Deep Robust Unsupervised Multi-Modal Network

Yang Yang,[1] **Yi-Feng Wu,**[1] **De-Chuan Zhan,**[1] **Zhi-Bin Liu,**[2] **Yuan Jiang**[1]
[1]National Key Laboratory for Novel Software Technology, Nanjing University, [2]Tecent
{yangy, wuyf, zhandc, jiangy}@lamda.nju.edu.cn, lewiszbliu@tencent.com

Abstract

In real-world applications, data are often with multiple modalities, and many multi-modal learning approaches are proposed for integrating the information from different sources. Most of the previous multi-modal methods utilize the modal consistency to reduce the complexity of the learning problem, therefore the modal completeness needs to be guaranteed. However, due to the data collection failures, self-deficiencies, and other various reasons, multi-modal instances are often incomplete in real applications, and have the inconsistent anomalies even in the complete instances, which jointly result in the inconsistent problem. These degenerate the multi-modal feature learning performance, and will finally affect the generalization abilities in different tasks. In this paper, we propose a novel Deep Robust Unsupervised Multi-modal Network structure (DRUMN) for solving this real problem within a unified framework. The proposed DRUMN can utilize the extrinsic heterogeneous information from unlabeled data against the insufficiency caused by the incompleteness. On the other hand, the inconsistent anomaly issue is solved with an adaptive weighted estimation, rather than adjusting the complex thresholds. As DRUMN can extract the discriminative feature representations for each modality, experiments on real-world multi-modal datasets successfully validate the effectiveness of our proposed method.

Introduction

With the development of data collection techniques, a huge number of multi-modal data can be collected from different channels, e.g., the articles are always with image and text information, the videos include the image, audio and text. And multi-modal learning approaches aim to utilize these multiple information, in which different modalities can complement each other to improve the generalization abilities of the whole learners, e.g., Yang et al. (2015) extracted informative features of weak modality with the auxiliary strong modalities; Arora, Mianjy, and Marinov (2016) studied the partial least square problem as a stochastic optimization problem.

Considering the labeling costs, subspace embedding based unsupervised multi-modal methods, have been attracted many researches, which mainly aim to obtain a discriminative latent subspace shared by multiple modalities. As a consequence, with the learned subspace, it is straightforward to conduct the subsequent tasks, such as clustering, retrieval, etc. Recently, incorporated with the deep models, the deep unsupervised multi-modal methods are also proposed, e.g., Ngiam et al. (2011) proposed the multi-modal deep auto-encoder to learn a shared representation between different modalities; Kan, Shan, and Chen (2016) used a multi-modal deep network MvDN to seek a non-linear discriminant and modal-invariant representations. It is notable that previous multi-modal learning methods mainly concentrate on utilizing the consistency principle between different modalities, which can reduce the complexity of the learning problem, therefore, it is necessary to ensure that all instances have complete modal information and consistency between different modalities.

While in real applications, note that there are many reasons for incompleteness, including data collection failures caused by the damage of data sensors, data corruptions from network communication, data privacy policies, etc, e.g., the articles may miss the images or texts as shown in Fig. 1. Existing multi-modal learning approaches cannot directly apply on the incomplete modal situation unless removing the incomplete instances, yet the model trained will clearly loses information. Aiming at this issue, there are some preliminary investigations, Shao et al. (2016) learned the latent feature matrices for each incomplete modality and pushes them towards a common consensus; Yang et al. (2018) utilized the extrinsic information from unlabeled data against the insufficiencies brought by the incomplete modal issues. However, these methods are mainly linear methods, which are difficult to extend to non-linear situation, and rarely consider the inconsistent anomalies in the complete situation.

On the other hand, considering the noise effects, even the complete instances are not necessarily consistent, which can be defined as "inconsistent anomalies". Note that different from standard single-modal anomalies, there are two cases of inconsistent anomalies defined as (Zhao and Fu 2015), i.e., class-anomalies exhibit inconsistent characteristics across different modalities; feature-anomalies exhibit inconsistency on all modalities as shown in Fig. 5. Thus, Iwata and Yamada (2016) proposed probabilistic latent variable models for multi-modal anomaly detection; Fan et al. (2017) considered the confidence levels of both modalities and instances. While these methods are always with many hyper-

parameters for adjusting, which are unprocurable.

In conclusion, there are many reasons for modal inconsistency in real applications, including the incompleteness and the inconsistent anomalies. To solve these problems, DRUMN utilizes the deep energy based model for each modality to handle the heterogeneous incomplete multi-modalities, while maximizing the consistency among the homogeneous multi-modalities simultaneously. With the heterogeneous information and consistency constraints, the incompleteness problem can be relieved. Meanwhile, an adaptive weight estimation for inconsistent anomalous instances can be naturally embedded into our proposed approach, thus it can also adopt for eliminating the influence of anomalies rather than setting thresholds as traditional methods.

In the following of this paper, we start with a brief review of related works. Then give the DRUMN approach and the experimental results. After that, we conclude the paper.

Related Work

The exploitation of multi-modal learning has attracted much attention recently. And the basic assumption behind these methods is the consistent principle, while realistic applications are hard to satisfy this assumption. In this paper, our method concentrates on deep robust multi-modal feature embedding in an unsupervised scenario for handling the inconsistent problem. Therefore, our work is related to unsupervised multi-modal feature embedding and robust multi-modal learning.

Most unsupervised multi-modal methods are mainly based on subspace learning, which fully utilize the consistency between multiple modalities to find a discriminative shared subspace, e.g., Shrivastava et al. (2015) proposed CCCM, which enforces the consistency across all available modalities; Rupnik and Shawe-Taylor (2010) proposed the multi-modal CCA (MCCA) to find a common subspace for different modalities. Considering that deep networks can learn nonlinear feature representations without suffering from the drawbacks of nonparametric models, Andrew et al. (2013) used the Deep Canonical Correlation Analysis (DCCA) to learn complex nonlinear transformations for two modalities; Wang et al. (2015) proposed the deep canonically correlated auto-encoders (DCCAE), which combines the DCCA and deep auto-encoders in one unified framework. These multi-modal methods mainly utilize the modal consistency with complete modalities. However, multi-modal instances are always with incomplete features and exist inconsistent anomalies.

Therefore, many researchers have devoted to learning robust multi-modal methods. For incomplete problem, Li, Jiang, and Zhou (2014) established a latent representation where the different modalities of the same example are close to each other; Shao, He, and Yu (2015) proposed the MIC based on weighted nonnegative matrix factorization with $L_{2,1}$ regularization. However, these methods are mainly linear models, which are difficult to learn more discriminative feature representation, and rarely consider the inconsistent anomalies. On the other hand, handling the anomalous data is a relatively new topic, Iwata and Yamada (2016) proposed probabilistic latent variable models for multi-modal anomaly detection; Zhao and Fu (2015) proposed a novel dual-regularized multi-modal outlier detection method. However, the performance of these methods is subject to the hyper-parameters, which are sensitive to noise.

In this paper, a novel Deep Robust Unsupervised Multi-modal Network (DRUMN) is proposed, which considers both the incomplete and anomalous multi-modal data in a unified framework to solve the inconsistent problem. Specifically, DRUMN utilizes the deep energy based model for each modality to handle the heterogeneous multi-modal data, while maximizing the consistency between the homogeneous multi-modal examples simultaneously, besides, an adaptive weight estimation method is utilized for inconsistent anomalies detection, which considers the energy variance of the sample uncertainty sequence, rather than setting thresholds of sample uncertainty for eliminating both class and feature anomalies jointly. Finally, more discriminative feature can be achieved.

Proposed Method

In the incomplete multi-modal setting, an instance can be characterized by multiple modal features. Suppose we are given N examples with K modalities. The $k-$th modality of $i-$th instance $\mathbf{x}_i$ can be represented as $\mathbf{x}_{i^k} \in \mathbb{R}^{d_k}$, where d_k is the dimension of the $k-$th modality. It is notable that each instance may has complete or partial modalities, suppose we have N_c homogeneous examples with complete modal features, meanwhile, we have N_k heterogeneous instances for each modality. Thus, the whole dataset can be represented as $\mathcal{D} = \{\mathbf{X}_c, X_1, X_2, \cdots, X_K\}$, where $\mathbf{X}_c = \{(\mathbf{x}_{i^1}, \mathbf{x}_{i^2}, \cdots, \mathbf{x}_{i^K})\}_{i=1}^{N_c} \in \mathbb{R}^{N_c \times d}$ denotes the examples presenting in all modalities, $d = d_1 + d_2 + \cdots + d_K$, $X_k \in \mathbb{R}^{N_k \times d_k}$ denotes the incomplete examples presenting in the $k-$th modality. Without any loss of generality, we consider two modalities in this paper, i.e., image and text.

The Formulation of DRUMN

The goal of DRUMN is to learn discriminative feature representations for each modality with the incomplete and inconsistent anomalous multi-modal data. In this section, we mainly introduce the concrete steps on how to construct a robust deep network. There are several different setting:

Threshold Based Deep Network

Combing the canonical correlation analysis (CCA) and reconstruction-based objective, Wang et al. (2015) proposed a deep multi-modal method DCCAE, which consists two distinct auto-encoder networks for different modalities, and optimizes the combination of canonical correlation between the learned bottleneck representations and the auto-encoder reconstruction errors. Considering that some examples only have partial modalities, e.g., some articles about " Strike of Kings" only have contents or images information as shown in Fig. 1. Thus, the CCA term maximizes the mutual information between different projected modal latent representations of the complete instances, while auto-encoder network of each modality can be used to minimize the reconstruction error of all the instances, including the complete and incomplete instances. Without any loss of generality, DCCAE can

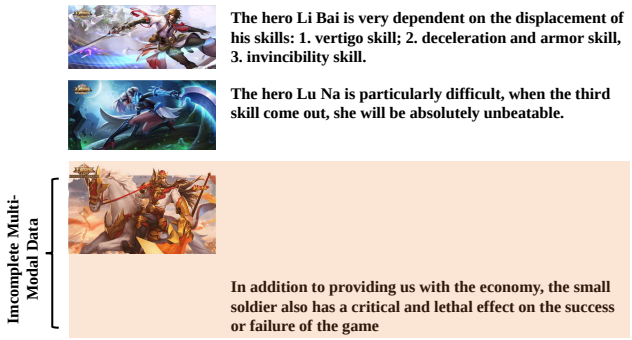

Figure 1: An illustration of the inconsistency due to incompleteness. The data are from the "Strike of Kings" forum paragraph, which can be represented with image and content information. The incomplete modal instances are with yellow shallows.

be reformulated as:

$$\begin{aligned}\min_{\Theta_{f_k},\Theta_{r_k},U_k} &\sum_{m\neq n}^{K}\sum_{i=1}^{N_c} -\frac{1}{N_c}\operatorname{tr}(U_m^\top f_m(\mathbf{x}_{i^m}) f_n(\mathbf{x}_{i^n})^\top U_n) + \\ &\lambda\sum_{k=1}^{K}\sum_{i=1}^{N_c+N_k}\|\mathbf{x}_{i^k} - r_k(f_k(x_{i^k}))\|_F^2 \\ s.t.\quad & U_m^\top(\frac{1}{N_c} f_m(\mathbf{x}_{i^m}) f_m(X_{i^m})^\top + \gamma_m I)U_m = I \\ & \mathbf{u}_{m_i}^\top f_m(\mathbf{x}_{i^m}) f_n(\mathbf{x}_{j^n})^\top \mathbf{u}_{n_j} = 0 \quad i \neq j\end{aligned} \tag{1}$$

where $\Theta_{f_k}, \Theta_{r_k}$ are the weight parameters of encoder network f_k and decoder network r_k of the $k-$th modality, $U_k = [\mathbf{u}_{k_1}, \mathbf{u}_{k_2}, \cdots, \mathbf{u}_{k_L}]$ are the CCA directions that project to the output space, L is the shared dimension of latent space. $\gamma_m > 0$ is regularization parameter for same covariance estimation (Hardoon, Szedmak, and Shawe-Taylor 2004), the $U_k^\top f_k(\mathbf{x}_{i^k})$ is the final projection mapping for testing. $\lambda > 0$ is the trade-off parameter. Therefore, considering the trade-off between the consistent term of different modalities, and the information captured in the bottleneck representations from auto-encoder term, DCCAE can learn more discriminative feature representations.

On the other hand, note that real-world data always contain inconsistent entries that result in the unreliable multi-modal data, as a matter of fact, the affections of anomalies become one of the barriers for modeling robust models. Inconsistent multi-modal anomalies have two varieties, as shown in Fig. 2, paragraph 3 is a class-anomaly since the content and image are not consistent, while paragraph 4 is a feature-anomaly since it is an unrelated advertisement. To solve this problem, a deep structure energy based approach (DSEBMs) (Zhai et al. 2016) for anomaly detection can be adopted naturally in the improved DCCAE framework. Specifically, the energy function is the output of deterministic deep neural networks, i.e., MLP, CNN, RNN, the energy function can be represented of L-layers deep EBM structure

Figure 2: An illustration of the inconsistency due to the inconsistent anomalies. The data are from "Strike of Kings" forum paragraph. Para. 1 and Para. 2 are normal paragraph with consistent image and content information. The anomalous instances are with yellow shallows. Para. 3 is a class anomaly, in which different modalities are inconsistent. Para. 4 is a feature anomaly, which is an irrelevant adverting paragraph.

as:

$$\begin{aligned}E_k(\mathbf{x}_{i^k};\theta) &= \frac{1}{2}\|\mathbf{x}_{i^k} - b'\|_2^2 - \sum_{j=1}^{N_l} h_{l,j} \\ s.t.\quad & h_l = g(f(h_{l-1})), \quad l \in \{1, \cdots, L\}\end{aligned} \tag{2}$$

where N_l is the dimensionality of the $l-$th layer, $g(x)$ is the function $log(1+e^x)$, and $f(h_{l-1})$ is the particular operator for different networks, i.e., for fully connect network, $f(h_{l-1}) = W_l^\top h_{l-1} + b_l$, where $W_l \in \mathbb{R}^{N_{l-1}\times N_l}$, $b_l \in \mathbb{R}^{N_l}$, h_{l-1} is the output of the previous layer. $b' \in \mathbb{R}^{d_k}$, and the term $\|\mathbf{x}_{i^k} - b'\|_2^2$ acts as a prior, which punishes the probability of the inputs that far away from b'.

It is notable that the Restricted Boltzmann Machine (RBM), as one of the most well known EBM model, is proved closely related to a variant of auto-encoders method DAEs (Vincent et al. 2010). Particularly, Vincent (2011) showed that using score matching (SM) (Hyvarinen 2005), an alternative method to MLE, and can be used to estimate EBM. Consequently, training RBM is equivalent to a one-layer DAE. In detail, SM minimizes the following objective function: $J(\theta) = \frac{1}{2}\int_x p_x(\mathbf{x})\|\Psi(\mathbf{x};\theta) - \Psi(\mathbf{x}))\|$, where $p_x(\mathbf{x})$ is the true data distribution which is unknown, $\Psi(x;\theta) = \nabla_x \log p(\mathbf{x};\theta) = -\nabla_x E(\mathbf{x};\theta)$ and $\Psi_x(\mathbf{x}) = \nabla_x \log p_x(\mathbf{x})$ are the score function of the model and the true density function, respectively. Vincent (2011) showed that by approximating the $p_x(\mathbf{x})$ with the Parzen window density, minimizing $J(\theta)$ is the same form as an auto-encoder in DAEs, defined as:

$$r(f(\mathbf{x};\Theta_f);\Theta_r) = \mathbf{x} - \nabla_x E(\mathbf{x};\Theta) \tag{3}$$

Where $\Theta = \{\Theta_f, \Theta_r\}$, and $E(\mathbf{x};\Theta)$ is the energy function defined in Eq. 2. Thus, Eq. 3 can replace the network structure of auto-encoder term in Eq. 1 equivalently, which can detect anomaly more easily.

To perform the anomaly detection, we can select following two criteria: first, the samples that are assigned probability lower than the pre-defined threshold, i.e., $E(\mathbf{x};\Theta) < E_{th}^k$; another criterion is based on the reconstruction error, i.e., $\|\mathbf{x} - r(f(\mathbf{x};\Theta_f);\Theta_r)\|_F^2 = \|\nabla_x E(\mathbf{x};\Theta)\|_F^2 \geq Er_{th}^k$, in which high reconstruction errors correspond to examples whose energy has large gradient norms. In other words, instances with low energy probabilities and high reconstruction errors are viewed as anomalies.

Thus, when expanding to incomplete and anomalous multi-modal data setting, combining Eq. 1, Eq. 2 and Eq. 3 in a unified framework, we can formulate the threshold based deep framework as:

$$\min_{\Theta_{f_k},\Theta_{r_k},U_k} \sum_{m\neq n}^{K}\sum_{i=1}^{N_C} \ell_{co}^{m,n} + \lambda \sum_{k=1}^{K}\sum_{i=1}^{N_C+N_k} \ell_{re}^{k} \qquad (4)$$

where

$\ell_{co}^{m,n} = \max\{0, \frac{1}{N_c}\operatorname{tr}(U_m^\top f_m(\mathbf{x}_{i^m}) f_n(\mathbf{x}_{i^n})^\top U_n) - Eco_{th}^{m,n}\}$,

the $Eco_{th}^{m,n}$ is the hyper-parameter for eliminating the class anomalies, considering the two criteria mentioned in the EBM model, we can formulate the loss of reconstruction as

$\ell_{re}^{k} = \max\{0, E(\mathbf{x}_{i^k};\Theta) - E_{th}^k\} + \max\{0, Er_{th}^k - \|\nabla_{\mathbf{x}} E_k(\mathbf{x}_{i^k};\Theta)\|_F^2\}$,

the E_{th}^k and Er_{th}^k are the hyper-parameters for eliminating the feature anomalies. However, we find that there is large number of hyper-parameters for adjusting, i.e., $\frac{K\times(K+3)}{2}$, which is unprocurable.

Weighted Based Deep Network

To overcome the disadvantage of the threshold based deep model, we put forward a novel Deep Robust Unsupervised Multi-modal Network approach (DRUMN), which can eliminate inconsistent anomaly configurations naturally by adaptive learning the weights of different instances on different modalities.

Without any loss of generality, aiming at learning the discriminative feature representation for each modality, we can utilize the framework of Eq. 4 for handling the incomplete multi-modal data. Meanwhile, for the second target, we wish to adaptively update the weight of each instance on different modalities rather than adjusting the thresholds manually. Inspired from the active learning setting, the prediction variance can be used to measure the uncertainty of each sample for either regression or classification problems (Schein and Ungar 2007), and in the unsupervised setting, we can refer to the energy variance instead. Considering the anomaly setting, the instances with low energy variances are always easy instances which are more convinced, or anomalies that are always hard to be notarized, and in order to gain more information at each iteration, we prefer to choose the samples with high energy variances, which are more uncertain. Since the energy variances are estimated online, the weights can be calculated based on the estimated variances plus the confidence interval as (Chang, Learned-Miller, and McCallum 2017). Thus, the energy variance can be formulated as:

$$\omega_{i^k} = \hat{std}_{i^k}^{conf}(H)$$
$$\hat{std}_{i^k}^{conf} = \sqrt{\hat{var}(E_{H_{i^k}^{t-1}}(\mathbf{x}_{i^k};\Theta)) + \frac{\hat{var}(E_{H_{i^k}^{t-1}}(\mathbf{x}_i;\Theta))^2}{|H_{i^k}^{t-1}|-1}} \qquad (5)$$

where $\hat{var}(E_{H_{i^k}^{t-1}}(\mathbf{x}_{i^k};\Theta))$ is the energy variance estimated by history energy $H_{i^k}^{t-1}$ of $i-$th instance on $k-$th modality, and $|H_{i^k}^{t-1}|$ is the number of stored energy probability, we define as 8 in the experiments. Similarly, the correlation weight can be defined as $\gamma_{i^{m,n}}$ between $m-$th modality and $n-$th modality of $i-$th instance, i.e., $\gamma_{i^{m,n}} = \sqrt{\hat{var}(C_{H_{i^m,i^n}^{t-1}}(\mathbf{x}_{i^m},\mathbf{x}_{i^n})) + \frac{\hat{var}(C_{H_{i^m,i^n}^{t-1}}(\mathbf{x}_{i^m},\mathbf{x}_{i^n}))^2}{|H_{i^m,i^n}^{t-1}|-1}}$, where $C(\cdot)$ is the mutual information function.

Note that in the multi-modal setting, ω_{i^k} can be used to eliminate the feature-anomalies, which are with low weights on all the modalities, $\gamma_{i^{m,n}}$ can be used to eliminate the class-anomalies, which are with low weights between $m-$th modality and $n-$th modality of $i-$th instance. Substitute $\omega_{i^k}, \gamma_{i^{m,n}}$ into Eq. 1, we have the final formulation:

$$\begin{aligned}\min_{\Theta_{f_k},\Theta_{r_k},U_k} &\sum_{m\neq n}^{K}\sum_{i=1}^{N_C} -\gamma_{i^{m,n}} \frac{1}{N_C}\operatorname{tr}(U_m^\top f_m(\mathbf{x}_{i^m}) f_n(\mathbf{x}_{i^n})^\top U_n) \\ &+ \lambda \sum_{k=1}^{K}\sum_{i=1}^{N_C+N_k} \omega_{i^k}\|\mathbf{x}_{i^k} - r_k(f_k(x_{i^k}))\|_F^2 \\ s.t.\quad &\text{the same as the constraints of Eq. 1}\end{aligned} \qquad (6)$$

Optimization

The objective couples all training samples through the whitening constraints as in Eq. 1, thus, standard stochastic gradient descent cannot be applied. Wang et al. (2015) proved that DCCA can still be optimized efficiently when the gradient is estimated with a sufficiently large mini-batch, which owing to large mini-batch contains enough information for estimating the covariances. The whole procedure is summarized in Algorithm 1.

Experiments

In this section, we first introduce the datasets in brief and then give the empirical results of DRUMN and compared methods.

Datasets and Configurations

DRUMN can learn more discriminative feature representations for each modality by considering the incomplete and anomalous multi-modal data in a unified framework. With the learned features, we can conduct further tasks, i.e., retrieval, clustering. In this section, we will provide the empirical investigations and performance comparison of DRUMN on cross-modal retrieval and anomaly detection. In particular, we experiment on 4 public real-world datasets, i.e., FLICKR25K, IAPR TC-12, WIKI and NUS-WIDE, and 1 real-world incomplete multi-modal dataset, i.e., WKG Game-Hub.

Figure 3: Examples of retrieval results from WKG Game-Hub dataset, the top row examples are the "T $\rightarrow$ I" and the bottom row examples are the "I $\rightarrow$ T". There are four example queries on the left, and top 5 results are shown on the right (correct results are with blue, otherwise with red).

Algorithm 1 The pseudo code of DRUMN

Input:

- Dataset: $\mathcal{D} = \{\mathbf{X}_c, X_1, X_2, \cdots, X_K\}$
- Parameter: λ, S
- MaxIter: T, learning rate: $\{\alpha_t\}_{t=1}^{T}$

Output:

- Feature embedding network: f_k

1: **repeat**
2: Create Batch: Randomly pick up $|S|$ examples from $\mathcal{D}$ without replacement;
3: Calculate the ω_{ik}, $\gamma_{i^{m,n}} \leftarrow$ Eq. 5;
4: Calculate the loss $L \leftarrow$ Eq. 6;
5: Obtain the derivative $\frac{\partial L}{\partial U_k}$, $\frac{\partial L}{\partial \Theta_{r_k}}$, $\frac{\partial L}{\partial \Theta_{f_k}}$, update parameters $U_k, \Theta_{r_k}, \Theta_{f_k}$;
6: **until** converge or reach the max-iter

- FLICKR25K: (Huiskes and Lew 2008) consists of 25,000 image-text pairs collected from Flickr website. The text is represented as a 1386-dimensional bag-of-words vector;
- IAPR TC-12: (Escalante et al. 2010) consists of 20,000 image-text pairs. The text is represented as a 2912-dimensional bag-of-words vector;
- WIKI: (Rasiwasia et al. 2010) has 2,866 documents extracted from Wikipedia. We represent the text information by 7343-dimensional vectors based on TF-IDF;
- NUS-WIDE: (Chua et al. 2009) selects 195,834 image-text pairs that belong to the 21 most frequent concepts. The text is represented as a 1000-dimensional bag-of-words vector;
- WKG Game-Hub: consists of 32,222 image-text pairs collected from the Game-Hub of " Strike of Kings". The content is represented as a 300-dimensional word2vector vector.

Table 1: MAP of the first 50 rank list of 4 real-world datasets. The best performance is bolded.

Methods	I $\rightarrow$ T			
	FLICKR	IAPR TC-12	WIKI	NUS
DCCA	.5839	.4069	.1711	.4737
DCCAE	.5906	.4409	.1740	.4983
Corr-AE	.5573	.3370	.1780	.4683
Corr-Cross-AE	.5583	.4311	.1770	.4815
Corr-Full-AE	.6376	.3481	.1654	.4460
DRUMN-Thres	**.6416**	.4323	.1567	.5009
DRUMN	.6266	**.4605**	**.1781**	**.5626**
Methods	T $\rightarrow$ I			
	FLICKR	IAPR TC-12	WIKI	NUS
DCCA	.5807	.4472	.1647	.4572
DCCAE	.4518	.4347	.1660	.5021
Corr-AE	.4576	.3423	.1627	.4765
Corr-Cross-AE	.5048	.4568	.1699	.4868
Corr-Full-AE	.4750	.3877	.1643	.4490
DRUMN-Thres	.5097	.4578	.1722	.4776
DRUMN	**.6589**	**.4974**	**.1887**	**.5548**

For each dataset, we randomly select 20% data for the test (query) set and the remaining instances are used for training. Considering that the FLICKR25K, IAPR TC-12, WIKI and NUS-WIDE are completely in raw data, we first conduct the experiments on completely data, then conduct more experiments on segmented incomplete data as in (Yang et al. 2018). To demonstrate the generalization ability, we also experiment on the real-world incomplete multi-modal dataset, i.e., WKG Game-Hub, which contains 27,276 instances with two modalities, and 4946 instances appear with

Table 2: MAP of the first 50 rank list of WKG Game-Hub. The best performance is bolded.

Methods	I → T		T → I	
	L > 1	L > 3	L > 1	L > 3
DCCA	.3259	.2400	.3080	.2172
DCCAE	.3205	.2486	.3212	.2415
Corr-AE	.2682	.2370	.2108	.1445
Corr-Cross-AE	.3378	.2276	.3415	.2148
Corr-Full-AE	.2911	.2031	.1923	.1522
DRUMN-Thres	.2729	.1829	.2574	.1577
DRUMN	**.3594**	**.2629**	**.3448**	**.2495**

single modalities.

To verify the learned feature representations of our method, we examine the task of cross-modal retrieval. More specifically, given a text query, we aim to find images that are most relevant to it. For each query in the test set, we rank the retrieval results using the learned feature representations, without any loss of generality, we adopt the Euclidean ranking protocol, which ranks the retrieval results according to their Euclidean distances to the given query point, in an increasing order. And we adopt the widely used Mean Average Precision (MAP) (Zhu, Shao, and Yu 2014) metric to measure the accuracy of the Euclidean ranking protocol. The deep network for image encoder is implemented the same as Resnet50 (He et al. 2015). 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 in the training phase is tuned in $\{0.1, 0.2, \cdots, 0.9\}$. When the variation between the objective value of Eq. 6 is less than 10^{-4} between iterations, we consider DRUMN converges.

Compared methods

DNN-based unsupervised multi-modal methods are compared in the experiments: DCCA (Andrew et al. 2013), DCCAE (Wang et al. 2015), Corr-AE (Feng, Wang, and Li 2014), Corr-Cross-AE (Feng, Wang, and Li 2014), Corr-Full-AE (Feng, Wang, and Li 2014) and DRUMN-Thres. Since the DCCA can only handle complete modalities, thus we use the incomplete data for initialization. Note that all compared methods are modified as threshold based models for fairness. In detail, the compared methods are listed as:

- DCCA: computes representations of the two modalities by passing them through multiple stacked layers of nonlinear transformation;
- DCCAE: consists two auto-encoders and optimizes the combination of canonical correlation between the learned bottleneck representations and the reconstruction errors of the auto-encoders;
- Corr-AE: first use two uni-modal auto-encoders to learn higher level image feature and text feature respectively, and then use CCA to learn a common representation space on the learned features;
- Corr-Cross-AE: combines immediately cross-modal auto-encoders and CCA;
- Corr-Full-AE: combines full auto-encoders and CCA;
- DRUMN-Thres: threshold based deep network structure, as mentioned in section 2.

Cross Modal Retrieval

The MAP (pre@50) results, which calculate the MAP of the first 50 rankings as (Jiang and Li 2017), the results of DRUMN and other compare methods on 4 real-world datasets are reported in Table 1. Here, "T", "I" represent the text and image separately, e.g., "I → T" denotes the case where the query is image and the retrieval result is text, and "T → I" denotes the case where the query is text and the retrieval result is image. From the Table 1, it reveals that for all datasets, DRUMN almost consistently achieve the significant superior retrieval performance on all datasets comparing to other methods except for the "I → T" on FLICKR25K.

To further verify the effectiveness of DRUMN, we conduct more experiments on the real incomplete multiple modal dataset WKG Game-Hub and record the results in the Table 2. Considering that WKG Game-Hub is a multi-label dataset, which exists the label imbalance problem. On MAP calculation, we measure the similarity between the query instances and ranking results by considering the sharing labels larger than 1 (L>1) or 3 (L>3) labels. From the Table 2, it reveals that DRUMN also achieves the significantly superior retrieval performance.

Figure 3 shows several illustrative examples of the retrieval results on the WKG Game-Hub dataset. Qualitatively, it can observe that DRUMN captures the general latent feature representation represented in both the images and the texts. It is notable that most results shown are correct.

Influence on No. of Incomplete Multi-Modal Data

It is notable that the 4 real world datasets are complete. In order to explore the influence of the ratio of the incomplete modalities on performance, extensive experiments are conducted. In this section, the parameters in each investigation are fixed as the optimal values selected in above investigations, while the ratio of the incomplete data varies in $\{0\%, 10\%, 30\%, \cdots, 90\%\}$, with 20% as interval. Results on 4 datasets, i.e., FLICKR25K, IAPR TC-12, WIKI, NUS-WIDE, are recorded in Figure 4. From these figures, it clearly shows that DRUMN achieves the best on most datasets. Besides, we can also find that DRUMN achieves superiorities from high incomplete ratio, and the performance of DRUMN decreases slower than the compared methods.

Anomaly Detection

DRUMN also considers the anomalous multi-modal data except learning discriminative feature representation. Figure 5 shows several illustrative examples of the anomaly detection results on the WKG Game-Hub dataset. Qualitatively, it reveals that DRUMN can detect both the class and feature anomalies compared to DRUMN-Thres method, in detail, the class and feature anomalies detected by DRUMN

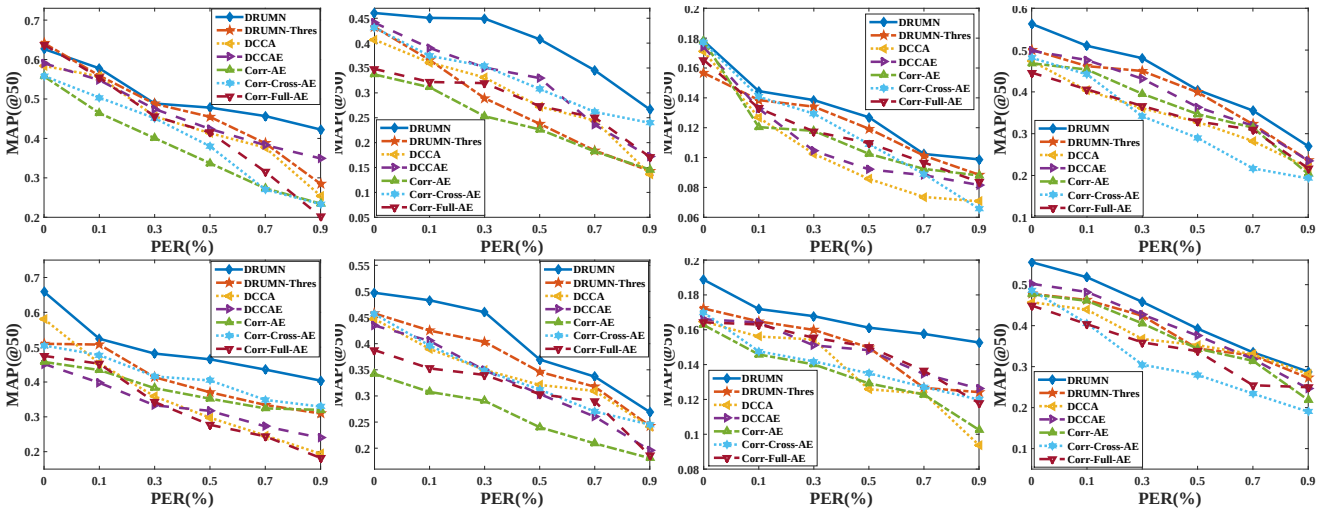

Figure 4: The MAP(@50) results of 4 real-world datasets. The four columns from left to right are FLICKR25K, IAPR TC-12, WIKI, NUS-WIDE datasets in order, the top row examples are the "I $\rightarrow$ T" and the bottom row examples are the "T $\rightarrow$ I". PER (partial example ratio) is the ratio of incomplete examples.

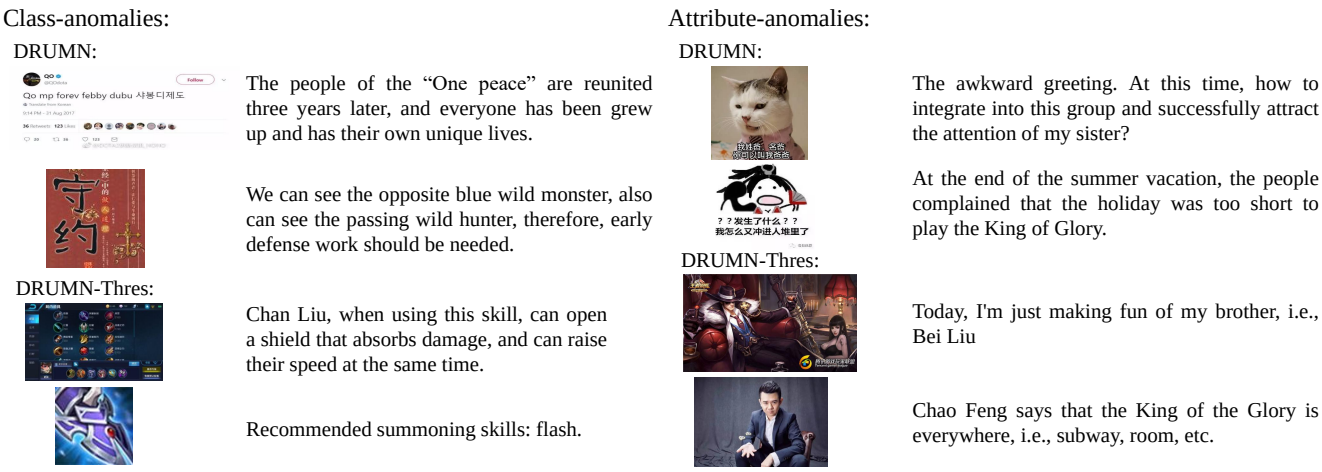

Figure 5: Examples of class and feature anomalies of DRUMN and DRUMN-Thres from WKG Game-Hub dataset. Left column are class anomalies and the right column are attribute anomalies

are all corrected, while the first result in class-anomalies of DRUMN-Thres is wrong.

Conclusion

Feature incompleteness and inconsistent class/feature-anomaly in multi-modal learning scenarios are often take place in real applications, which result in the inconsistent problem. In this work, we propose a Deep Robust Unsupervised Multi-modal Network (DRUMN) for utilizing the extrinsic heterogeneous information, and learning latent representation via a deep unified network. Besides, the model itself can also tackle the inconsistent anomaly problem with adaptively weight estimation. Experiments on real-world multi-modal datasets successfully validate the effectiveness of our proposed method. How to fully incorporate the supervised information into semi-supervised scenario is an interesting future work.

Acknowledgments

The National Key R&D Program of China (2018YFB1004300), NSFC (61773198, 61632004).

References

Andrew, G.; Arora, R.; Bilmes, J. A.; and Livescu, K. 2013. Deep canonical correlation analysis. In *ICML*, 1247–1255.

Arora, R.; Mianjy, P.; and Marinov, T. V. 2016. Stochastic optimization for multiview representation learning using partial least squares. In *ICML*, 1786–1794.

Chang, H.; Learned-Miller, E. G.; and McCallum, A. 2017. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *NIPS*, 1003–1013.

Chua, T.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. NUS-WIDE: a real-world web image database from national university of singapore. In *CIVR*.

Escalante, H. J.; Hernandez, C. A.; Gonzalez, J. A.; Lopez-Lopez, A.; andEduardo F. Morales, M. M.; Sucar, L. E.; Pineda, L. V.; and Grubinger, M. 2010. The segmented and annotated IAPR TC-12 benchmark. *CVIU* 114(4):419–428.

Fan, Y.; Liang, J.; He, R.; Hu, B.; and Lyu, S. 2017. Robust localized multi-view subspace clustering. *CoRR* abs/1705.07777.

Feng, F.; Wang, X.; and Li, R. 2014. Cross-modal retrieval with correspondence autoencoder. In *ACMMM*, 7–16.

Hardoon, D. R.; Szedmak, S. R.; and Shawe-Taylor, J. R. 2004. *Canonical Correlation Analysis: An Overview with Application to Learning Methods*. MIT Press.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Huiskes, M. J., and Lew, M. S. 2008. The MIR flickr retrieval evaluation. In *MIR*, 39–43.

Hyvarinen, A. 2005. Estimation of non-normalized statistical models by score matching. *JMLR* 6:695–709.

Iwata, T., and Yamada, M. 2016. Multi-view anomaly detection via robust probabilistic latent variable models. In *NIPS*, 1136–1144.

Jiang, Q., and Li, W. 2017. Deep cross-modal hashing. In *CVPR*, 3270–3278.

Kan, M.; Shan, S.; and Chen, X. 2016. Multi-view deep network for cross-view classification. In *CVPR*, 4847–4855.

Li, S.; Jiang, Y.; and Zhou, Z. 2014. Partial multi-view clustering. In *AAAI*, 1968–1974.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *ICML*, 689–696.

Rasiwasia, N.; Pereira, J. C.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A New Approach to Cross-modal Multimedia Retrieval. In *ACMMM*, 251–260.

Rupnik, J., and Shawe-Taylor, J. 2010. Multi-view canonical correlation analysis. In *KDD*, 1–4.

Schein, A. I., and Ungar, L. H. 2007. Active learning for logistic regression: an evaluation. *ML* 68(3):235–265.

Shao, W.; He, L.; Lu, C.; and Yu, P. S. 2016. Online multi-view clustering with incomplete views. In *BigData*, 1012–1017.

Shao, W.; He, L.; and Yu, P. S. 2015. Multiple incomplete views clustering via weighted nonnegative matrix factorization with $l_{2,1}$ regularization. In *ECML/PKDD*, 318–334.

Shrivastava, A.; Rastegari, M.; Shekhar, S.; Chellappa, R.; and Davis, L. S. 2015. Class consistent multi-modal fusion with binary features. In *CVPR*, 2282–2291.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; and Manzagol, P. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR* 11:3371–3408.

Vincent, P. 2011. A connection between score matching and denoising autoencoders. *Neural Computation* 23(7):1661–1674.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On deep multi-view representation learning. In *ICML*, 1083–1092.

Yang, Y.; Ye, H.-J.; Zhan, D.-C.; and Jiang, Y. 2015. Auxiliary information regularized machine for multiple modality feature learning. In *IJCAI*, 1033–1039.

Yang, Y.; Zhan, D.; Sheng, X.; and Jiang, Y. 2018. Semi-supervised multi-modal learning with incomplete modalities. In *IJCAI*, 2998–3004.

Zhai, S.; Cheng, Y.; Lu, W.; and Zhang, Z. 2016. Deep structured energy based models for anomaly detection. In *ICML*, 1100–1109.

Zhao, H., and Fu, Y. 2015. Dual-regularized multi-view outlier detection. In *IJCAI*, 4077–4083.

Zhu, F.; Shao, L.; and Yu, M. 2014. Cross-modality submodular dictionary learning for information retrieval. In *CIKM*, 1479–1488.