**	0.033±0.002

methods	Coverage ↓				SubAcc ↑			
	FCVID	MSRA	ML2000	TAOBAO	FCVID	MSRA	ML2000	TAOBAO
BR	12.309±0.667	36.008±0.530	2.580±0.106	14.653±0.334	0.525± 0.021	0±0	0.354±0.034	0.031±0.012
CC	16.15±0.685	41.88±0.719	2.62±0.094	24.28±1.198	0.422± 0.023	0±0	0.341±0.036	0.169±0.036
DMP	1.790±0.096	12.683±0.303	1.553±0.060	5.062±0.329	0.650±0.015	0.067±0.008	0.635±0.015	0.267±0.064
CS3G	1.726±0.168	13.481±0.317	1.636±0.063	5.772±0.315	0.575±0.019	0.053±0.003	0.573±0.028	0.072±0.019
MCC	**1.618±0.123**	16.481±0.254	1.548±0.013	5.223±0.212	0.524±0.020	0.076±0.008	0.663±0.032	0.219±0.027
M3LA	1.953±0.102	**12.258±0.287**	**1.532±0.069**	**5.032±0.736**	**0.703±0.017**	**0.096±0.007**	**0.677±0.031**	**0.382±0.043**

TABLE III

To validate effectiveness of modality prediction layer, the result(mean±std) of M3LA, M3LA-order(no modality prediction layer), M3LA-W_{co}(no correlation matrix) and compared approach on ML2000 dataset. Including 10 commonly used evaluation criteria on multi-label learning, we bolded the best resutl of each criteria, and ↑ / ↓ indicate the lower / higher value of indicator, the better performance.

methods	macro-AUC ↑	micro-AUC ↑	example-AUC ↑	macro-F1 ↑	micro-F1 ↑	example-F1 ↑	RankingLoss ↓	hammingLoss ↓	Coverage ↓	SubAcc ↑
DMP	0.943±0.010	**0.952±0.003**	0.923±0.014	0.782±0.019	0.788±0.019	0.775±0.020	0.076±0.014	0.101±0.009	1.553±0.060	0.635±0.015
CS3G	0.924±0.009	0.925±0.009	0.907±0.012	0.742±0.022	0.744±0.021	0.705±0.024	0.092±0.012	0.119±0.010	1.636±0.063	0.573±0.028
MCC	0.898±0.012	0.932±0.010	0.887±0.014	0.781±0.022	0.767±0.023	0.765±0.022	0.082±0.012	0.105±0.012	1.548±0.013	0.663±0.032
M3LA-order	0.917±0.068	0.932±0.031	0.924±0.071	0.753±0.022	0.729±0.013	0.699±0.137	0.075±0.071	0.123±0.058	1.590±0.409	0.560±0.183
M3LA-W_{co}	0.918 ±0.057	0.923 ±0.342	0.928 ±0.056	0.733 ±0.068	0.771 ±0.126	0.775 ±0.138	0.071 ±0.056	0.116 ±0.066	1.583 ±0.381	0.588 ±0.196
M3LA	**0.944±0.010**	0.939±0.004	**0.932±0.012**	**0.811±0.018**	**0.812±0.018**	**0.818±0.021**	**0.067±0.012**	**0.092±0.010**	**1.532±0.069**	**0.677±0.031**

For each dataset, we randomly select 90% for training set, and remaining instances are used for testing, and the batch size is set as 64, the dimension of hidden state is set as 256. Besides, we set the learning rate as 0.01, and we set the decay rate as 0.99 to avoid the excessive learning rate. $\lambda_3 = 0.0001$, $\lambda_1 = \lambda_2 = 0.5$, After dozens of experiments, we suggest that setting the epoch to 300. In order to verify the robustness of our approach, we repeat experiment ten times following above setting on each dataset with the implementation of an environment on NVIDIA 1080TI GPUs server.

B. Evaluation Measures

In order to comprehensively evaluate our approach, we use both the example-based metrics and label-based metrics. example-based metrics work by evaluating the learning algorithm's performance on each test example separately and then returning the mean value across the test data, while label-based metrics work by evaluating the learning algorithm's performance on each class label separately, and then returning the macro/micro-averaged value across all class labels [19].

For example-based metrics, we use Subset Accuracy which evaluates the proportion of correctly classified examples, e.g, for a given multi-label sample, if the predicted label set exactly matches the true label set of the sample, then the sample is considered to be correctly classified. Hamming Loss indicates the proportion of error samples in all labels, and the smaller the value, the stronger the classification ability of the model. Coverage evalutes how many steps are needed on average, to move down the ranked label list so as to cover all the relevant labels of the example. RankingLoss evalutes the fraction of reversely ordered label pairs. In other words, it indicates the case where unrelated label are more relevant than related label.

For label-based metrics, AUC and F1-score are used, AUC is the area under ROC curve, which has the ability to objectively the comprehensive prediction for each category, and F1-score is definited based on harmonic mean of precision and recll, both of them are widely used evaluation indicators.

C. Experiments Results

We compared some traditional and advanced appoaches in multi-modal multi-label learning, e.g. BR [9], CC [10], MCC [1], DMP [8], CS3G [2]. BR and CC are classic multi-label learning algorithm, in this compared experiments, we treat all modalities as a single modal as the input of BR and CC. MCC is a extends algorithm based on Classifier Chians which can make each modality interactions. DMP is a end-to-end serialized adaptive decision approach based on neural network that aims to reduce the modal extraction cost. Meanwhile, DMP can balance the classification performance and modal feature extraction cost. And it is notable that the DMP is a multi-modal single-label algorithm, but it can also be used for multi-modal multi-label learning problem as long as we simply modify the way of label prediction. CS3G is designed for scholarships and subsidies allocation, it is also a general multi-modal multi-label learning algorithm which can handle types of interactions among multiple label.

In order to compare these approaches objectively, we repeated experiment for ten times, the data set division according to IV-A, and average value of each evaluation metrics and standard deviation are shown in Table II.The results shows that, BR and CC, the traditional multi-label algorithm's performance is poor, a major reason is that they do not take into account the relation between modalities and labels. DMP and MCC considered the effect of modalities, and CS3G considered the correlation among labels, their performance is much better than the traditional method. However, our method considers both and it can be found that our method has achieved the best results on almost all dataset with different performance measures, which validates the effectiveness of our appoach solving multi-modal multi-label learning problem. We note that the F1 and SubAcc of the MSRA dataset are lower than other datasets because there are 50 labels , and the criterion for the SubAcc is overly strict that required the prediction results for each of the 50 labels are the same as the ground-truth. In contrast, the ML2000 dataset has only 5 labels and the evaluation results are much better. Another thing that we are surprised about is, for all dataset, our proposed approach has a significant improvement in some performance measures(e.g F1-socre, HammingLoss, SubAcc). This also shows that M3LA approach is a high-competitive multi-modal multi-label learning method.

D. Modality prediction and Label correlation Exploitation

We consider that different input order of modalities can affect performance for different instances, in other words, there is a relation between modality and labels that lables have varying degrees of denpendence on each modality. To verify this, we conducted a comparative experiment on the ML2000 dataset. The results of experiment are shown in Table III, M3LA-order indicates that we input modalities in order in decoding phase. It is obvious that the modality prediction part in our model has a great impact on performance.

Meanwhile, we also conducted a comparative experiment on the ML2000 dataset to identify the impact of the correlation matrix W_{co} on our model. We slightly modified our model where the M3LA-W_{co}indicates the M3LA model without correlation matrix W_{co}. And the results are shown in Table III. It can be found that the performance of the model for removing the correlation matrix W_{co} is far less than the original M3LA model.

V. CONCLUSION

In this work, we innovatively combined the Multi-modal Multi-label learning with Encoder-Decoder structure, and proposed a novel approach named M3LA. On one hand, M3LA can model the correlation among labels that solved the problem of association between labels. On the other hand, benefit from the relation between modality and labels is innovatively considered that the model can always choose the most appropriate modality to predict each label. Experiments on 4 widely used benchmark datasets and comparation with the model which have not modality prediction show the effectiveness of our model. However, in real-world, instances may have different numbers of modalities, so how to extend the appoach to this scenario is a very promising work 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 Collaborative Innovation Center of Novel Software Technology and Industrialization at Nanjing University.

References

[1] Zhang Yi, Zeng Cheng, Cheng Hao, Wang Chongjun and Zhang Lei, “Many could be better than all: A novel instance-oriented algorithm for Multi-modal Multi-label problem”, ICME 2019.

[2] H. Ye, D. Zhan, X. Li and Z. Huang, and Y. Jiang, College Student Scholarships and Subsidies Granting: A Multi-modal Multi-label Approach, 2016 IEEE 16th International Conference on Data Mining (ICDM).

[3] Yang Yang, Yi Feng, DeChuan Zhang and Yuan Jiang, “Complex Object Classification: A Multi-Modal Multi-Instance Multi-Label Deep Network with Optimal Transport. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018.

[4] Ilya Sutskever, Oriol Vinyals and Quoc V. Le, “Sequence to Sequence Learning with Neural Networks,” Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada.

[5] Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate”, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015.

[6] KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau and Yoshua Bengio, On the Properties of Neural Machine Translation: Encoder-Decoder Approaches, CoRR, 2014.

[7] Volodymyr Mnih, Nicolas Heess, Alex Graves and Koray Kavukcuoglu, ”Recurrent Models of Visual Attention”, Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada.

[8] Yang Yang, DeChuan Zhan, Ying Fan and Yuan Jiang, ”Instance Specific Discriminative Modal Pursuit: A Serialized Approach”, Proceedings of The 9th Asian Conference on Machine Learning, ACML 2017, Seoul, Korea, November 15-17, 2017.

[9] Matthew R. Boutell, Jiebo Luo, Xipeng Shen and Christopher M. Brown, ”Learning multi-label scene classification”, Pattern Recognition, vol. 37, 2004.

[10] Jesse Read, Bernhard Pfahringer, Geoff Holmes and Eibe Frank, ”Classifier chains for multi-label classification”, Machine Learning, vol. 85 pp. 333-359.

[11] Eva Gibaja and Sebasti Ventura, ”A Tutorial on Multilabel Learning”, ACM Comput. Surv, vol. 47 pp.1-38, 2015.

[12] Amanda Clare and Ross D. King, ”Knowledge Discovery in Multi-label Phenotype Data”, Principles of Data Mining and Knowledge Discovery, 5th European Conference, PKDD 2001, Freiburg, Germany, September 3-5, 2001, Proceedings.

[13] CamTu Nguyen, DeChuan Zhan and ZhiHua Zhou, ”Multi-Modal Image Annotation with Multi-Instance Multi-Label”, IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013.

[14] L. Song et al., ”A Deep Multi-Modal CNN for Multi-Instance Multi-Label Image Classification”, IEEE Transactions on Image Processing, vol.27, 2018.

[15] ChunLiang Li and HsuanTien Lin, ”Condensed Filter Tree for Cost-Sensitive Multi-Label Classification”, ICML 2014, vol. 32, pp. 423–431.

[16] YuGang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue and ShihFu Chang, ”Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks”, IEEE Trans. Pattern Anal. Mach. Intell. vol. 33, pp. 353–367.

[17] Tie Liu et al., ”Learning to Detect a Salient Object”, ICML 2014, vol. 32, pp. 423–431.

[18] MinLing Zhang and ZhiHua Zhou, ”ML-KNN: A lazy learning approach to multi-label learning”, Pattern Recognition, 2007, vol. 40, pp. 2038–2048.

[19] MinLing Zhang and ZhiHua Zhou, ”A Review on Multi-Label Learning Algorithms”, IEEE Trans. Knowl. Data Eng, 20014, vol. 26, pp. 1819–1837.