Comprehensive Semi-Supervised Multi-Modal Learning

Yang Yang[1] , **Ke-Tao Wang**[1] , **De-Chuan Zhan**[1*] and **Hui Xiong**[2]
Yuan Jiang[1]

[1]National Key Laboratory for Novel Software Technology, Nanjing University
[2]Rutgers University
{yangy, wangkt, zhandc, jiangy}@lamda.nju.edu.cn, hxiong@rutgers.edu

Abstract

Multi-modal learning refers to the process of learning a precise model to represent the joint representations of different modalities. Despite its promise for multi-modal learning, the co-regularization method is based on the consistency principle with a sufficient assumption, which usually does not hold for real-world multi-modal data. Indeed, due to the modal insufficiency in real-world applications, there are divergences among heterogeneous modalities. This imposes a critical challenge for multi-modal learning. To this end, in this paper, we propose a novel Comprehensive Multi-Modal Learning (CMML) framework, which can strike a balance between the consistency and divergency modalities by considering the insufficiency in one unified framework. Specifically, we utilize an instance level attention mechanism to weight the sufficiency for each instance on different modalities. Moreover, novel diversity regularization and robust consistency metrics are designed for discovering insufficient modalities. Our empirical studies show the superior performances of CMML on real-world data in terms of various criteria.

1 Introduction

In most real-world applications, data are collected from diverse sources and exhibit heterogeneous properties, e.g., during driverless driving, cars collect information from different sensors; in medical testing, doctors collect information from different inspections, etc. These variable group information is defined as *Multiple Modality* in literature. In contrast to single modal learning, multi-modal learning mainly exploits the consistent and complementary properties among different modalities and improves the learning performance, which attracts increasing attentions and is widely studied in terms of algorithms and theories [Xu *et al.*, 2013; Sridharan and Kakade, 2008; Wang and Zhou, 2013]. Formally, existing algorithms can be categorized into two classes: 1) co-training style; 2) co-regularize style.

Co-training [Blum and Mitchell, 1998] style methods always use complementary principle to label unlabeled data for each other. Relatively, co-regularize [Brefeld *et al.*, 2006] style methods exploit unlabeled data with consistency principle. In the literature on multi-modal learning, these methods always assume that each modality is sufficient for classification independently. With this assumption, theoretically, Balcan *et al.* [2004] proved that if each modal classifier is never "confident but wrong", the ε-expansion can guarantee the success of co-training; Sridharan and Kakade [2008] presented an information theoretic framework for co-regularization which showed excess error between the output hypothesis of co-regularization and the optimal classifier. Nevertheless, in many practical applications, it is very difficult to meet the sufficiency assumption. Consequently, using the previous multi-modal methods directly will degenerate the performance on the contrary, even without the effect of single modality [Tao *et al.*, 2018]. Considering this realistic problem, Wang and Zhou [2013] proved that if different modalities have large diversities, co-training algorithm may be able to improve the learning performance, while co-regularize based methods will lose efficacy. Yet co-training has a lot of hyper-parameters for adjusting, and it is difficult to satisfy the label noise assumption in reality. Thus, the applicabilities of these style methods are very limited in practice.

In this paper, we are committed to reformulate co-regularize based methods. The modal insufficiency mainly leads to the divergences among different modalities. These are partial instances with consistent modalities, while other instances are with diverse modalities, i.e., the modalities are inconsistent, and exist strong/weak modalities [Yang *et al.*, 2015]. Therefore, forcing the predictions of different modalities to be consistent as previous methods will lead to simply minimize the disagreements rather than the optimal classifiers, result in increasing learning confusion and reducing diversity. And in turn, it will degrade single modal performance. Ensemble results are also affected. To solve this much more challenging but practical problem, we present the analysis on multi-modal learning with insufficient modalities, i.e., the consistent and divergent modalities should be treated differently, and propose a novel Comprehensive Multi-Modal Learning (CMML) framework. Specifically, CMML processes the insufficiency with the instance level attention mechanism. On this basis, a diversity regularization is applied to

*Contact Author

discover complementarities among modalities, and a novel robust consistency metric is designed considering the divergent multi-modal data. In consequence, different modalities can predict accurately for the same task while preserving the diversity.

2 Related Work

Multi-modal approaches aim to improve the single modal and overall performance by using the heterogeneities of different modalities. Crucially, this can be done by using unlabeled data, which is out of reach with single modality. Thus, multi-modal learning always denotes to semi-supervised learning. The majority multi-modal methods can be divided into two categories: co-training style and co-regularize style.

Co-training [Blum and Mitchell, 1998] is one of the earliest multi-modal methods, which uses initial labeled data to learn two weak hypotheses and allows them to label confident instances for each other, thus improving each modal performance. Blum and Mitchell [1998] proved that co-training can boost the performance of weak classifiers to arbitrarily high level with sufficient modalities. This assumption is too strong to meet in real applications, i.e., the classifier may not correctly make prediction. Thus, Wang and Zhou [2013] proved that co-training may also be effective under the insufficient setting with large diversity. Yet the co-training is still difficult for tuning the hyper-parameters and it ignores the consensus principle, which limits the practical applicability of such methods.

Another style methods is co-regularize [Brefeld *et al.*, 2006], which directly minimizes the disagreements over different modal predictions. The intuition of these methods is that the optimal classifiers of various modalities are compatible with each other. Co-regularize style algorithms are widely researched such as co-regularize [Brefeld *et al.*, 2006], ARM [Yang *et al.*, 2015] and SLIM [Yang *et al.*, 2018b]. It is worthy noting that the basic assumption is that different modalities can provide almost the same predictions. Unfortunately, in practical applications, it is unreasonable that each modality can provide sufficient information, in other words, there exist divergences among different modalities. Consequently, the single modal and overall performance may even degrade.

As a matter of fact, learning comprehensive multi-modal methods considering the modal insufficiency is a relatively new topic. Intensively, we propose a novel Comprehensive Multi-Modal Learning (CMML) framework considering both the consistency and diversity properties in an unified framework. CMML can well consider the sufficiency of each instance on different modalities, while novel diversity regularization and robust consistency metric are designed cogently for promoting the performance.

3 Proposed Method

Insufficient multi-modal data leads to the divergences among modalities, thus we need to reflect the diversity and robust consistency among different modalities on the basis of measuring the sufficiency, rather than simply use consistency regularization as before.

3.1 Notations

Without any loss of generality, there are N instances with multi-modal information, including N_l labeled examples, i.e., $X_l = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \cdots, (\mathbf{x}_{N_l}, \mathbf{y}_{N_l})\}$, and N_u unlabeled instances, i.e., $X_u = \{\mathbf{x}_{N_l+1}, \cdots, \mathbf{x}_N\}$. $\mathbf{x}_i = \{\mathbf{x}_i^1, \mathbf{x}_i^2, \cdots, \mathbf{x}_i^M\}$ with M modalities, $m-$th modality is with d_m-dimensional representation, $\mathbf{y} \in \{0,1\}^C$, C is the class number. The goal is to learn M discriminative models for every modalities: $f_m : \mathcal{R}^{d_m} \rightarrow \mathcal{R}^C$. Basically, the learning targets are: 1) better single modal result, the result of each modality will not degrade; 2) better ensemble result, the overall result is better than the best single modal result.

3.2 Sufficiency Measure

The supervise loss term of previous method is always constituted by mean/max voting with different modal classifiers, while is leaves the importance of each instance on different modalities without consideration. Therefore, we utilize the instance level attention mechanism to automatically learn the attention weights. Thus, the loss can be reformulated as:

$$L_s = \sum_{i=1}^{N_l} \sum_{j=1}^{M} \ell(\alpha_{i,j} f_j(\mathbf{x}_i), \mathbf{y}), \tag{1}$$

here considering the distinguished performance of deep models, f utilizes the deep model for each modality in this paper, e.g., convolutional neural network for image modality, long short term memory for text modality. $\alpha_{i,j}$ is the attention weight for $i-$th instance on $j-$th modality, which can be learned by an extra attention network: $\alpha_{i,j} = \frac{h(f_j(\mathbf{x}_i^j))}{\sum_{m=1}^{M} h(f_m(\mathbf{x}_i^m))}$, $h(\cdot)$ is the extra neural network, i.e., we utilize two layer shallow fully connected here. The α can represent the instance level sufficiency if the instance volume is large as [Wang and Zhou, 2013]. At the end of each round, the weights α are normalized as $\sum_j \alpha_{i,j} = 1$. Moreover, it is notable that the attention can also be applied on the feature embedding layer, i.e., $\hat{f} = \alpha_{i,j} \mathbf{x}_{ij}^{l_p} W$, where $\mathbf{x}_{ij}^{l_p}$ is the embedding for $\mathbf{x}_{ij}$ of feature representation layer, W denotes the fully connected matrix to the prediction layer.

3.3 Diversity Measure

The previous co-regularize based methods usually require the prediction probabilities of different modalities to be consistent. While in the insufficient setting, the multi-modal ensemble result using consistency regularization is even worse than single modal result. This is ultimately because minimizing the disagreement among modalities will result in the loss of the divergence. In other words, for the insufficient multi-modal data, partial data of each modality may contain the knowledge that other modalities lack, while the consistency regularization will lose the complementary information instead, i.e., weak modality will affect the strong modality to some extent in the learning procedure. Therefore, similar to ensemble learning, the divergences among different modalities can be regarded as the diversities among different modal classifiers.

Without any loss of generality, given M trained classifiers $F = \{f_m(\mathbf{x}_i)\}_{m=1}^M$ for all modalities. In this work, we

measure diversities of different modalities based on pairwise difference as the diversity [Li *et al.*, 2012], and the definition is given as follows:

Definition 1 *Given a set of N examples $\mathcal{D} = \{\{(\mathbf{x}_i^m, \mathbf{y}_i^m)\}_{i=1}^{N_l}, \{(\mathbf{x}_i^m)\}_{i=N_l+1}^{N}\}_{m=1}^{M}$, the diversities of different modalities on $\mathcal{D}$ is defined as:*

$$Com(F) = \frac{1}{\sum_{1\le i\ne j\le M}} \sum_{1\le i\ne j\le M} sim(f_i, f_j), \quad (2)$$

where $sim(\cdot,\cdot)$ is the pairwise difference between two modalities as:

$$sim(f_i, f_j) = \frac{1}{N}\sum_{k=1}^{N} \cos(f_i(x_k^i), f_j(x_k^j)),$$

where the sim can be any convex function here, we utilize the cosine here, thus the difference $sim(f_i, f_j)$ falls into the interval $[-1, 1]$, and equals to $1/-1$ only if two modal classifiers always make the same/opposite predictions on the same instances, and the larger sim, the larger Com. In consequence, since the diversity is based on the pairwise differences, it reveals that the smaller $Com(F)$, the larger diversity of the modal classifier set F. Moreover, different from defining the diversity in the parameter space [Yu *et al.*, 2011], here the diversity measure defines in the output space, thus can cover various kinds of individual classifier.

On the other hand, it is easy to find that this diversity measure is closely related to complementary measure, which aims to utilize the preponderant information of each modality to improver the performance. Thereby, intergrading multiple modalities with diversity regularization can describe the data more comprehensively and accurately in the insufficient setting. Moreover, the supervised loss and diversity regularization are also adversarial here. In detail, the supervised loss means that the classifiers of different modalities must predict similarly on the same label, while the diversity regularization increases the diversities over various modalities. Thus, considering the tradeoff between empirical error and diversity, better generalization performance can be expected.

3.4 Robust Consistency Measure

The consistency principle maximizes agreement on multiple modalities, and it has been demonstrated the connection between the consensus of two modal hypotheses respectively and their error rates as [Dasgupta *et al.*, 2001]:

$$P(f_1 \ne f_2) \ge \max\{P_{err}(f_1), P_{err}(f_2)\}$$

which reflects that the probability of the disagreement of two independent hypotheses upper bounds the error rate of either hypothesis. Thus by minimizing the disagreement of the two hypotheses, the error rate of each hypothesis will be minimized. However, the basic assumption behind the theory is that each modality is sufficient for prediction, while different modal feature representations are always insufficient in practice. And it has been proved that co-regularization based methods, which utilize the consistency principle, prefers to output false hypotheses rather than output optimal classifiers [Wang and Zhou, 2013].

To solve this problem, we turn to utilize the robust regression loss, i.e., Huber Loss [Huber and others, 1964], instead of the square loss in previous methods. For consistency calculation, we also adopt cosine for convenient as the diversity measure mentioned in last section. The modified huber loss can be defined as:

$$H_\delta(f_i, f_j) = \begin{cases} \frac{1}{2}(2-\cos(f_i,f_j))^2, & |2-\cos(f_i,f_j)| \le \delta \\ \delta|2-\cos(f_i,f_j)| - \frac{1}{2}\delta^2, & otherwise \end{cases} \quad (3)$$

Differing from the squared loss which has a disadvantage of the tendency being dominated by outliers (disagree on different modalities), huber loss is quadratic for small values of $|2-\cos(f_i,f_j)|$, linear for large $|2-\cos(f_i,f_j)|$, with equal values and slopes of the different sections at the two points where $|2-\cos(f_i,f_j)| = \delta$, and it has the differentiable extension, the δ is set as 1 in the experiments. In result, the robust consistency metric will more incline to constrain the instances with consistent modalities, neglecting the inconsistent instances. Thus, with the Eq. 3, the consistency regularization can be rewritten as:

$$R_\delta(F) = \frac{1}{\sum_{1\le i\ne j\le M}} \sum_{1\le i\ne j\le M} H_\delta(f_i, f_j), \quad (4)$$

3.5 Comprehensive Multi-Modal Learning

In this section, we will reformulate the co-regularize based methods in insufficient setting with the three paradigms mentioned above. The basic intuition of previous co-regularize style methods is that the complexity of learning problem can be reduced by eliminating hypotheses from each modality that do not agree [Sridharan and Kakade, 2008]. Representative multi-modal semi-supervised learning method as co-regularize is:

$$\min_{f^j} \sum_{i=1}^{N_l}\sum_{j=1}^{M} \ell(f_j, \mathbf{y}) + \|f_j\|_F^2 + \sum_{k=N_u+1}^{N_l+N_u}\sum_{i,j=1}^{M} \lambda\|f_i - f_j\|_F^2, \quad (5)$$

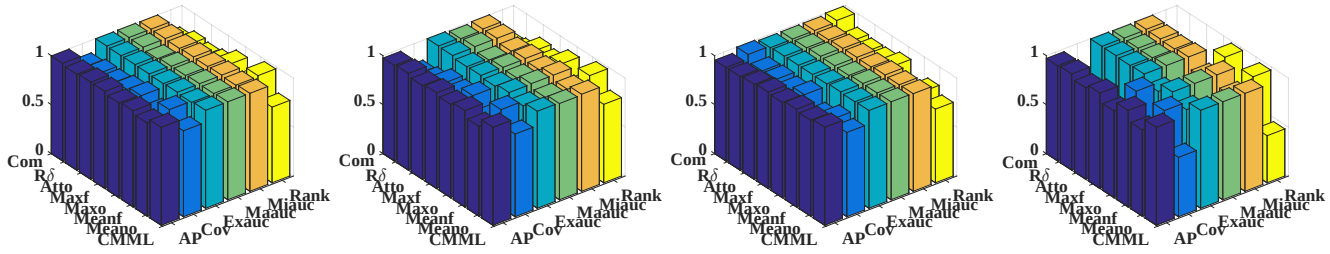
As mentioned above, it reveals that co-regularize will lose efficacy under the insufficient scenario. While in CMML, the attention mechanism can calculate the instance level insufficiency, the robust consistency metric and diversity measure can process consistent or divergent modalities separately, thus it can make full use of multi-source data. Therefore, in the more realistic setting, combing the Eq. 1, Eq. 2 and Eq. 4, previous approaches can be rewritten dramatically:

$$\begin{aligned} \arg\min_{f^i} & \sum_{i=1}^{N_l}\sum_{j=1}^{M} \ell(\alpha_{i,j} f_j(\mathbf{x}_i), \mathbf{y}) \\ & + \|f_j\|_F^2 + Com(F) + \lambda R_\delta(F). \end{aligned} \quad (6)$$

it is notable that we utilize the deep network for different modalities in this paper.

4 Experiment

In this section, we validate the effectiveness of our proposed CMML approach. Specifically, most of current large-scale



(*a*) FLICKR25K (*b*) IAPR TC-12 (*c*) NUS-WIDE (*d*) MS-COCO

Figure 1: Ablation Study on 4 benchmark datasets. The x-axis represents the variant methods, the y-axis represents the different criteria, and the z-axis represents the results. (AP, Cov, Exauc, Maauc, Miauc and Rank represent Coverage, example AUC, Macro AUC, Micro AUC, Average Precision and Ranking Loss.)

multi-modal datasets are image-text multi-label classification for deep networks, thus we conduct the experiments on real-world multi-label task.

4.1 Datasets and Configurations

Without any loss of generality, we first experiment on 4 public real-world datasets, i.e., FLICKR25K [Huiskes and Lew, 2008], IAPR TC-12 [Escalante *et al.*, 2010], MS-COCO [Lin *et al.*, 2014] and NUS-WIDE [Chua *et al.*, 2009]. Besides, we also experiment on 1 real-world complex article dataset, i.e., WKG Game-Hub [Yang *et al.*, 2018a]:

- **FLICKR25K**: consists of 25,000 images collected from Flickr website, each image is associated with several textual tags. The text for each instance is represented by a 1386-dimensional bag-of-words vector. Each point is manually annotated with 24 labels. We select 23,600 pairs that belong to the 10 most frequent concepts;
- **IAPR TC-12**: consists of 20,000 image-text pairs which are annotated 255 labels. The text for each point is represented by a 2912-dimensional bag-of-words vector;
- **NUS-WIDE**: contains 260,648 web images, and images are associated with textual tags where each point is annotated with 81 concept labels. We select 195,834 image-text pairs that belong to the 21 most frequent concepts. The text for each point is represented by a 1000-dimensional bag-of-words vector;
- **MS-COCO**: contains 82,783 training, 40,504 validation image-text pairs which belong to 91 categories. We select 38,000 image-text pairs that belong to the 20 most frequent concepts. Text is represented by 2912-dimensional bag-of-words vector;
- **WKG Game-Hub**: consists of 13,750 articles collected from the Game-Hub of " Strike of Kings" with 1744 concept labels. We select 11,000 image-text pairs that belong to the 54 most frequent concepts. Each article contains several images and content paragraphs. The text for each point is represented by a 300-dimensional word2vector vector.

For each dataset, we randomly select 33% of the data for test set and the remaining instances are used for training. And for training data, we randomly choose 30% as the labeled data, and the left 70% as unlabeled ones. Image encoder is implemented with Resnet18 [He *et al.*, 2015], the text utilizes fully connected network. The parameter λ in the training phase is tuned in $\{0.1, 0.2, \cdots, 0.9\}$. When the variation between the objective values of Eq. 6 is less than 10^{-4} in iterations, we consider CMML converges. For all compared methods, the parameters are tuned as original papers. 6 common multi-label measurement criteria are recorded, i.e., Coverage, Ranking Loss, Average Precision, Macro AUC, example AUC and Micro AUC. We run the following experiments with the implementation of an environment on NVIDIA K80 GPUs server, and our model can be trained about 290 images per second with a single K80 GPGPU.

4.2 Compared Methods

To evaluate the performance of our proposed approach, firstly, we adopt ablation study to verify the effectiveness of the proposed CMML. Specifically, we define 5 different varieties of CMML (Att_f), i.e., Att_o, Max_f, Max_o, $Mean_f$, $Mean_o$, in which $Att/Max/Mean$ mean attention mechanism/max pooling/mean pooling, f represents the feature embedding layer and o represents label output layer. Thus, e.g., Att_f denotes that we utilize the attention mechanism on the feature embedding layer for different modalities, then with a fully connected matrix for final prediction and so on. Besides, for validate the effectiveness of the diversity measure and consistency measure, we construct another 2 comparison methods using only one of the measurements alone, i.e., Com, R_δ.

Moreover, we compare CMML with the state-of-the-art multi-modal/multi-label/multi-modal multi-label methods. For multi-modal comparing methods, we treat each label independently, i.e., for each label, a method trains classifiers using different modalities; for multi-label comparison methods, we concatenate multi-modal data as the unified input. In detail, multi-modal methods include: Co-trade [Zhang and Zhou, 2011], Co-regularize [Brefeld *et al.*, 2006], WNH [Wang *et al.*, 2013], ARM [Yang *et al.*, 2015], SLIM [Yang *et al.*, 2018b]; multi-label method includes: DeepMIML [Feng and Zhou, 2017]; multi-modal multi-label methods include: CS3G [Ye *et al.*, 2016], M3DN [Yang *et al.*, 2018a]:

- **Co-trade:** is a novel co-training algorithm by reliably communicating labeling information between different modalities;

Methods	Coverage ↓				Macro AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
CoReg	12.326±1.469	18.608±.695	8.320±.189	8.712±1.116	.490±.060	.503±.003	.502±.002	.460±.004
CoTra	N/A	18.240±.914	9.168±1.523	N/A	N/A	.499±.003	.496±.007	N/A
WNH	20.110±.400	18.804±.183	7.274±.172	6.282±.185	.692±.017	.554±.007	.696±.009	.753±.015
SLIM	13.988±.178	16.526±.080	6.188±.057	3.473±.504	.821±.001	.656±.002	.809±.001	.907±.000
CS3G	14.423±.278	13.508±.155	5.034±.086	5.273±.050	.663±.001	.589±.001	.684±.001	.685±.002
DeepMIML	13.815	11.281	3.808	3.712	.629	.678	.864	.844
ARM	17.228	12.307	6.652	6.600	.620	.722	.733	.630
M3DN	**3.947**	8.324	6.119	2.764	.892	.836	.838	.898
CMML	8.299	**8.204**	**3.428**	**2.494**	**.934**	**.853**	**.889**	**.936**

Methods	Ranking Loss ↓				Example AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
CoReg	.482±.081	.413±.018	.270±.009	.297±.053	.537±.087	.586±.017	.731±.010	.703±.053
CoTra	N/A	.381±.027	.340±.096	N/A	N/A	.619±.028	.670±.092	N/A
WNH	.224±.009	.402±.006	.237±.010	.187±.003	.762±.019	.597±.006	.763±.010	.812±.004
SLIM	.120±.001	.338±.002	.179±.002	.071±.001	.877±.001	.661±.003	.821±.003	.919±.001
CS3G	.139±.003	.250±.002	.152±.003	.159±.002	.860±.003	.749±.002	.847±.003	.840±.003
DeepMIML	.124	.212	.093	.086	.876	.788	.907	.914
ARM	.177	.245	.190	.183	.822	.754	.809	.816
M3DN	.108	.142	.112	.119	.899	.858	.898	.881
CMML	**.043**	**.135**	**.076**	**.045**	**.957**	**.864**	**.924**	**.955**

Methods	Average Precision ↑				Micro AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
CoReg	.335±.100	.235±.007	.590±.010	.414±.031	.552±.039	.582±.012	.687±.019	.713±.026
CoTra	N/A	.303±.015	.433±.161	N/A	N/A	.605±.018	.641±.069	N/A
WNH	.363±.065	.247±.002	.341±.005	.362±.007	.763±.019	.575±.006	.743±.007	.782±.007
SLIM	.578±.001	.382±.008	.669±.003	.806±.002	.881±.001	.645±.002	.824±.002	.912±.001
CS3G	.619±.009	.442±.003	.679±.005	.484±.002	.856±.003	.730±.001	.828±.002	.829±.001
DeepMIML	.654	.476	.771	.763	.864	.776	.913	.915
ARM	.496	.417	.660	.613	.804	.738	.805	.791
M3DN	.698	**.637**	.691	.634	.858	.863	.877	.878
CMML	**.837**	.614	**.814**	**.854**	**.956**	**.867**	**.922**	**.959**

Table 1: Comparison results of CMML. 6 common criteria are recorded. The best performance for each criterion is bolded. ↑ / ↓ indicate the larger/smaller, the better of the criterion.

- **Co-regularize:** minimizes the disagreement over different modal data with the unlabeled data;
- **WNH:** combines all modal values from different modalities together and then uses $l_{2,1}$-norm to regularize the modality selection process and finally gives the results;
- **ARM:** extracts the discriminative feature subspace of weak modality while regularizing the strong predictor;
- **SLIM:** exploits more extrinsic information from unlabeled data for classification and clustering;
- **DeepMIML**: exploits deep neural network to generate instance representation for MIML;
- **CS3G:** handles types of interactions between multiple labels and utilizes the data from different modalities;
- **M3DN:** models the deep independent network for each modality, and imposes the modal consistency on bag-level prediction by requiring that bag-based prediction of different modalities generate similar label correlation.

4.3 Ablation Study

We first explore which variant method will yield better results, and the results are listed in Fig. 1. Due to page limitation, we only list 4 public datasets here. The results reveal that the CMML, i.e., Att_f, achieves the best on most datasets for different performance measures, while it reveals that it achieves better performance to use both the diversity regularization and robust consistency metric than using them independently. This verifies the effectiveness of attention mechanism, diversity regularization and robust consistency metric for solving the insufficient problem.

4.4 Multi-label Classification

CMML is firstly compared with the state-of-the-art methods on 4 benchmark datasets, results are listed in Tab. 1. Results of deep learning methods only give the best results as [LeCun

Methods	Coverage ↓	Macro AUC ↑	Ranking Loss ↓	Example AUC ↑	Average Precision ↑	Micro AUC ↑
Image	40.049	.751	.148	.851	.560	.855
Text	59.016	.635	.194	.805	.479	.802
All	53.899	.717	.166	.834	.541	.837
CoReg	74.685±1.201	.488±.003	.290±.006	.711±.006	.337±.010	.705±.007
CoTra	74.548±.628	.488±.001	.298±.010	.703±.009	.323±.006	.695±.011
WHN	74.431±.242	.546±.007	.356±.004	.643±.004	.284±.004	.613±.006
SLIM	60.923±.369	0.809±.001	.205±.001	.794±.001	.518±.002	.797±.001
CS3G	60.958±.186	.581±.001	.208±.001	.791±.001	.453±.001	.793±.001
DeepMIML	36.318	.855	.100	.899	.637	.908
ARM	72.903	.587	.278	.722	.410	.712
M3DN	31.432	.732	.180	.828	.409	.880
CMML$_I$	30.865	.885	.082	.917	.718	.925
CMML$_T$	54.442	.674	.177	.822	.499	.823
CMML	**29.263**	**.893**	**.077**	**.922**	**.763**	**.931**

Table 2: Comparison results of CMML with compare methods on WKG Game-Hub dataset. 6 common criteria are recorded. The best performance is bolded. ↑ / ↓ indicate the larger/smaller, the better of the criterion.

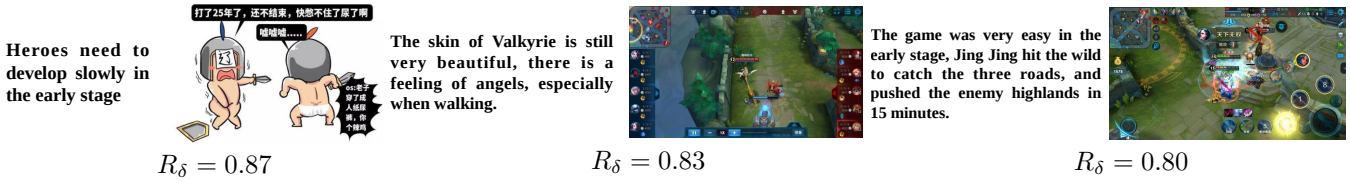


Figure 2: Examples of inconsistent instances using the robust consistency metric from WKG Game-Hub dataset.

et al., 2015], other comparison methods give the results with mean and std. The notation "N/A" means a method cannot give a result in 60 hours. From the results, it is obviously that our CMML can achieve the best performance in most datasets with different performance measures, except the Coverage on FLICKR25K, which reveals that the CMML approach is a high-competitive multi-modal learning method considering the modal insufficiency.

Besides, for the real-world complex article classification dataset, comparison results against compared methods are listed in Tab. 2. Similarly, it can be found that CMML approach also gets the best results on overall criteria, which validates the effectiveness of our method in solving complex article classification problem.

Considering limitation of the paper, we only give the comparison results of single modality on WKG-Hub dataset, which is with modal insufficiency in practice. In Tab. 2, the first partition is the results using only image or text information, and ensemble results. The last partition is the results of CCML method for image (CMML$_I$), text (CMML$_T$) predictions and ensemble (CMML) results. We can find that not only the results of single modalities are not degraded on various criteria, but also the overall results are also improved, thus achieved our goal, i.e., better single modal and overall performance. While other multi-modal methods (e.g., CoReg, CoTra, WNH, SLIM) are lower than single modal results on some criteria without considering the modal insufficiency.

4.5 Overcome Insufficiency

Insufficiency leads to divergences of different modalities, and there will exist inconsistent instances. However, the robust consistency measure can overcome the insufficiency. Thus, we conduct more experiments, and locate the inconsistent instances with the δ in maximum order. Fig. 2 exhibits several illustrative examples of the inconsistent instances. Qualitatively, we find that using robust consistency measure will achieve better performance from the classification results, in addition, it also demonstrates the phenomenon from the sorted examples.

5 Conclusion

In real-world applications, multi-modal data are often insufficient, which leads to the failure of previous multi-modal learning methods based on sufficiency assumption. In this paper, under the insufficient scenario, we developed a novel Comprehensive Multi-Modal Learning (CMML) framework. Specifically, we proposed the attention mechanism to learn the instance level sufficiency for each instance on different modalities. Also, novel diversity regularization and robust consistency metrics are designed for discovering insufficient modalities. Finally, experiments on real-world data obtain remarkable results for our method in terms of various criteria.

Acknowledgments

This work is partially supported by National Key R&D Program of China (2018YFB1004300), NSFC(61773198), NSF

IIS-1814510, and Collaborative Innovation Center of Novel Software Technology and Industrialization of NJU, Jiangsu. De-Chuan Zhan is the corresponding author.

References

[Balcan *et al.*, 2004] Maria-Florina Balcan, Avrim Blum, and Ke Yang. Co-training and expansion: Towards bridging theory and practice. In *NIPS*, pages 89–96, British Columbia, Canada, 2004.

[Blum and Mitchell, 1998] Avrim Blum and Tom M. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pages 92–100, Madison, Wisconsin, 1998.

[Brefeld *et al.*, 2006] Ulf Brefeld, Thomas Gartner, Tobias Scheffer, and Stefan Wrobel. Efficient co-regularised least squares regression. In *ICML*, pages 137–144, Pittsburgh, Pennsylvania, 2006.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In *CIVR*, Santorini Island, Greece, 2009.

[Dasgupta *et al.*, 2001] Sanjoy Dasgupta, Michael L. Littman, and David A. McAllester. PAC generalization bounds for co-training. In *NIPS*, pages 375–382, British Columbia, Canada, 2001.

[Escalante *et al.*, 2010] Hugo Jair Escalante, Carlos A. Hernandez, Jesus A. Gonzalez, Aurelio Lopez-Lopez, Manuel Montes-y-Gomez, Eduardo F. Morales, Luis Enrique Sucar, Luis Villasenor Pineda, and Michael Grubinger. The segmented and annotated IAPR TC-12 benchmark. *CVIU*, 114(4):419–428, 2010.

[Feng and Zhou, 2017] Ji Feng and Zhi-Hua Zhou. Deep MIML network. In *AAAI*, pages 1884–1890, San Francisco, California, 2017.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[Huber and others, 1964] Peter J Huber et al. Robust estimation of a location parameter. *AMS*, 35(1):73–101, 1964.

[Huiskes and Lew, 2008] Mark J. Huiskes and Michael S. Lew. The MIR flickr retrieval evaluation. In *SIGMM*, pages 39–43, British Columbia, Canada, 2008.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[Li *et al.*, 2012] Nan Li, Yang Yu, and Zhi-Hua Zhou. Diversity regularized ensemble pruning. In *ECML/PKDD*, pages 330–345, Bristol, UK, 2012.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, Zurich, Switzerland, 2014.

[Sridharan and Kakade, 2008] Karthik Sridharan and Sham M. Kakade. An information theoretic framework for multi-view learning. In *COLT*, pages 403–414, Helsinki, Finland, 2008.

[Tao *et al.*, 2018] Hong Tao, Chenping Hou, Xinwang Liu, Tongliang Liu, Dongyun Yi, and Jubo Zhu. Reliable multi-view clustering. In *AAAI*, pages 4123–4130, New Orleans, Louisiana, 2018.

[Wang and Zhou, 2013] Wei Wang and Zhi-Hua Zhou. Co-training with insufficient views. In *ACML*, pages 467–482, Canberra, Australia, 2013.

[Wang *et al.*, 2013] Hua Wang, Feiping Nie, and Heng Huang. Multi-View Clustering and Feature Learning via a Structured Sparsity. In *ICML*, pages 352–360, Atlanta, GA, 2013.

[Xu *et al.*, 2013] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *CoRR*, abs/1304.5634, 2013.

[Yang *et al.*, 2015] Yang Yang, Han-Jia Ye, De-Chuan Zhan, and Yuan Jiang. Auxiliary information regularized machine for multiple modality feature learning. In *IJCAI*, pages 1033–1039, Buenos Aires, Argentina, 2015.

[Yang *et al.*, 2018a] 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 *KDD*, pages 2594–2603, London, UK, 2018.

[Yang *et al.*, 2018b] Yang Yang, De-Chuan Zhan, Xiang-Rong Sheng, and Yuan Jiang. Semi-supervised multi-modal learning with incomplete modalities. In *IJCAI*, pages 2998–3004, Stockholm, Sweden, 2018.

[Ye *et al.*, 2016] 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 *ICDM*, pages 559–568, Barcelona, Spain, 2016.

[Yu *et al.*, 2011] Yang Yu, Yu-Feng Li, and Zhi-Hua Zhou. Diversity regularized machine. In *IJCAI*, pages 1603–1608, Catalonia, Spain, 2011.

[Zhang and Zhou, 2011] Min-Ling Zhang and Zhi-Hua Zhou. Cotrade: Confident co-training with data editing. *TSMC*, 41(6):1612–1626, 2011.