MAXIMAL CORRELATION EMBEDDING NETWORK FOR MULTILABEL LEARNING WITH MISSING LABELS

Lu Li[1], Yang Li[1], Xiangxiang Xu[2], Shao-Lun Huang[1], Lin Zhang[1]

[1]Tsinghua-Berkeley Shenzhen Institute, Tsinghua University
[2]Department of Electronic Engineering, Tsinghua University
{lilou16@mails., yangli@sz., xuxx14@mails., shaolun.huang@sz., linzhang@}tsinghua.edu.cn

ABSTRACT

Multilabel learning, the problem of mapping each data instance to a subset of labels, appears frequently in many real-world applications. However, obtaining complete label annotation for every instance requires tremendous efforts, especially when the label set is large. As a result, multilabel learning with missing labels remains as a common challenge. Existing works either cannot handle missing labels or lack nonlinear expressiveness and scalability to large label set. In this paper, we present a novel end-to-end solution for multilabel learning with missing labels. Our algorithm, Maximal Correlation Embedding Network learns a low dimensional label embedding using an encoder-decoder architecture. It exploits label similarity through a maximal correlation regularization in the embedded label space to reduce the classification bias due to missing labels. A series of experiments on popular multilabel datasets demonstrate that our approach outperforms state of the art, both in complete data and partially observed data.

Index Terms— Multilabel learning algorithms, multilabel classification, missing labels, embedding network, max correlation regularization

1. INTRODUCTION

One of the most common machine learning problems in multimedia is multi-class classification, where one object is mapped to a single class label. However, most applications in real life have a multilabel nature, i.e. object classes are not mutually exclusive and one object can be assigned to multiple labels. For example, in image annotation, an image can be annotated with several tags. In text categorization, one document can be attached with more than one topic. In the multilabel scenario, many labels are correlated. For instance, in an image scene classification problem, with the prior knowledge that one image is related to *ocean*, we can deduce that it's probably related to *beach* and definitely not to *fitting room*. Therefore, how to effectively utilize label correlation is an important question in multilabel learning.

In addition to the difficulty caused by label correlation, multilabel learning faces another prominent challenge: missing labels. A sufficient amount of training data with accurate labels is required in supervised learning, but it costs huge efforts to obtain the full label set for each sample. Many times people who annotate the data may drop some relevant labels unconsciously. For example, an image containing concepts *building, plaza, sky, street* may be only marked with *building* and *plaza* but be left out with *sky* and *street*.

Existing works in multilabel classification are based on task decomposition or low rank label transformation. For the task decomposition approaches [1, 2], the multilabel problem is decomposed into a set of binary classification problems. However, they either ignore label correlations, or have poor scalability to a large label set. Low rank label transformaton approaches are widely used to handle label correlation. The main idea is that the original label can be embedded into a lower dimensional space in which the dependencies between labels can be removed but all the principal information is remained. These methods then obtain predictions by projecting the compressed vectors back to the original label space. The low rank constraint has been implemented in many different ways, such as matrix decomposition [3, 4] and alternating optimization [5, 6]. Nevertheless, these methods are rather limited in nonlinear expressiveness and some are not capable of handling missing labels.

The missing label issue in multilabel classification has often been studied using label similarity in recent works. With the aid of label similarity, the assignment of one certain label can be inferred from its close labels corresponding with top largest similarities. Some approaches aim to recover missing information by label propagations [7], label reconstructions [8] or transductive learning [9]. Others add a regularization on the label manifold [10, 11]. Yet the nonlinear expressiveness in their proposed models is limited and some still do not scale well to large-scale dataset.

To address the challenge of label correlation, missing label and scalability, we propose *Maximal Correlation Embedding Network* (MCEN), an embedding network for multilabel classification. It extracts label correlation by learning low di-

mensional label embeddings and solves the missing label issue by regularizing label similarity in the embedded space. The main advantages of our work are as follows:

1. We integrate the low rank label transformation into an end-to-end model, which can be compatible with all kinds of internal/fine-grained network architecture. Therefore, our model has strong nonlinear representation ability and scalability to different label cardinality.

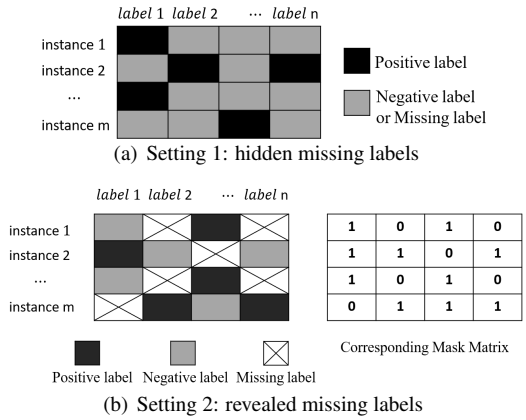
2. The proposed label similarity regularization can handle missing label by maximizing the total correlation among labels with theoretical support.

We did a wide range of experiments on several popular multilabel datasets. In the situation with complete or missing labels, our method performs better than the state of the art consistently.

2. RELATED WORKS

Driven by the broad application prospects, a lot of works have been done to pursuit a better performance in multilabel problems. For the traditional task decomposition methods, two representative algorithms are Binary Relevance (BR) [1] and Classifier Chain (CC) [2]. BR decomposes the task into multiple independent binary classification problems and learns every binary model separately. However, BR ignores the label correlations and this may lead to dropping important information. CC is an improved model based on BR, and it transfers all the independent classifiers into a chain in a specified order. For each binary model, the results from all the previous classifiers are extended to the feature space as 0/1 attributes. However, only partial label correlation is used for each classifier in this model. Both BR and CC can not scale well to large label set.

Low rank label transformation approaches overcome the flaws of task decomposition. These methods can handle label correlations by restricting the dimension of latent label space. Conditional Principal Label Space Transformation (CPLST) [4] is a representative model using a low rank constraint. CPLST does dimension reduction for labels through a transformation matrix, which is constrained to have lower rank than the original label space. Implemented using SVD decompositon, CPLST has a limitation that it is incapable to handle missing labels since SVD requires a complete label matrix. Besides, it can not deal with large-scale data. Hsiang-Fu proposed a method named LEML [5] in the standard empirical risk minimization framework. It's also subjected to the low rank constraint, and CPLST can be regarded as a special case of LEML when squared-L2 loss is adopted. Although LEML is more general, it can not deal with highly nonlinear data. Even with the help of kernel extension, it's nowhere near as flexible and powerful as blocks in neural network such as relu activation or CNN architecture. Another approach,



(a) Setting 1: hidden missing labels

(b) Setting 2: revealed missing labels

Fig. 1. Two settings for multilabel learning with missing labels

LCML, has demonstrated the ability to handel both label correlations and missing labels [6]. It's built on a probabilistic framework using label transformation concepts, but it also lacks ability to express highly nonlinear space.

To make use of label similarity, Hao-Chen et al. [10] presented a regularization term to alleviate the label incompleteness, and it's based on the smoothness assumption involved with label similarity and instance similarity. Yue et al. [11] formulated a new regularization by forcing the predicted label matrix to maintain the same label correlations as in the training label matrix. While the success of these regularization terms rely on a good approximation to label similarity, which can be hard to learn. Moreover, their proposed methods cannot learn highly nonlinear features either.

3. PROBLEM FORMULATION AND NOTATIONS

Given a training data set $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$ with m training instance-label pairs, in which $x^{(i)} \in \mathbb{R}^d$ is a real value vector representing one instance and $y^{(i)}$ is the corresponding label vector. Specifically, $y^{(i)} \in \{0,1\}^n$, and n is the cardinality of whole label set. For any $j \in [n]$, $y_j^{(i)} = 1$ indicates that the i^{th} sample belongs to the concept of j^{th} label and 0 otherwise. In the rest of this paper, we would use X to denote a batch of input data instances and Y for the corresponding labels. For the special case with missing labels, we consider two possible situations as explained in the Fig. 1. A common setting assumes that the positions of missing labels can be obtained, so in this paper we use another notation mask to record this information. Matrix mask is in the same shape of label matrix Y, in particular, $\mathrm{mask}_j^{(i)} = 0$ means that the label of concept j is missing for the i^{th} sample and $\mathrm{mask}_j^{(i)} = 1$ means the label is not missing (positive or negative). Another setting also appears in practice where the missing labels are mixed up with negative ones. In other words, only positions assigned with value 1 are undoubtedly positive labels while those assigned with value 0 could be either negative or miss-

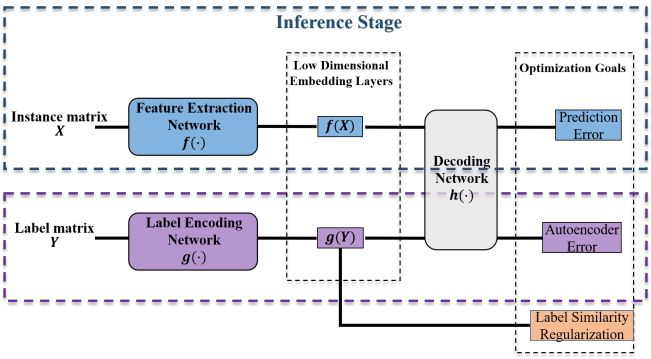


Fig. 2. Maximal correlation embedding network designation

ing labels. In real life, these two settings are both possible, so our work will take both into consideration.

4. THE PROPOSED METHOD

4.1. Maximal correlation embedding network

In the field of natural language processing, embedding is widely used [12]. Through the embedding process, we obtain continuous vectors rather than one-hot representations for each word, and words with similar semantics would be closer to each other. Andrea Frome et al. [13] first integrates the idea of embeddings in NLP to multilabel learning problems. They first get the dense vector embeddings for all candidate labels from a pre-training process with extra corpus and then apply those vectors in subsequent classifiers. This kind of operation extremely relies on the quality of corpus from other sources rather than the intrinsic logic of the task itself. Moreover, it requires huge computing resources.

To overcome these shortcomings, we adopt the setting in NLP that words often appear in the same context (within distance of a certain number of words) should be more similar, and propose a method to learn label embeddings in which labels frequently co-occur within one instance are closer to each other in the embedded label space.

We design the Maximal Correlation Embedding Network (MCEN) to address the multilabel learning problem as shown in Fig. 2. In additional to the classification module shown in the blue dotted box, MCEN integrates an autoencoder module to learn label representations in purple dotted box. The label similarity regularization attached to the label embedding layer is the key component to handle missing labels. The final optimization goals (described in subsection 4.4) include three kind of losses: prediction error, autoencoder error (descried in subsection 4.2) and label similarity regularization (described in subsection 4.3).

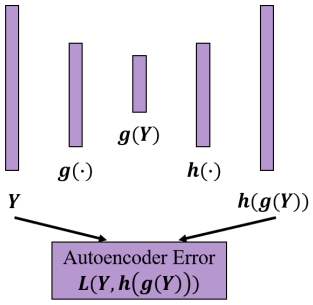


Fig. 3. The architecture of undercomplete autoencoders

4.2. Handling label correlations

Previous works such as [4] attempt to handle label correlation by label matrix decomposition. In particular, for label matrix Y, they consider a low rank transformation matrix V and then minimize the term

$$\left\|Y - YV^TV\right\|_{\mathrm{F}}^2 \tag{1}$$

Instead of doing this way, we adopt the undercomplete autoencoders [14] shown in Fig. 3 to be the label encoding-decoding network. The nature of undercomplete autoencoders, i.e. the latent feature dimension is less than the input dimension, forces the model to capture correlations among labels. Compared with (1), the autoencoder error defined by

$$L(Y, h(g(Y))) = \|Y - h(g(Y))\|_{\mathrm{F}}^2$$

has stronger nonlinear expressiveness because the transformation functions $g(\cdot)$ and $h(\cdot)$ could be any nonlinear functions.

The autoencoder module aims at extracting correlations among labels, while at the same time we should connect it with the classification module. The idea is to share the parameters in the decoding network of both modules. The feature extraction network can adopt any network structure for inputs such as CNN for images and so on. The latent features $f(X)$ must have the same dimension with label embedding vectors $g(Y)$ to make it possible to share the decoding function $h(\cdot)$. Finally we can get the logits $h(f(X))$ for classification. For a particular instance $x^{(i)}$, the corresponding prediction for the j^{th} label concept is denoted as

$$\hat{y}_j^{(i)} = \frac{1}{1 + e^{-[h(f(x^{(i)}))]_j}}$$

where $[h(f(x^{(i)}))]_j$ is the j^{th} component of the n dimensional logits.

Rather than the softmax loss function for multi-class classification, we use the summation of sigmoid-cross-entropy loss for each label as the prediction error

$$L(Y, h(f(X))) = -\sum_{j=1}^{n}\sum_{i=1}^{m} l(y_j^{(i)}, \hat{y}_j^{(i)})$$

where $l(y_j^{(i)}, \hat{y}_j^{(i)})$ is the log-loss between prediction $\hat{y}_j^{(i)}$ and ground truth $y_j^{(i)}$, and n, m is the number of labels and instances respectively.

4.3. Handling missing labels

As we mentioned before, extracting the information of label similarity can mitigate the effects caused by missing labels. We proposed a novel label similarity regularization term for the label embedding vectors. The main idea is to maximize total correlation among embedding vectors of labels, and it's proved to be equivalent to force the correlations between labels proportional to their co-occur counts within the same instance. A well-known correlation measure is the HGR (Hirschfeld-Gebelein-Rényi) maximal correlation:

$$\max_{\substack{\mathbb{E}[f]=0,\mathrm{Cov}[f]=I\\ \mathbb{E}[g]=0,\mathrm{Cov}[g]=I}} \mathbb{E}[f^{\mathrm{T}}(X)g(Y)]$$

It can find highly related features between two random variables X and Y.

To generalize the idea of maximal correlation to multiple random variables, Huang et al. [15] proposed the Generalized Maximal Correlation (GMC). For jointly distributed random variables $Y_1, Y_2 \dots Y_n$, the definition of GMC is defined as:

$$\begin{aligned} \underset{g_1,g_2,\dots g_n}{\text{maximize}} \quad & \mathbb{E}\left[\sum_{i\neq j}(g_i(Y_i))^{\mathrm{T}}(g_j(Y_j))\right] \\ \text{subject to} \quad & \mathbb{E}[g_i(Y_i)] = 0,\ i = 1,\dots,n. \\ & \mathbb{E}\left[\sum_{i=1}^{n}\|g_i(Y_i)\|^2\right] = 1 \end{aligned}$$

However, solving the problem with constraints is rather difficult. So, in our paper we transform the original GMC into a relaxed unconstrained optimization problem:

$$\underset{g_1,g_2,\dots g_n}{\text{maximize}} \quad \mathbb{E}\left[\sum_{i\neq j}(g_i(Y_i))^{\mathrm{T}}(g_j(Y_j))\right] - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\|g_i(Y_i)\|^2\right] \tag{2}$$

In our multilabel classification problem, we use n binary random variables $Y_1, Y_2 \dots Y_n$ to represent each label component in a multilabel random vector $\vec{Y}$. Through the label encoding network $g(\cdot)$, the representation $g(\vec{Y})$ for each label vector $\vec{Y} \triangleq (Y_1, Y_2 \dots Y_n)$ could be derived from a linear combination of embedding components corresponding with positive labels. It's defined as

$$g(\vec{Y}) = \sum_{j=1}^{n} g_i(Y_i), \qquad g_i(Y_i) = \begin{cases} v_i & Y_i = 1 \\ 0 & Y_i = 0 \end{cases}$$

where $g_i(Y_i)$ is the function of component variable Y_i and v_i stands for the embedding vector for positive value.

We rewrite the optimization objective in (2) as a label similarity regularization $\Omega(Y)$, derivation details may be found in Appendix A in the supplemental material.

$$\Omega(Y) = -\frac{1}{m}\left(\sum_{i=1}^{n}\sum_{j\neq i} S_{i,j}\cdot v_i^{\mathrm{T}}v_j - \frac{1}{2}\sum_{i=1}^{n} S_{i,i}\|v_i\|^2\right)$$

in which $S_{i,j}$ indicates how many instances have both positive label i and positive label j:

$$S_{i,j} = \sum_{k=1}^{m} y_i^{(k)} \cdot y_j^{(k)}$$

The practical purpose of $\Omega(Y)$ is rather straightforward. It guarantees that the relative magnitude of $v_i^{\mathrm{T}}v_j$ is proportional to the coexistence counts between label i and label j. In other words, labels that frequently co-occur within the same instance will be closer in the embedded label space.

Guided by the theoretical proof in [15], our regularization is actually finding the optimal feature functions which could maximize the total correlation among multiple labels. What'more, in the cases with missing labels, the co-occurrence counts will fall down proportionately so that this regular term still works.

4.4. Optimization goals for two missing label settings

4.4.1. Setting 1: hidden missing labels

In this situation as in Fig. 1 (a), the missing labels are mixed up with negative labels. The final optimization goal is to minimize all the three losses we mentioned above. Considering that different losses have different scales so we use two hyper-parameters α and β to weight autoencoder error and label similarity regularization respectively. Then the final loss function is

$$\text{Loss}_1 = L(Y, h(f(X))) + \alpha \cdot L(Y, h(g(Y))) + \beta \cdot \Omega(Y)$$

4.4.2. Setting 2: revealed missing labels

Recall that the matrix mask records the information of missing positions as in Fig. 1 (b), we can use it to eliminate the corresponding loss in missing positions. Under this setting, the prediction error and autoencoder error turn into

$$\widetilde{L}(Y, h(f(X)), \text{mask}) = -\sum_{j=1}^{n}\sum_{i=1}^{m} \text{mask}_j^{(i)} \cdot l(y_j^{(i)}, \hat{y}_j^{(i)})$$

$$\widetilde{L}(Y, h(f(Y)), \text{mask}) = \|(Y - h(g(Y))) \circ \text{mask}\|^2$$

where $\circ$ means element-wise product operation. Consequently, the final loss becomes

$$\begin{aligned} \text{Loss}_2 = {} & \widetilde{L}(Y, h(f(X)), \text{mask}) + \\ & \alpha \cdot \widetilde{L}(Y, h(f(Y)), \text{mask}) + \beta \cdot \Omega(Y) \end{aligned}$$

5. EXPERIMENTS

Table 1 shows all the data sets used in our experiments where APL stands for Average Positive Labels in a sample. These data sets include images, texts and medical data and the label

Table 1. Data sets description

Dataset	Instance		Feature	Labels	APL	Type
	Train	Test				
corel5k	4500	499	1000	260	3.4	image
espgame	18689	2081		268	4.7	
iaprtc12	17665	1962		291	5.7	
pascal07	5011	4952		20	1.5	
mirflicker	12500	12500		38	4.7	
bibtex	4880	2515	1836	159	2.4	text
delicious	12920	3185	500	983	19.0	
nuswide	161789	107859	1134	1000	5.8	image
yeast	1500	500	103	14	4.2	medical

Table 2. F1-score performance on complete labels

Dataset	MCEN		LCML	CPLST
	$\alpha, \beta = 0$	$\alpha, \beta \neq 0$		
corel5k	0.2526	**0.2752**	0.2071	0.1980
espgame	0.2603	**0.2682**	0.2273	0.2219
iaprtc12	0.3524	**0.3541**	0.2405	0.2304
pascal07	0.3707	**0.3945**	0.3493	0.3181
mirflickr	0.5082	**0.5083**	0.4665	0.4608

Table 3. Top-3 precision and AUC on complete labels

Dataset	MCEN		LEML	
	Top-3	AUC	Top-3	AUC
bibtex	35.94	0.8870	**36.53**	**0.9015**
delicious	**61.82**	**0.9038**	61.23	0.8827
nuswide	**18.29**	**0.8081**	16.00	0.7718

cardinality varies in a wide range. For the first five image data sets, we use their 1000 dimensional SIFT features[1], and the remaining features are from a open source multilabel learning library called Mulan[2].

To demonstrate the effectiveness of our proposed method, MCEN will be compared with a traditional method CPLST [4] and the state of the art methods, including LCML [6], LEML [5] and SSWL [10].

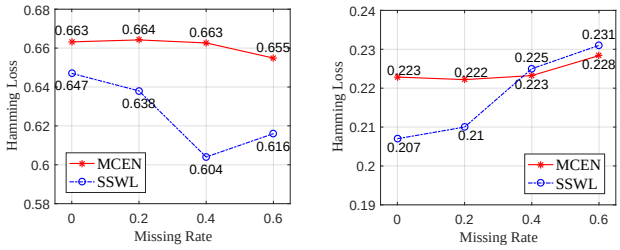
Several popular evaluation metrics for multilabel classification are used in our paper: F1-score, top-k precision, average Area Under Curve (AUC) and hamming loss. See Appendix B in the supplemental material for details.

Experimental Setup. We use two hidden layers with relu activation and batch normalization for the feature extraction network and label encoding network. We adjust the dimension of embedding layer to be smaller than the cardinality of labels in different data sets. Grid search strategy is used to select optimal values for hyper parameters α and β in the final losses. We minimize the loss using ADAM optimizer. To simulate cases with missing labels, we randomly drop out some labels and observe the performance by varying the missing rate.

[1]http://lear.inrialpes.fr/people/guillaumin/data.php
[2]http://mulan.sourceforge.net/datasets-mlc.html

Table 4. F1-score performance on 50% missing labels

Dataset	MCEN		LCML
	$\alpha, \beta = 0$	$\alpha, \beta \neq 0$	
corel5k	0.2508	**0.2770**	0.2380
espgame	0.2359	**0.2448**	0.2212
iaprtc12	0.3216	**0.3283**	0.2309
pascal07	0.3160	0.3283	**0.3426**
mirflickr	0.4823	**0.4857**	0.4367



(a) F1-score (b) Hamming Loss

Fig. 4. Performance tendency on dataset yeast

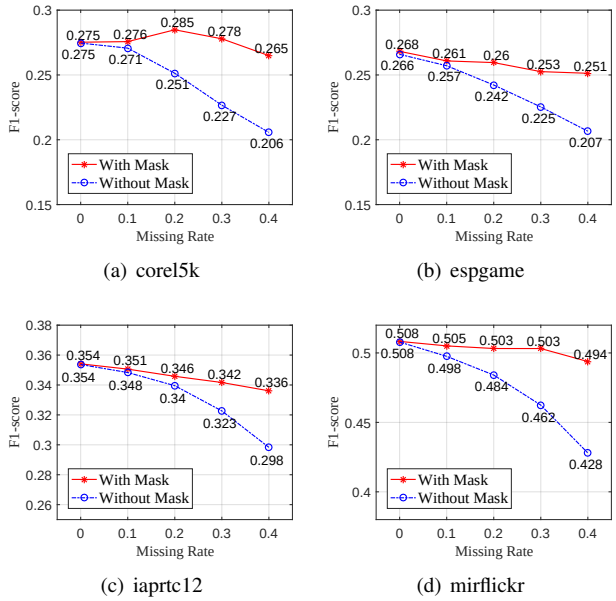
5.1. Performances on complete labels

We first performed experiments on data sets with full labels. Table 2 shows the compared results between our method (MCEN) with LCML and CPLST under F1-score metric and table 3 shows the comparision of MCEN and LEML with top-3 precision and AUC for evaluation. The results across different data sets and different evaluation criteria demonstrate that our proposed approach outperforms all the competing methods, especially it improves the F1-score by **24%** than the second best method on average. Table 2 also proves that the performance will improve by using autoencoder error and label similarity regularization ($\alpha, \beta \neq 0$).

5.2. Performances on missing labels

We randomly drop a percentage of the labels to simulate the situation with missing labels. First, we consider the setting with revealed missing labels. Table 4 shows the F1-score on 50% missing labels (every label has 50chance of missing). The results prove that our method performs better even in missing label data. On Corel5k, it even outperforms the complete label case. One possible reason is that the number of training examples is so small that complete labels lead to over-fitting. The method SSWL studied the performance by varying the missing rate, compared with it, our method shows excellent stability when dealing with large missing rate, as shown in Fig.4. Using MCEN, F1-score drops **3.5%** lower and hamming loss increases **9.1%** lower than SSWL when missing rate climbed to 0.6.

For the setting with hidden missing labels, all competing methods are not applicable. so we conducted a series of experiments using MCEN to observe the difference between



(a) corel5k (b) espgame

(c) iaprtc12 (d) mirflickr

Fig. 5. Performance tendency under two settings

two settings mentioned in subsection 4.4. For convenience, we will use "without mask" to denote the first setting and "with mask" to denote the second one since we use a mask to adjust the final loss in the second setting. The result is shown in Fig.5. It shows that the performance with hidden missing labels is acceptable when the missing rate is less than 0.2 while the information of missing positions can make the model more robust to missing rate.

6. CONCLUSION

In this paper, we studied the multilabel learning problem especially for cases with missing labels. To the best of our knowledge, we proposed the first end-to-end architecture to address this problem. We designed a maximal correlation embedding network which integrates a undercomplete autoencoder and a novel label similarity regularization to handle label correlation and missing labels. Our experiments demonstrate that our method outperforms the state of the art both in data sets with complete labels and missing labels.

ACKNOWLEDGMENTS

The research was funded by the Natural Science Foundation of China 61807021, Shenzhen Science and Technology Research and Development Funds (JCYJ20170818094022586, GJHZ20170314112258560) and Innovation and entrepreneurship project for overseas high-level talents of Shenzhen (KQJSCX2018032714403783).

7. REFERENCES

[1] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas, "Mining multi-label data," in *Data mining and knowledge discovery handbook*, pp. 667–685. Springer, 2009.

[2] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, pp. 333, 2011.

[3] Farbound Tai and Hsuan-Tien Lin, "Multilabel classification with principal label space transformation," *Neural Computation*, vol. 24, no. 9, pp. 2508–2542, 2012.

[4] Yao-Nan Chen and Hsuan-Tien Lin, "Feature-aware label space dimension reduction for multi-label classification," in *Advances in Neural Information Processing Systems*, 2012, pp. 1529–1537.

[5] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon, "Large-scale multi-label learning with missing labels," in *International conference on machine learning*, 2014, pp. 593–601.

[6] Wei Bi and James T Kwok, "Multilabel classification with label correlations and missing labels.," in *AAAI*, 2014, pp. 1680–1686.

[7] Xiaojin Zhu and Andrew B Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[8] Zijia Lin, Guiguang Ding, Mingqing Hu, Jianmin Wang, and Xiaojun Ye, "Image tag completion via image-specific and tag-specific linear sparse reconstructions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1618–1625.

[9] Xiangnan Kong, Michael K Ng, and Zhi-Hua Zhou, "Transductive multilabel learning via label set propagation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 3, pp. 704–719, 2013.

[10] Hao-Chen Dong, Yu-Feng Li, and Zhi-Hua Zhou, "Learning from semi-supervised weak-label data," AAAI, 2018.

[11] Yue Zhu, James T Kwok, and Zhi-Hua Zhou, "Multi-label learning with global and local label correlation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1081–1094, 2018.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al., "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, 2013, pp. 2121–2129.

[14] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio, *Deep learning*, vol. 1, MIT press Cambridge, 2016.

[15] Shao-Lun Huang, Lin Zhang, and Lizhong Zheng, "An information-theoretic approach to unsupervised feature selection for high-dimensional data," in *Information Theory Workshop (ITW), 2017 IEEE*. IEEE, 2017, pp. 434–438.