Complex Object Classification: A Multi-Modal Multi-Instance Multi-Label Deep Network with Optimal Transport*

Yang Yang
National Key Laboratory for Novel Software Technology, Nanjing University
Nanjing
yangy@lamda.nju.edu.cn

Yi-Feng Wu
National Key Laboratory for Novel Software Technology, Nanjing University
Nanjing
wuyf@lamda.nju.edu.cn

De-Chuan Zhan*
National Key Laboratory for Novel Software Technology, Nanjing University
Nanjing
zhandc@nju.edu.cn

Zhi-Bin Liu
Tencent WXG
ShenZhen
lewiszbliu@tencent.com

Yuan Jiang
National Key Laboratory for Novel Software Technology, Nanjing University
Nanjing
jiangy@lamda.nju.edu.cn

ABSTRACT

In real world applications, complex objects are usually with multiple labels, and can be represented as multiple modal representations, e.g., the complex articles contain text and image information as well as are with multiple annotations. Previous methods assume that the homogeneous multi-modal data are consistent, while in real applications, the raw data are disordered, i.e., the article is constituted with variable number of inconsistent text and image instances. To solve this problem, Multi-modal Multi-instance Multi-label (M3) learning provides a framework for handling such task and has exhibited excellent performance. Besides, how to effectively utilize label correlation is also a challenging issue. In this paper, we propose a novel Multi-modal Multi-instance Multi-label Deep Network (M3DN), which learns the label prediction and exploits label correlation simultaneously based on the Optimal Transport, by considering the consistency principle between different modal bag-level prediction and the learned latent ground label metric. Experiments on benchmark datasets and real world WKG Game-Hub dataset validate the effectiveness of the proposed method.

KEYWORDS

Multi-modal, Multi-instance, Multi-label, Optimal Transport

ACM Reference Format:
Yang Yang, Yi-Feng Wu, De-Chuan Zhan*, Zhi-Bin Liu, and Yuan Jiang. 2018. Complex Object Classification: A Multi-Modal Multi-Instance Multi-Label Deep Network with Optimal Transport. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3220012

1 INTRODUCTION

With the development of data collection techniques, objects always can be represented with multiple modal features, e.g., in the forum of the famous mobile game " Strike of Kings", the articles are with image and content information, while the articles belong to multiple categories if they are observed from from different aspects, e.g., an article can belong to "Wukong Sun" (Game Heroes) as well as "golden cudgel" (Game Equipment) from the images, while can be categorized as "game strategy", "producer name" from contents and so on. The major challenge for addressing such problem is how to jointly model the multiple types of heterogeneities in a mutually beneficial way. To address this problem, multi-modal multi-label learning approaches utilize multiple modal information, among which require that modal-based classifiers generate similar predictions, Huang et al. proposed a multi-label conditional restricted Boltzmann machine (ML-CRBM), which uses multiple modalities to obtain shared representations under the supervision [18]; Yang et al. proposed a novel graph-based model for learning with both label and feature heterogeneities [32]. However, a real-world object may contain with variable number of inconsistent multi-modal instances, e.g., the article usually contains multiple images and content paragraphs, in which each image or content paragraph can be regarded as an instance, yet the relationships between the images and contents have not marked.

To solve this problem, several Multi-modal Multi-instance Multi-label methods are proposed. Nguyen et al. proposed M3LDA consisting of a visual-label part, a textual-label part,

[*] * is the corresponding author

KDD '18, August 19–23, 2018, London, United Kingdom
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5552-0/18/08. . . $15.00
https://doi.org/10.1145/3219819.3220012

and a label topic part, in which the topic decided by the visual information and the topic decided by the textual information should be consistent, leading to the label assignment [25]; Nguyen et al. developed a multi-modal MIML framework based on hierarchical Bayesian Network, and derived an effective learning algorithm based on variational inference [24]. Nevertheless, previous approaches rarely consider the correlation between labels, Yang and He learned a hierarchical multi-latent space, which can simultaneously leverage the task relatedness, modal consistency and the label correlation to improve the learning performance [30]; Huang and Zhou proposed the ML-LOC approach which allows label correlation to be exploited locally, where global discrimination fitting and local correlation sensitivity are incorporated into a unified framework [17]; Frogner et al. developed a loss function with ground metric for multi-label learning, based on the wasserstein distance, which provides a natural notion of dissimilarity for probability measures [12]. The label similarity matrix or the ground metric acts as an important role in measuring the label correlation. Previous works mainly assumed that there exists some prior knowledge as cost matrix [12, 28]. Yet since there may be no a direct or simple semantic information among labels, leaving the confidence of the label similarity matrix or ground metric without considering.

In this work, aiming at simultaneously learning the label prediction and exploring label correlation, we proposed a novel Multi-modal Multi-instance Multi-label Deep Network (M3DN), which models the independent deep network for each modality, and imposes the modal consistency on bag-level, by requiring the bag-based prediction of different modalities should generate similar label correlation on the same bag. Specifically, we input the heterogeneous bag of multi-modal instances separately, and make label prediction with the novel bag-concept for different modalities. Besides, based on Optimal Transport (OT) theory [29], M3DN adopts optimal transport distance to measure the quality of prediction, which provides a more meaningful measure in multi-label tasks by capturing the geometric information of the underlying label space. On the other hand, it is notable that the raw ground metric is not confidently calculated by the raw data, and we cast the label correlation exploration as a latent ground metric learning problem. Consequently, M3DN could automatically learn the predictors for different modalities and the latent shared ground metric.

The main contributions of this paper are summarized in following three points:

- A novel Multi-modal Multi-instance Multi-label Deep Network (M3DN), which 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;
- Considering label correlation exploration as a latent ground metric learning problem between different modalities, rather than calculating with the prior raw knowledge with less confidence;
- A superior performance on real-world applications, comprehensively evaluation on the performance and obtaining consistently superior performances stably.

Section 2 is related work, our approach is presented in Section 3. Section 4 reports our experiments. Finally, Section 5 gives the conclusion.

2 RELATED WORK

The exploitation of multi-modal multi-instance multi-label learning has attracted much attention recently. In this paper, our method concentrates on deep multi-label classification with inconsistent multi-modal multi-instance data, while considering the label correlation using optimal transport technique. Therefore, our work is related to multi-modal multi-instance multi-label learning and the optimal transport.

Multi-modal learning deals with data from multiple modalities, i.e., multiple feature sets. The goal is to improve performance or reduce the sample complexity. Meanwhile, multi-modal multi-label learning has been well studied, e.g., Fang and Zhang proposed a multi-modal multi-label learning method based on the large margin framework, which maps the multi-modal data into low-dimensional feature space and simultaneously maximizes the dependency between features and labels [10]. Yang et al. modeled both the modal consistency and the label correlation in a graph-based framework [31]. The basic assumption behind of these methods is that the multi-modal data are consistent on the ground truth, however, in real application, the multi-modal data are always heterogeneous on the instance-level, e.g., the articles have variable number of inconsistent images and text paragraphs, the videos have variable length of inconsistent audio and image frames, these instances only have consistency on the bag level, rather than instance level. Thus, multi-modal multi-instance multi-label learning is proposed recently, Nguyen et al. developed a multi-modal MIML framework based on hierarchical Bayesian Network, and derived an effective learning algorithm based on variational inference [24]; Feng and Zhou exploited deep neural network to generate instance representation for MIML and can be extended to multi-modal scenario. Nevertheless, previous approaches rarely consider the label correlation [11].

Considering the label correlation, several multi-label learning methods are proposed [3, 36, 39]. Recently, Optimal Transport (OT) [29] is developed to measure the difference between two distributions based on given ground metric. The distance defined by OT is the Wasserstein distance or Earth Mover distance, and has been widely used in computer vision and image processing fields, e.g., Qian et al. proposed a novel method that exploits knowledge in both data manifold and feature correlation, which adopts an approximation of Earth Mover Distance (EMD) as metric [26]. Ye et al. developed a modified Bregman ADMM approach for computing the approximate discrete Wasserstein barycenter of large clusters [34]. Courty et al. proposed a regularized unsupervised optimal transportation model to perform the alignment of the representations in the source and target domains [6].

Previous work mainly assumed that there exists some prior knowledge for cost matrix. However, it is not general in real application considering the information or domain knowledge deficient.

Therefore, to solve these problems, we proposed a novel Multi-modal Multi-instance Multi-label Deep Network (M3DN), by learning the label prediction and exploring label correlation simultaneously, we can improve the label prediction performance as a result. Specifically, M3DN inputs bag of instances to different modal deep networks, based on Optimal Transport (OT) theory, bag-level label predictions are required consistent with the learned label metric. On the other hand, considering the uncertainty of the prior knowledge, we cast the label correlation exploration as a latent ground metric learning problem. Consequently, M3DN could automatically learn the predictors for different modalities and the latent shared ground metric.

3 PROPOSED METHOD

3.1 Notation

Considering the multi-instance extension of the multi-modal multi-label framework. Suppose we are given N bags of instances, let $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}$ denotes the label set, $\mathbf{y}_i \in \mathbb{R}^L$ is the label vector of $i-$th bag, where $y_{i,j} = 1$ denotes positive class, and $y_{i,j} = 0$ otherwise. On the other hand, suppose we are given K modalities, without any loss of generality, we consider two modalities in our paper, i.e., images and contents. Let $\mathcal{D} = \{([\mathbf{X}_1^1, \mathbf{X}_1^2], \mathbf{y}_1), ([\mathbf{X}_2^1, \mathbf{X}_2^2], \mathbf{y}_2), \cdots, ([\mathbf{X}_N^1, \mathbf{X}_N^2], \mathbf{y}_N)\}$ represents the training dataset, where $\mathbf{X}_i^1 = \{\mathbf{x}_{i,1}^1, \mathbf{x}_{i,2}^1, \cdots, \mathbf{x}_{i,m_i}^1\}$ denotes the bag representation of m_i instances of $\mathbf{X}_i^1$, similarly, $\mathbf{X}_i^2 = \{\mathbf{x}_{i,1}^2, \mathbf{x}_{i,2}^2, \cdots, \mathbf{x}_{i,n_i}^2\}$ is the bag representation of n_i instances of $\mathbf{X}_i^2$, it is notable that different bags of different modalities may contain variable number of instances. The goal is to generate a learner to annotate new bags based on its inputs $\mathbf{X}^1, \mathbf{X}^2$, e.g., annotates a new complex article with its images and contents.

3.2 Optimal Transport

Traditionally, several measurements as Kullback-Leibler divergences, Hellinger, total variation can be utilized to measure the similarity between two distributions. However, these measurements are with little effect when the probability space has the geometrical structures. On the other hand, Optimal transport [29], also known as Wasserstein distance or earth mover distance [35], defines a reasonable distance between two probability distribution over the metric space. Intuitively, the Wasserstein distance is the minimum cost of transporting the pile of one distribution into the pile of another distribution. Therefore, the Wasserstein distance is more powerful in such situations by considering the pairwise cost.

Definition 1. *(Transport Polytope) For two probability vectors r and c in the simplex $\sum_L$, $U(r,c)$ is the transport polytope of r and c, namely the polyhedral set of $L \times L$ matrices,*

$$U(r,c) = \{P \in \mathbb{R}_+^{L\times L} | P\mathbf{1_L} = \mathbf{r}, \mathbf{P}^\top \mathbf{1_L} = \mathbf{c}\}$$

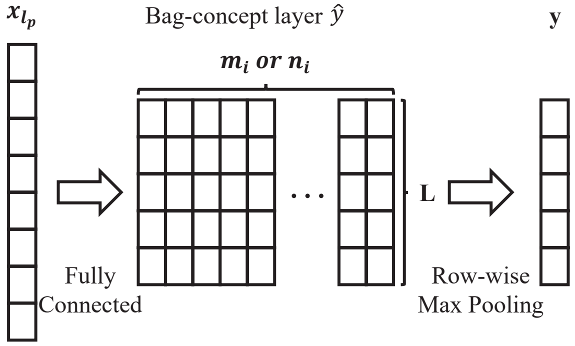


Figure 2: The schematic of the bag-concept layer, with the output feature representations of a bag of instances, we can acquire the bag-concept layer, in which each column represents corresponding prediction of each instance. Eventually, the final label prediction is calculated by row-wise max pooling.

Definition 2. *(Optimal Transport) Given a $L \times L$ cost matrix M, the total cost of mapping from r to c using a transport matrix (or coupling probability) P can be quantified as $\langle P, M\rangle$. The optimal transport (OT) problem is defined as,*

$$d_M(r,c) = \min_{P\in U(r,c)} \langle P, M\rangle$$

Theorem 1. *d_M defined in Def. 2 is a distance on $\sum_L$ whenever M is a metric matrix [29].*

3.3 Multi-Modal Multi-instance Multi-label Deep Network (M3DN)

Parallel Deep Network

In this section, we propose the Multi-Modal Multi-instance Multi-label Deep Network (M3DN) framework, which aims to learn the label prediction and explore label correlation simultaneously. M3DN models deep networks for different modalities and imposes the modal consistency by requiring that bag-based output label predictions of different modalities generate similar latent label correlations on the homogeneous bag representation.

Specifically, the raw articles can be divided into two modal bags of heterogeneous instances, i.e., the image bag with 4 images and content bag with 5 text paragraphs as shown in Fig. 1, while only the homogeneous bags share the same multiple labels. Each instance $\mathbf{x}^1(\mathbf{x}^2)$ in different modal bag can be calculated among several layers and can be finally represented as $\mathbf{x}_{l_{p1}}(\mathbf{x}_{l_{p2}})$, without any loss of generality, we use the convolutional neural network for images and the fully connected networks for text. Then, the output features are fully connected with the bag-concept layer, all parameters including deep network facts and fully connected weights can be organized as $\Theta_1 = \{\theta_{l_1}, \theta_{l_2}, \cdots, \theta_{l_{p1-1}}, W_1\}(\Theta_2 = \{\theta_{l_1}, \theta_{l_2}, \cdots, \theta_{l_{p2-1}}, W_2\})$. Concretely, once the label predictions of the instances for a bag $\mathbf{X}_i^v$ are obtained, we propose a

]

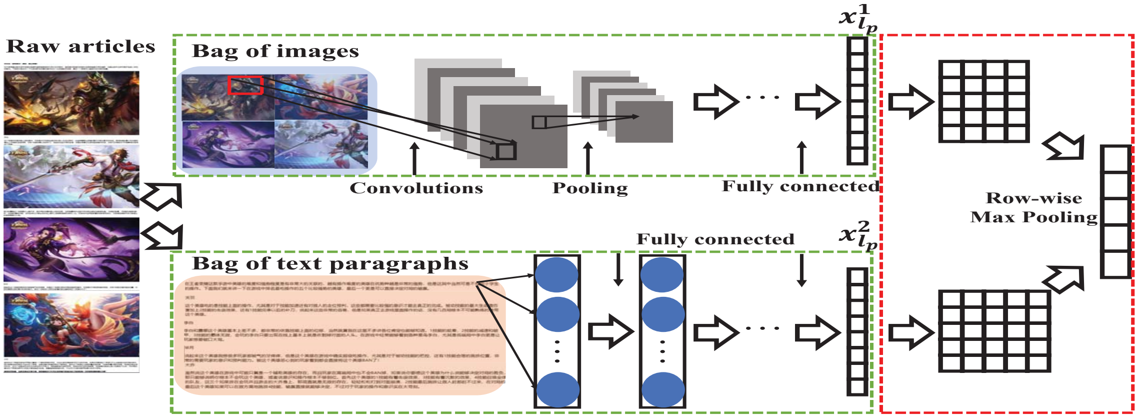


Figure 1: The flowchart of the M3DN, the raw articles can be divided into two homogeneous modal bag with variable number of heterogeneous instances, i.e., the image bag with four images and content bag with 5 text paragraphs. The instances of different modalities can be calculated with different deep networks, and finally represented as $\mathbf{x}^1_{l_p}$ or $\mathbf{x}^2_{l_p}$, the output features are fully connected with the labels, and we can get the bag-concept layer for different modalities. Eventually, we can acquire the final prediction by mean-max pooling the bag-concept layer of different modalities.

fully connected 2D layer (bag-concept layer) of size $m_i(n_i) \times L$ as shown in Fig. 2, in which each column represents corresponding prediction of each instance in the image/content bag. Formally, for a given bag of instances $\mathbf{X}_i^v$, the (k, j)-th node in the 2D bag-concept layer represents the prediction score between the instance $\mathbf{x}_{i,j}^v$ and the $k-$th label. That is, the j-column has the following form of activation:

$$\hat{\mathbf{y}}_j^v = g(W_v \mathbf{x}_{i,j}^v + b_v) \tag{1}$$

Here, $g(\cdot)$ can be any convex activation function, and we use softmax function here. As for the bag-concept layer, we utilize the row-wise max pooling: $f_v(i) = max(\hat{\mathbf{y}}_{i,\cdot})$. Finally, the final prediction value is:$f = \frac{f_1+f_2}{2}$.

The Formulation

However, fully connection to the label output rarely considers the relationship between labels. Recently, Optimal Transport (OT) theory [29] is used in multi-label learning, which captures the geometric information of the underlying label space. Without any loss of generality, with the Def. 2 and Def. 1, the loss function implied in the parallel network structure can be formulated as:

$$\min_{P_v \in U(f(X_i^v),\mathbf{y}_i)} \sum_{v=1}^{2}\sum_{i=1}^{N} \langle P_v, M\rangle$$
$$s.t. \quad U(f(\mathbf{X}_i^v), \mathbf{y}_i) = \{P_v \in \mathbb{R}_+^{L\times L} | P_v \mathbf{1}_L = f(X_i^v), P_v^\top \mathbf{1}_L = \mathbf{y}_i\} \tag{2}$$

Where M is the shared latent cost matrix. However, this method assumes that there exists the prior knowledge for constructing the cost matrix M. Yet since there may be no direct or informative information among labels in real application, which leads to the weak cost matrix M and poor classification performance.

Therefore, we can define learning the cost metric as an optimization problem. Optimizing the cost metric directly is difficult and consuming, thus, [8, 40] proposed to formulate the cost metric learning problem with the side information, i.e., the label similarity matrix S as in the [40], and [8] has proved that learning the cost metric matrix M, which computes corresponding optimal transport distance d_M between pairs of labels, agrees with the side information. More precisely, this criterion favors matrix M, for which the distance $d_M(r; c)$ is small for pairs of similar histograms r and c (corresponding $S(r; c)$ is large) and large for pairs of dissimilar histograms (corresponding $S(r; c)$ is small). Consequently, optimizing M can be turned to optimize the S. Finally, the goal of M3DN can be tuned to learn label predictor and explore label correlation simultaneously.

In detail, we first introduce the connection between the nonlinear transformation and pseudo-metric:

Definition 3. *With the nonlinear transformation $\emptyset(\cdot)$, the Euclidean distance after the transformation can be denoted as:*

$$D_\emptyset(r, c) = \|\emptyset(r) - \emptyset(c)\|_2.$$

And [21] proved that $D_\emptyset$ satisfies all properties of a well-defined pseudo-metric in the original input space.

Theorem 2. *For a pseudo-metric M defined in Def. 3 and histograms $r, c \in \sum_L$, the function $(r, c) \to \mathbf{1}_{r\neq c} d_M(r, c)$ satisfies all four distance axioms, i.e., non-negativity, symmetry, definiteness and sub-additivity (triangle inequality) as in [8].*

Thus, M can be tuned to learn the kernel S defined by the non-linear transformation $\emptyset(\cdot)$:

$$S_{ij} = S(\mathbf{y}_i, \mathbf{y}_j) = \emptyset(\mathbf{y}_i)^\top \emptyset(\mathbf{y}_j) \tag{3}$$

where the $\mathbf{y}_i$ represents the label vector of $i-$th instance. Besides, it is notable that the cost matrix M is computed

Algorithm 1 The pseudo code of learning the predictors

```
Input:
    • Sampled Batch Dataset: ${[X_i^1, X_i^2], \mathbf{y}}_{i=1}^n$, kernelized similar matric $S^t$, current mapping $f_1, f_2$
    • Parameter: $\lambda$
Output:
    • Gradient of the target mapping: $\partial L/\partial f_1, \partial L/\partial f_2$
Calculate $M \leftarrow$ Eq. 4
Initialize $K = exp(-\lambda M - 1)$, $\nabla \leftarrow \mathbf{0}$
for $v = 1, 2$ do
    for $i = 1, 2, \cdots, n$ do
        $u_i^v \leftarrow \mathbf{1}$
        while $u_i^v$ not converged do
            $u_i^v \leftarrow f_v(\mathbf{x}_i^v) \oslash (K(\mathbf{y}_i^v \oslash K^\top u_i^v))$
        end while
        $\nabla^{f_v} \leftarrow \nabla^{f_v} + \frac{logu_i^v}{\lambda} - \frac{logu_i^{v\top}\mathbf{1}}{\lambda L} \cdot \mathbf{1}$
    end for
end for
```

as $M_{ij} = D_\emptyset^2(\mathbf{y}_i, \mathbf{y}_j)$, while the kernel S is defined as Eq. 3. Thus, the relation between M and S can be derived as:

$$M_{ij} = S_{ii} + S_{jj} - 2S_{ij}. \tag{4}$$

The non-linear mapping preserves pseudo metric properties in Def. 3, therefore it only needs a projection to positive semi-definite matrix cone when learning the kernel matrix S. Thus, we can avoid the projection to metric space which is very complicated and costly. Therefore, we propose to conduct the label predictions and label correlation exploration simultaneously based on substituted optimal transport, combining Eq. 4, Eq. 2 can be reformulated as:

$$\begin{aligned} &\min_{S, P_v \in U(f(X_i^v), \mathbf{y}_i)} \sum_{v=1}^{2}\sum_{i=1}^{N}\langle P_v, M\rangle + \lambda_1 r(S, S_0) \\ s.t.\quad & U(f(\mathbf{X}_i^v), \mathbf{y}_i) = \{P_v \in \mathbb{R}_+^{L\times L} | P_v\mathbf{1}_L = f(X_i^v), P_v^\top\mathbf{1}_L = \mathbf{y}_i\} \\ & S \in \mathcal{S}_+, \quad M_{ij} = S_{ii} + S_{jj} - 2S_{ij} \end{aligned} \tag{5}$$

where λ_1 is a trade-off parameter, $\mathcal{S}_+$ denotes the set of positive semi-definite matrix. $\lambda_1 r(S, S_0)$ can be any convex regularization. This regularizer is $\mathcal{S}_+ \times \mathcal{S}_+ \to \mathcal{R}_+$, allows us to exploit prior knowledge on the kernelized similar matrix, encoded by a reference matrix S_0, since typically no strong prior knowledge is available, we use $S_0 = \mathcal{Y}' \times \mathcal{Y}$. Following common practice [15], we make use of the asymmetric Burg divergence, which yields:

$$r(S, S_0) = \mathrm{tr}(SS_0^{-1}) - logdet(SS_0^{-1}) - p$$

Optimization

The 1st term in Eq. 5 involves the product of the predictors f and the cost matrix S, which makes the formulation not joint convex. Consequently, the formulation cannot be optimized easily. We provide the optimization process below:

Fix S**, Optimize** f_1, f_2**:** When updating f_1, f_2 with a fixed S, the 2nd term of Eq. 5 is not relevant to f_1, f_2, and the Eq. 5 can be reformulated as follows:

$$\begin{aligned} &\min_{P_v \in U(f(X_i^v), \mathbf{y}_i)} \sum_{v=1}^{2}\sum_{i=1}^{N}\langle P_v, M\rangle \\ s.t.\quad & U(f(\mathbf{X}_i^v), \mathbf{y}_i) = \{P_v \in \mathbb{R}_+^{L\times L} | P_v\mathbf{1}_L = f(X_i^v), P_v^\top\mathbf{1}_L = \mathbf{y}_i\} \end{aligned} \tag{6}$$

The empirical risk minimization function of Eq. 6 can be optimized by stochastic gradient descent. However, it requires to evaluate the descent direction for the loss, with respect to the predictor f, while computing the exact subgradient is quite costly, especially when it is in the constraints.

To solve this problem, similar to [12], the loss is a linear program, and the subgradient can be computed using Lagrange duality. Therefore, we use primal-dual approach to compute the gradient by solving the dual LP problem, from [1], we know that the dual optimal α is, in fact, the subgradient of the loss of training sample $(\mathbf{X}^v, \mathbf{y})$ with respect to its first argument f_v. However, it is costly to compute the exact loss directly. In the [7], Sinkhorn relaxation is adopted as the entropy regularization to smooth the transport objective, which results in a strictly convex problem that can be solved through Sinkhorn matrix scaling algorithm, at a speed that is several orders of magnitude faster than that of transport solvers [7].

For a given training bag of instances $([\mathbf{X}^1, \mathbf{X}^2], \mathbf{y})$, the dual LP of Eq. 6 is:

$$d_M(f_v(\mathbf{X}^v), y) = \max_{\alpha, \beta \in C_M} \alpha^\top f(\mathbf{X}_i^v) + \beta\mathbf{y}, \tag{7}$$

where $C_M = \{\alpha, \beta \in \mathbb{R}^L : \alpha_i + \beta_j < M_{i,j}\}$.

Definition 4. *(Sinkhorn Distance) Given a* $L \times L$ *cost matrix* M*, and histograms* $(r, c) \in \sum_L$*. The Sinkhorn distance is defined as:*

$$\begin{aligned} d_M^\lambda(r, c) &= \min_{P^\lambda \in U(r,c)} \langle P^\lambda, M\rangle \\ P^\lambda &= \arg\min_{P \in U(f(X_i^v), \mathbf{y}_i)} \langle P, M\rangle - \frac{1}{\lambda}H(P) \end{aligned} \tag{8}$$

where $H(P) = -\sum_{i=1}^L\sum_{j=1}^L p_{ij}logp_{ij}$ is the entropy of P, and $\lambda > 0$ is entropic regularization coefficient.

Based on the Sinkhorn theorem, we could conclude that the transportation matrix can be written in the form of $P^\star = diag(u)Kdiag(v)$, where $K = exp(-\lambda M - 1)$ is the element-wise exponential of $\lambda M - 1$. Besides, $u = exp(\lambda\alpha)$ and $v = exp(\lambda\beta)$.

Therefore, we adopt the well-known Sinkhorn-Knopp algorithm, which is used in [7, 8] to update the target mapping f_v given the ground metric, in which f_v can be defined as Eq. 1. The detailed procedure is summarized in Algorithm 1, then with the help of Back Propagation technique, gradient descent could be adopted to update the network parameters.

Fix f_1, f_2**, Optimize** S**:** When updating S with the fixed f_1, f_2, the sub-problem can be rewritten as following:

$$\begin{aligned} &\min_S \sum_{v=1}^{2}\sum_{i=1}^{N}\langle P, M\rangle + \lambda_1 r(S, S_0) \\ s.t.\quad & K \in \mathcal{S}_+, \quad M_{ij} = S_{ii} + S_{jj} - 2S_{ij}. \end{aligned} \tag{9}$$

Algorithm 2 The pseudo code of M3DN

```
Input:
  • Dataset: $\mathcal{D} = \{[X_i^1, X_i^2], \mathbf{y}\}_{i=1}^N$
  • Parameter: $\lambda_1$, $\lambda$
  • maxIter: $T$, learning rate: $\{\alpha_t\}_{t=1}^T$
Output:
  • Classifiers: $f_1, f_2$
  • Label similar matric: $S, M$
Initialize $S_0 \leftarrow \mathcal{Y}' \times \mathcal{Y}$
while true do
    Create Batch: Randomly pick up $n$ examples from $\mathcal{D}$ without replacement
    Calculate $S^{t+1} \leftarrow$ Eq. 10, Eq. 11
    Calculate $\partial L/\partial f_1^t, \partial L/\partial f_2^t \leftarrow$ Alg. 1
    Weight Propagation step: Obtain the derivative $\partial f_1^t/\partial\Theta_1$, $\partial f_2^t/\partial\Theta_2$;
    Update parameters $\Theta_1, \Theta_2$
    $Func_{obj}^{t+1} \leftarrow$ calculate obj. value in Eq. 5 with $F^{t+1}$
    if $\|Func_{obj}^{t+1} - Func_{obj}^t\| \leq \epsilon$ or $t \geq T$ then
        Break;
    end if
end while
```

This sub-problem has closed-form solution. The differential can be formulated as:

$$S = (\hat{P} + S_0^{-1} - p)^{-1} \tag{10}$$

where

$$\hat{P} = \begin{cases} -2P_{ij}, \ when \quad i \neq j, \\ \sum_{k \neq i}^{L} (P_{ik} + P_{ki}), when \quad i = j \end{cases}$$

Then, we project S back to positive semi-definite cone as:

$$S = \mathbf{Proj}(S) = Umax(\sigma, 0)U^\top \tag{11}$$

where **Proj** is a projection operator, U and σ correspond to the eigenvectors and eigenvalues of S. The whole procedure is summarized in Algorithm 2.

4 EXPERIMENTS

In this section, we validate the effectiveness of our proposed M3DN approach. We first compare M3DN on real-world multi-label datasets as benchmarks, then present the assessment of complex article classification collected from the WKG Game-Hub of Tencent.

4.1 Datasets and Configurations

M3DN can learn more discriminative multi-modal feature representation on bag level for multi-label classification, while considering the label correlation between different labels. Thus, in this section, we will provide the empirical investigations and performance comparisons of M3DN on multi-label classification and label correlation. Without any loss of generality, we experiment on 4 public real-world datasets, i.e., FLICKR25K [19], IAPR TC-12 [9], MS-COCO [23] and NUS-WIDE [5]. Besides, we also experiment on 1 real-world complex article dataset, i.e., WKG Game-Hub:

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 as a 1386-dimensional bag-of-words vector. Each point is manually annotated with 24 labels.

IAPR TC-12: consists of 20,000 image-text pairs which are annotated 255 labels. The text for each point is represented as 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 as a 1000-dimensional bag-of-words vector.

MS-COCO: contains 82,783 training, 40,504 validation image-text pairs which belong to 91 categories. The text for each point is represented as a 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. Each article contains several images and content paragraphs, the text for each point is represented as a 300-dimensional word2vector vector.

For each dataset, we randomly select 33% data for the test set and the remaining instances are used for training. For the 4 benchmark datasets, each image is divided into 10 regions using [13] as image bag, while the corresponding text tags also separate into several independent tags as text bag. For the WKG Game-Hub dataset, each article is denoted as an image bag and a content bag. The deep network for image encoder is implemented the same as Resnet-50 [14]. 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. The parameter λ_1 in the training phase is tuned in $\{0.1, 0.2, \cdots, 0.9\}$. When the variation between the objective value of Eq. 5 is less than 10^{-4} in iterations, we consider M3DN converges.

4.2 Compared methods

In our experiments, first, multi-modal multi-instance multi-label methods are compared, i.e., M3LDA [25], MIMLmix [24], besides, considering that M3DN can be degenerated into different settings, therefore, multi-modal multi-label methods, i.e., CS3G [33]; multi-instance multi-label methods, i.e., DeepMIML [11], M3MIML [38], MIMLfast [16]; and multi-label methods, i.e., SLEEC [2], Tram [22], ECC [27], ML-KNN [37], RankSVM [20], ML-SVM [4] are also compared. Specifically, for the multi-modal multi-label methods, all instances in each bag are additive average as the bag-level feature representation. For the multi-instance multi-label methods, all modalities of a dataset are concatenated together as a single

Table 1: Comparison results (mean ± std.) of M3DN with both compared methods on 4 benchmark datasets. 6 commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicate the larger/smaller the better of a criterion.

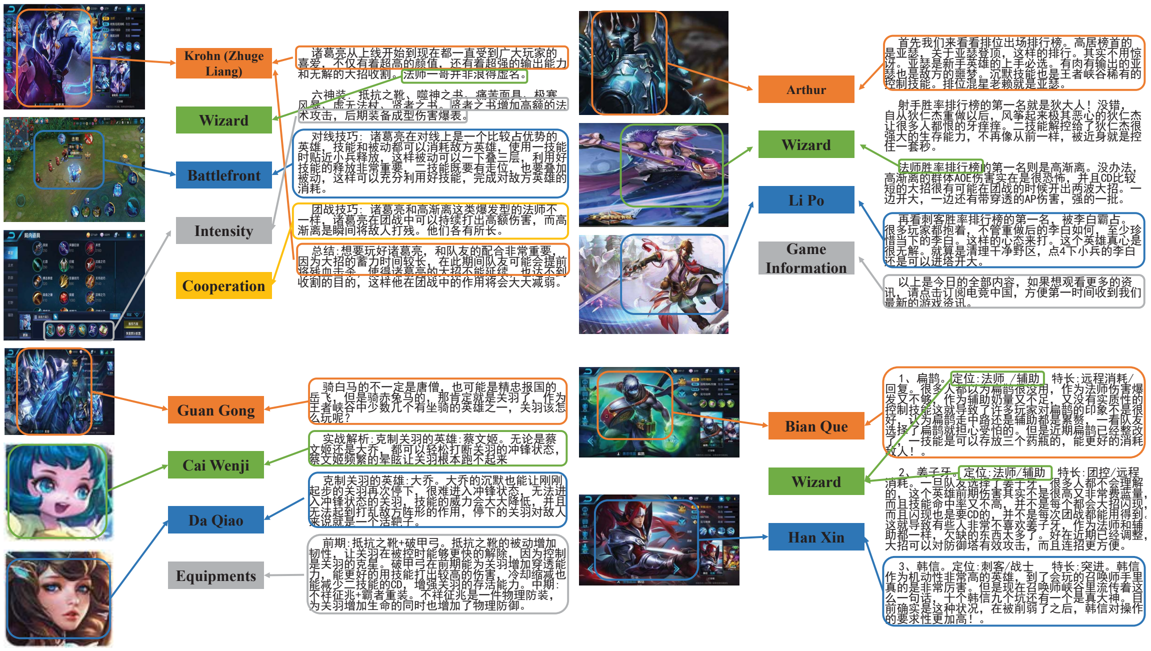
Methods	Coverage ↓				Macro AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
M3LDA	12.345±.214	11.620±.042	47.400±.622	6.670±.205	.532±.015	.526±.003	.507±.015	.509±.012
MIMLmix	17.114±1.024	15.720±.543	64.130±1.121	14.167±1.140	.472±.018	.554±.096	.471±.019	.493±.020
CS3G	8.168±.137	7.153±.178	50.138±2.146	8.028±.907	**.837±.007**	**.817±.006**	.717±.011	.530±.022
DeepMIML	9.242±.331	8.931±.421	27.358±.654	8.369±.119	.766±.035	.795±.022	**.827±0.006**	.823±.005
M3MIML	11.760±1.121	9.125±.553	42.420±.2.696	5.210±.920	.687±.087	.724±.033	.650±.032	.649±.084
MIMLfast	12.155±.913	12.711±.315	41.048±.831	8.634±.028	.524±.050	.485±.009	.506±.010	.522±.008
SLEEC	9.568±.222	9.494±.105	47.502±.448	7.390±.275	.706±.007	.675±.007	.661±.014	.620±.006
Tram	7.959±.187	8.156±.163	28.417±.945	9.934±.026	.780±.009	.746±.007	.776±.011	.493±.007
ECC	14.818±.086	14.229±.258	47.124±.675	7.941±.194	.532±.013	.484±.009	.630±.023	.634±.009
ML-KNN	10.379±.115	9.523±.072	27.568±.066	4.610±.062	.591±.008	.723±.006	.823±.003	.736±.008
RankSVM	11.439±.196	11.941±.078	37.300±.835	8.292±.054	.512±.019	.499±.009	.521±.033	.501±.001
ML-SVM	11.311±.158	11.755±.270	39.258±.294	7.890±.020	.503±.010	.502±.010	.497±.016	.561±.001
M3DN	**7.502±.129**	**6.936±.065**	**26.921±.320**	**4.599±.050**	.822 ±.009	.798±.002	.811±.004	**.826±.006**

Methods	Ranking Loss ↓				Example AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
M3LDA	.301±.009	.377±.002	.247±.001	.257±.006	.707±.008	.630±.005	.770±.006	.652±.009
MIMLmix	.609±.036	.675±.012	.609±.040	.583±.081	.391±.036	.325±.012	.391±.040	.417±.082
CS3G	.118±.005	.155±.005	.202±.009	.170±.032	.881±.005	.835±.005	.798±.009	.642±.032
DeepMIML	.149±.012	.166±.017	.089±.002	.164±.007	.791±.044	.834±.017	.911±.002	.835±.007
M3MIML	.271±.053	.250±.011	.191±.016	.284±.030	.729±.053	.751±.011	.811±.017	.717±.031
MIMLfast	.275±.033	.435±.021	.194±.006	.430±.009	.724±.033	.626±.013	.811±.005	.646±.009
SLEEC	.316±.009	.413.006	.455±.005	.512±.008	.843±.003	.761±.005	.796±.002	.713±.008
Tram	.132±.004	.203±.007	.117±.004	.456±.004	.867±.004	.797±.007	.883±.005	.591±.001
ECC	.804±.024	.928±.013	.461±.009	.617±.020	.642±.005	.529±.012	.775±.005	.697±.013
ML-KNN	.235±.005	.264±.004	.097±.002	.176±.003	.764±.005	.736±.004	.903±.001	.824±.003
RankSVM	.236±.006	.344±.001	.199±.098	.323±.008	.763±.006	.656±.001	.801±.098	.677±.001
ML-SVM	.232±.005	.337±.009	.179±.004	.314±.002	.768±.005	.662±.009	.822±.004	.686±.002
M3DN	**.108±.003**	**.151±.002**	**.085±.002**	**.117±.002**	**.891±.003**	**.850±.003**	**.915±.003**	**.883±.001**

Methods	Average Precision ↑				Micro AUC ↑			
	FLICKR25K	IAPR TC-12	MS-CoCo	NUS-WIDE	FLICKR25K	IAPRTC-12	MS-CoCo	NUS-WIDE
M3LDA	.371±.005	.311±.007	.399±.007	.338±.005	.693±.006	.609±.002	.773±.005	.657±.008
MIMLmix	.207±.038	.183±.008	.213±.041	.167±.020	.436±.024	.438±.060	.434±.026	.472±.015
CS3G	**.749±.008**	.622±.006	.542±.012	.597±.031	.867±.005	.827±.006	.738±.007	.557±.021
DeepMIML	.621±.027	.619±.025	.633±.005	.583±.008	.835±.009	.802±.017	.914±.002	.852±.003
M3MIML	.423±.056	.490±.020	.446±.030	.443±.076	.745±.034	.707±.017	.816±.020	.762±.020
MIMLfast	.432±.064	.339±.013	.413±.005	.365±.021	.712±.022	.540±.010	.745±.012	.630±.005
SLEEC	.608±.006	.473±.010	.565±.003	.392±.007	.824±.004	.736±.005	.795±.002	.701±.005
Tram	.653±.011	.523±.008	.494±.007	.336±.002	.842±.003	.782±.007	.883±.006	.554±.002
ECC	.416±.012	.278±.011	.462±.007	.438±.014	.646±.004	.514±.008	.779±.005	.702±.009
ML-KNN	.398±.006	.403±.010	.585±.002	.439±.006	.752±.005	.729±.003	.905±.002	.817±.004
RankSVM	.467±.005	.364±.004	.427±.066	.401±.001	.748±.005	.649±.004	.791±.093	.680±.003
ML-SVM	.466±.006	.367±.006	.441±.007	.443±.007	.753±.004	.656±.009	.825±.004	.724±.001
M3DN	.719±.006	**.634±.003**	**.680±.005**	**.691±.001**	**.876±.003**	**.834±.001**	**.918±.002**	**.877±.003**

Table 2: Comparison results (mean ± std.) of M3DN with compared methods on WKG Game-Hub dataset. 6 commonly used criteria are evaluated. The best performance for each criterion is bolded. ↑ / ↓ indicate the larger/smaller, the better of a criterion.

Methods	Coverage ↓ ($\times 10^3$)	Macro AUC ↑	Ranking Loss ↓	Example AUC ↑	Average Precision ↑	Micro AUC ↑
M3LDA	1.645±.056	.519±.005	.921±.004	.320±.007	.062±.004	.307±.005
MIMLmix	1.472±.118	.502±.030	.442±.008	.578±.008	.028±.013	.502±.030
CS3G	.424±.017	.550±.018	.364±.017	.651±.017	.241±.020	.619±.015
DeepMIML	.932±.025	.607±.010	.217±.003	.791±.002	.123±.007	.814±.003
M3MIML	N/A	N/A	N/A	N/A	N/A	N/A
MIMLfast	1.239±.072	.509±.024	.297±.022	.703±.022	.128±.019	.711±.027
SLEEC	1.603±.013	.506±.012	.855±.007	.393±.005	.050±.006	.381±.006
Tram	.902±.017	.499±.008	.115±.019	.354±.021	.064±.008	.064±.008
ECC	1.602±.020	.530±.004	.838±.019	.403±.015	.098±.005	.395±.011
ML-KNN	.873±.002	.613±.002	.195±.003	.805±.003	.156±.001	.828±.001
RankSVM	N/A	N/A	N/A	N/A	N/A	N/A
ML-SVM	.949±.029	.471±.006	.228±.010	.783±.008	.131±.003	.803±.007
M3DN	**.311±.032**	**.693±.005**	**.155±.018**	**.840±.018**	**.307±.001**	**.868±.013**



诸葛亮从上线开始到现在都一直受到广大玩家的喜爱，不仅有着超高的颜值，还有着超强的输出能力和无解的大招收割。法师一哥并非浪得虚名。

六神装：抵抗之靴、噬神之书、痛苦面具、极寒风暴、虚无法杖、贤者之书。贤者之书增加高额的法术攻击，后期装备成型伤害爆表。

对线技巧：诸葛亮在对线上是一个比较占优势的英雄，技能和被动都可以消耗敌方英雄，使用一技能时贴近小兵释放，这样被动可以一下叠三层，利用好技能的释放非常重要，二技能既要有走位，也要叠加被动，这样可以充分利用好技能，完成对敌方英雄的消耗。

团战技巧：诸葛亮和高渐离这类爆发型的法师不一样，诸葛亮在团战中可以持续打出高额伤害，而高渐离是瞬间将敌人打残。他们各有所长。

总结：想要玩好诸葛亮，和队友的配合非常重要，因为大招的蓄力时间较长，在此期间队友可能会提前将残血击杀，使得诸葛亮的大招不能延续，也达不到收割的目的，这样他在团战中的作用将会大大减弱。

首先我们来看看排位出场排行榜。高居榜首的是亚瑟，关于亚瑟登顶，这样的排行。其实不用惊讶。亚瑟是新手英雄的上手必选。有肉有输出的亚瑟也是敌方的噩梦。沉默技能也是王者峡谷稀有的控制技能。排位混星老赖就是亚瑟。

射手胜率排行榜的第一名就是狄大人！没错，自从狄仁杰重做以后，风筝起来极其恶心的狄仁杰让很多人都恨的牙痒痒。二技能解控给了狄仁杰很强大的生存能力，不再像从前一样，被近身就是控住一套秒。

法师胜率排行榜的第一名则是高渐离。没办法，高渐离的群体AOE伤害实在是很恐怖，并且CD比较短的大招很有可能在团战的时候开出两波大招。一边开大，一边还有带穿透的AP伤害，强的一批。

再看刺客胜率排行榜的第一名，被李白霸占。很多玩家都抱着，不管重做后的李白如何，至少珍惜当下的李白。这样的心态来打。这个英雄真心是很无解。就算是清理干净野区，点4下小兵的李白还是可以进塔开大。

以上是今日的全部内容，如果想观看更多的资讯，请点击订阅电竞中国，方便第一时间收到我们最新的游戏资讯。

骑白马的不一定是唐僧，也可能是精忠报国的岳飞，但是骑赤兔马的，那肯定就是关羽了，作为王者峡谷中少数几个有坐骑的英雄之一，关羽该怎么玩呢？

实战解析：克制关羽的英雄：蔡文姬。无论是蔡文姬还是大乔，都可以轻松打断关羽的冲锋状态，蔡文姬频繁的晕眩让关羽根本跑不起来

克制关羽的英雄：大乔。大乔的沉默也能让刚刚起步的关羽再次停下，很难进入冲锋状态，无法进入冲锋状态的关羽，技能的威力会大大降低，并且无法起到打乱敌方阵形的作用，停下的关羽对敌人来说就是一个活靶子。

前期：抵抗之靴+破甲弓。抵抗之靴的被动增加韧性，让关羽在被控时能够更快的解除，因为控制是关羽的克星。破甲弓在前期能为关羽增加穿透能力，能更好的用技能打出较高的伤害，冷却缩减也能减少二技能的CD，增强关羽的存活能力。中期：不祥征兆+霸者重装。不祥征兆是一件物理防装，为关羽增加生命的同时也增加了物理防御。

1、扁鹊。定位：法师 /辅助 特长：远程消耗/回复。很多人都以为扁鹊很没用，作为法师伤害爆发又不够，作为辅助奶量又不足，又没有实质性的控制技能这就导致了许多玩家对扁鹊的印象不是很好，认为扁鹊走中路还是辅助都是累赘，一看队友选择了扁鹊就担心受怕的。但是近期扁鹊已经整改了，一技能是可以存放三个药瓶的，能更好的消耗敌人！。

2、姜子牙。定位：法师/辅助 特长：团控/远程消耗。一旦队友选择了姜子牙，很多人都不会理解的，这个英雄前期伤害其实不是很高又非常费蓝量，而且技能命中率又不高，并不是每个都会大招闪现，而且闪现也是要CD的，并不是每次团战都能用得到。这就导致有些人非常不喜欢姜子牙，作为法师和辅助都一样，欠缺的东西太多了。好在近期已经调整，大招可以对防御塔有效攻击，而且连招更方便。

3、韩信。定位：刺客/战士 特长：突进。韩信作为机动性非常高的英雄，到了会玩的召唤师手里真的是非常厉害。但是现在召唤师峡谷里流传着这么一句话，十个韩信九个坑还有一个是真大神。目前确实是这种状况，在被削弱了之后，韩信对操作的要求性更加高！。

Figure 3: Sample test complex articles predictions of the WKG Game-Hub.

modal input. As to the multi-label learners, we first calculate bag-level feature representation for different modalities independently, then all modalities are concatenated together as a single modal input. In detail, the compared methods are listed as:

M3LDA: makes the topic decided by the visual information and the topic decided by the textual information to be consistent, leading to the label assignment;

MIMLmix: is a hierarchical Bayesian network, the instances are sampled from a mixture model where components are representations of labels in multiple modalities;

CS3G: handles types of interactions between multiple labels and utilizes the data from different modalities;

DeepMIML: exploits deep neural network to generate instance representation for MIML;

M3MIML: learns from multi-instance multi-label examples by maximum margin strategy;
MIMLfast: is a fast multi-instance multi-label method;
SLEEC: learns a small ensemble of local distance preserving embeddings, which can accurately predict infrequently occurring labels;
Tram: is a transductive multi-label classification algorithm via label set propagation;
ECC: state-of-the-art supervised ensemble multi-label method;
ML-KNN: is a kNN style multi-label classification algorithm which often outperforms other existing multi-label algorithms;
RankSVM: learns a ranking function for multi-label classification;
ML-SVM: proposes a new training strategy, i.e., cross-training, to build multi-label classifiers.

4.3 Benchmark Comparisons

M3DN is compared with other methods on 4 benchmark datasets to demonstrate the ability. Results of compared methods and M3DN are listed in Tab. 1. From the results, it is obvious that our M3DN approach can achieve the best or second performance on most datasets with different performance measures, which reveals that the M3DN approach is a high-competitive multi-modal multi-label learning method.

4.4 Complex Article Classification

In this subsection, M3DN approach is tested on the real-world complex article classification problem, i.e., WKG Game-Hub dataset. There are 13,570 articles in collection, image and text modalities are used to help the classification. Specifically, each article contains variable number of images and text paragraphs. Thus, each article can be divided into the image bag and text bag. Comparison results against compared methods are listed in Tab. 2, where notation “N/A” means a method cannot give a result in 60 hours. Similarly, 6 measurement criteria are used as in previous subsection, i.e., Coverage, Ranking Loss, Average Precision, Macro AUC, example AUC and Micro AUC. It can be found that our M3DN approach can get the best results overall criteria, which validates the effectiveness of our method solving the complex article classification problem. In addition, this approach has assisted successfully for Jan. article recall of the WKG Game-Hug.

Figure 3 shows 4 illustrative examples of the classification results on the WKG Game-Hub dataset. Qualitatively, the illustration of the predictions clearly discover the modal-instance-label relation on the test set.

4.5 Label Correlations Exploration

Considering that M3DN can learn label correlation explicitly. In this subsection, we examine the effectiveness of M3DN in label correlations exploration. Due to the page limitation, the exploration is conducted on the real-world dataset WKG Game-Hug. We randomly sampled 27 labels, and the ground metric learned by M3DN is shown in Figure 4, and we scale the original value in cost matrix into $[-1, 1]$. Red color indicates a positive correlation, and blue one indicates a negative correlation. We can see that the learned pairwise cost accords with intuitions. Taking a few examples, the cost between (Overwatch, Tencent) indicates a very small correlation, and this is reasonable since the game Overwatch has no correlation with the Tencent. While the cost between (Zhuge Liang, Wizard) indicates a very strong correlation, since Zhuge Liang belongs to the wizard role in the game.

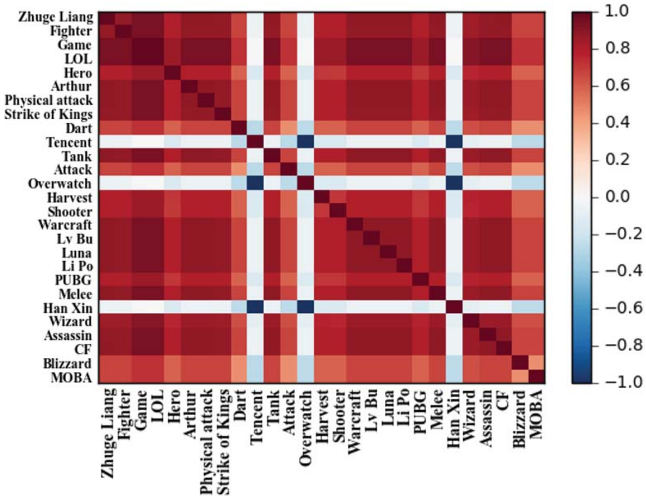


Figure 4: Illustration of learned label correlations for different datasets, and the value has been scaled in [-1,1]. Red color indicates a positive correlation, and blue one indicates a negative correlation.

5 CONCLUSION

Complex objects, i.e., the articles, the videos, etc, can always be represented with multi-modal multi-instance information, while with multiple labels. However, we usually only have bag-level consistency between different modalities, rather than the previous instance-level consistent. Therefore, Multi-modal Multi-instance Multi-label (M3) learning provides a framework for handling such task, on the other hand, previous methods rarely consider the label correlation. In this paper, we propose a novel Multi-modal Multi-instance Multi-label Deep Network (M3DN), which learns the label prediction and exploits label correlation simultaneously based on the Optimal Transport (OT) theory. Experiments on the real world benchmark datasets and special complex article dataset WKG Game-Hub validate the effectiveness of the proposed method. In the future, how to extend to semi-supervised scenario is a very interesting work.

ACKNOWLEDGMENT

This research was supported by the National Key R&D Program of China (2018YFB1004300), NSFC (61773198, 61632004), the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University, and Tencent fund.

REFERENCES

[1] Dimitris Bertsimas and John N Tsitsiklis. 1997. *Introduction to linear optimization*. Vol. 6. Athena Scientific Belmont, MA.

[2] Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. 2015. Sparse Local Embeddings for Extreme Multi-label Classification. In *Advances in Neural Information Processing Systems 28*. Quebec, Canada, 730–738.

[3] Wei Bi and James T. Kwok. 2014. Multilabel Classification with Label Correlations and Missing Labels. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*. Quebec, Canada, 1680–1686.

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

[5] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval*. Santorini Island, Greece.

[6] Nicolas Courty, Remi Flamary, Devis Tuia, and Alain Rakotomamonjy. 2017. Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 9 (2017), 1853–1865.

[7] Marco Cuturi. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems 26*. Lake Tahoe, Nevada, 2292–2300.

[8] Marco Cuturi and David Avis. 2014. Ground metric learning. *Journal of Machine Learning Research* 15, 1 (2014), 533–564.

[9] 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. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding* 114, 4 (2010), 419–428.

[10] Zheng Fang and Zhongfei (Mark) Zhang. 2012. Simultaneously Combining Multi-view Multi-label Learning with Maximum Margin Classification. In *Proceedings of the 12th International Conference on Data Mining*. Brussels, Belgium, 864–869.

[11] Ji Feng and Zhi-Hua Zhou. 2017. Deep MIML Network. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. San Francisco, California, 1884–1890.

[12] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso A. Poggio. 2015. Learning with a Wasserstein Loss. In *Advances in Neural Information Processing Systems 28*. Quebec, Canada, 2053–2061.

[13] Ross B. Girshick. 2015. Fast R-CNN. In *Proceedings of the 25th International Conference on Computer Vision*. Santiago, Chile, 1440–1448.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

[15] Judy Hoffman, Erik Rodner, Jeff Donahue, Brian Kulis, and Kate Saenko. 2014. Asymmetric and Category Invariant Feature Transformations for Domain Adaptation. *International Journal of Computer Vision* 109, 1-2 (2014), 28–41.

[16] Sheng-Jun Huang, Wei Gao, and Zhi-Hua Zhou. 2014. Fast Multi-Instance Multi-Label Learning. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*. Quebec, Canada, 1868–1874.

[17] Sheng-Jun Huang and Zhi-Hua Zhou. 2012. Multi-Label Learning by Exploiting Label Correlations Locally. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*. Ontario, Canada.

[18] Yan Huang, Wei Wang, and Liang Wang. 2015. Unconstrained Multimodal Multi-Label Learning. *IEEE Transactions Multimedia* 17, 11 (2015), 1923–1935.

[19] Mark J. Huiskes and Michael S. Lew. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval*. British Columbia, Canada, 39–43.

[20] Thorsten Joachims. 2002. Optimizing search engines using click through data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Alberta, Canada, 133–142.

[21] Dor Kedem, Stephen Tyree, Kilian Q. Weinberger, Fei Sha, and Gert R. G. Lanckriet. 2012. Non-linear Metric Learning. In *Advances in Neural Information Processing Systems 25*. Lake Tahoe, Nevada, 2582–2590.

[22] Xiangnan Kong, Michael K. Ng, and Zhi-Hua Zhou. 2013. Transductive Multilabel Learning via Label Set Propagation. *IEEE Transaction Knowledge Data Engineer* 25, 3 (2013), 704–719.

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the 13th European Conference Computer Vision*. Zurich, Switzerland, 740–755.

[24] Cam-Tu Nguyen, Xiaoliang Wang, Jing Liu, and Zhi-Hua Zhou. 2014. Labeling Complicated Objects: Multi-View Multi-Instance Multi-Label Learning. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*. Quebec, Canada, 2013–2019.

[25] Cam-Tu Nguyen, De-Chuan Zhan, and Zhi-Hua Zhou. 2013. Multi-Modal Image Annotation with Multi-Instance Multi-Label LDA. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. Beijing, China, 1558–1564.

[26] Wei Qian, Bin Hong, Deng Cai, Xiaofei He, and Xuelong Li. 2016. Non-Negative Matrix Factorization with Sinkhorn Distance. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. New York, NY, 1960–1966.

[27] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning* 85, 3 (2011), 333–359.

[28] Antoine Rolet, Marco Cuturi, and Gabriel Peyre. 2016. Fast Dictionary Learning with a Smoothed Wasserstein Loss. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Cadiz, Spain, 630–638.

[29] Cedric Villani. 2008. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.

[30] Pei Yang and Jingrui He. 2015. Model Multiple Heterogeneity via Hierarchical Multi-Latent Space Learning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. NSW, Australia, 1375–1384.

[31] Pei Yang, Jingrui He, Hongxia Yang, and Haoda Fu. 2014. Learning from Label and Feature Heterogeneity. In *Proceedings of the 14th IEEE International Conference on Data Mining*. Shenzhen, China, 1079–1084.

[32] Pei Yang, Hongxia Yang, Haoda Fu, Dawei Zhou, Jieping Ye, Theodoros Lappas, and Jingrui He. 2016. Jointly Modeling Label and Feature Heterogeneity in Medical Informatics. *ACM Transactions on Knowledge Discovery from Data* 10, 4 (2016), 39:1–39:25.

[33] Han-Jia Ye, De-Chuan Zhan, Xiaolin Li, Zhen-Chuan Huang, and Yuan Jiang. 2016. College Student Scholarships and Subsidies Granting: A Multi-modal Multi-label Approach. In *Proceedings of the 16th International Conference on Data Mining*. Barcelona, Spain, 559–568.

[34] Jianbo Ye, Panruo Wu, James Z. Wang, and Jia Li. 2017. Fast Discrete Distribution Clustering Using Wasserstein Barycenter With Sparse Support. *IEEE Transactions Signal Processing* 65, 9 (2017), 2317–2332.

[35] R Yossi, LJ Guibas, and C Tomasi. 1997. The earth mover's distance multi-dimensional scaling and color-based image retrieval. In *Proceeding of the ARPA image understanding workshop*.

[36] Wang Zhan and Min-Ling Zhang. 2017. Inductive Semi-supervised Multi-Label Learning with Co-Training. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. NS, Canada, 1305–1314.

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

[38] Min-Ling Zhang and Zhi-Hua Zhou. 2008. M3MIML: A Maximum Margin Method for Multi-instance Multi-label Learning. In *Proceedings of the 8th International Conference on Data Mining*. Pisa, Italy, 688–697.

[39] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26, 8 (2014), 1819–1837.

[40] Peng Zhao and Zhi-Hua Zhou. 2018. Label Distribution Learning by Optimal Transport. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. New Orleans, Louisiana.