Discovering Unknown Labels for Multi-Label Image Classification

Jun Huang
Anhui University of Technology
Yu Yan (1927271955@qq.com)
Anhui University of Technology
Xiao Zheng
Anhui University of Technology

Research Article

Keywords: multi-label learning, unknown labels, image classification

Posted Date: July 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1813664/v1

Discovering Unknown Labels for Multi-Label Image Classification

Jun Huang[1,2], Yu Yan[1*] and Xiao Zheng[1,2]

[1]School of Computer Science and Technology, Anhui University of Technology, Maanshan, 243032, Anhui, China.
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, 230088, Anhui, Country.

*Corresponding author(s). E-mail(s): 1927271955@qq.com;
Contributing authors: huangjun.cs@ahut.edu.cn;
xzheng@ahut.edu.cn;

Abstract

A multi-label learning (MLL) method can simultaneously process the instances with multiple labels, and many well-known methods have been proposed to solve various MLL-related problems. The existing MLL methods are mainly applied under the assumption of a fixed label set, i.e., the class labels are all observed for the training data. However, in many real-world applications, there may be some unknown labels outside of this set, especially for large-scale and complex datasets. In this paper, a multi-label classification model based on deep learning is proposed to discover the unknown labels for multi-label image classification. It can simultaneously predict known and unknown labels for unseen images. Besides, an attention mechanism is introduced into the model, where the attention maps of unknown labels can be used to observe the corresponding objects of an image and to get the semantic information of these unknown labels.

Keywords: multi-label learning, unknown labels, image classification

1 Introduction

The task of MLL is mainly to learn an efficient classification model from the training data in order to predict one or multiple possible labels for new data

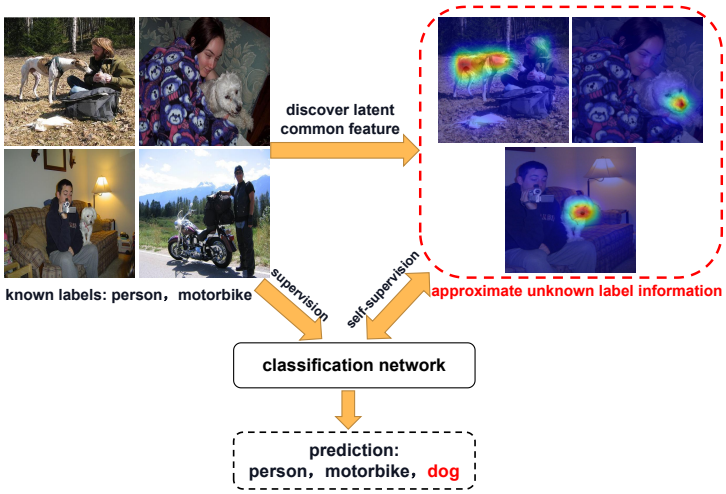

Fig. 1 The train set in the upper left corner has the known labels of "person" and "motorbike", used as the classification model's supervision. Moreover, the latent common features of clustering are obtained from the training set and then obtaining the approximate unknown label information. The attention map presents the objects corresponding to latent common features is in the upper right. The approximately unknown label information is used as self-supervision of classification network to guide network to learn unknown labels. After learning, the prediction of classification network contains the prediction value of unknown label "dog." Finally, the semantic information of unknown labels is obtained by observing the attention map.

samples. In the past decades, many methods have been proposed for MLL from different aspects, such as multi-view MLL [1, 2], partial label [3, 4], missing label [5–18]. For the training data, existing works mainly assume that all the labels are observed and known. However, in some applications, this assumption is not always true. Users only see part of the labels, and some unknown labels are hidden in the data. In addition, their semantic information and labelled results of samples are also all unknown for the training data. There are many possible reasons for this problem [19, 20], such as the high cost of labelling process, complexity of label semantics, limitation of human knowledge and data acquisition conditions.

Many kinds of researches have been proposed in the past few decades to solve the problem of new labels, such as online learning [21–25], class incremental learning [26–29], zero-shot learning [30–33], Generalized out-of-distribution detection [34–45], etc. In the zero-shot learning problem, the information of new labels is usually known during the training stage, including the semantic information and their total numbers. In the generalized out-of-distribution problem, unknown labels may only appear in the test stage. However, as shown in Figure 1, this paper aims to study the case with completely unknown labels, including the semantic information of labels and their labelling results of training samples. In addition, unknown labels may appear in both training and test stages.

Researchers have defined the learning with unknown labels as a new learning framework. Zhao et al. [46] proposed an ExML method in which unknown labels are recognized and found by adding new features space. However, this method is mainly designed to handle multi-class classification and can not be directly applied in MLL. Pham et al. [19] proposed a discriminatory probability model MIMLNC and classified all new instances as a new label. In their work, it is reported that discovering unknown labels hiding in the data can not only find interesting knowledge but also improve the performance of known labels. Zhou et al. [47] proposed DMNL methods. This method optimized a supervised loss and an unsupervised clustering regular item to model the classifier of known labels and to discover k latent new labels. Similarly, Huang et al. [48] proposed a DLCL method to discover unknown labels in the single-instance data. They used non-negative matrix decomposition (NMF) [49] to decompose the feature matrix into a coefficient matrix and a complete approximate label representation. At the same time, they built a classification model from the feature space to the complete label space. This method can improve the performance of known labels when unknown labels information discovered.

Benefited from the great success of deep neural networks in recent years, great efforts have been devoted to develop deep-based models to address multi-label image classification. Besides, deep models can extract high-order semantic features to obtain more excellent results than shallow models. However, the current methods to deal with the existence of unknown labels are based on shallow models. In this paper, we propose a deep model to handle MLL with the existence of unknown labels in the training data. The contributions of this paper are summarized as follows.

(a) We propose a deep model based on neural networks for multi-label image classification with unknown labels, and it can predict known and unknown labels simultaneously for unseen images. Moreover, the unknown label detection module can be applied in any existing deep classification model.
(b) We use unknown label information obtained to estimate the complete label relationships and to further improve the performance of known labels.
(c) We use attention maps to observe the regions corresponding to the unknown labels of an image and obtain the semantic information of unknown labels.

2 Related Work

Many problems are similar to our research in MLL, such as missing label, zero-shot learning, generalized out-of-distribution detection, online learning and class-incremental learning, and Figure 2 shows the differences between them according the label matrix.

Missing label learning: In missing label learning, the label set is a known closed set, but part of the labelling results of some examples are missing. Researchers have proposed many solutions from different aspects, for example, the missing label data is supplemented or restored first, and then a classifier is

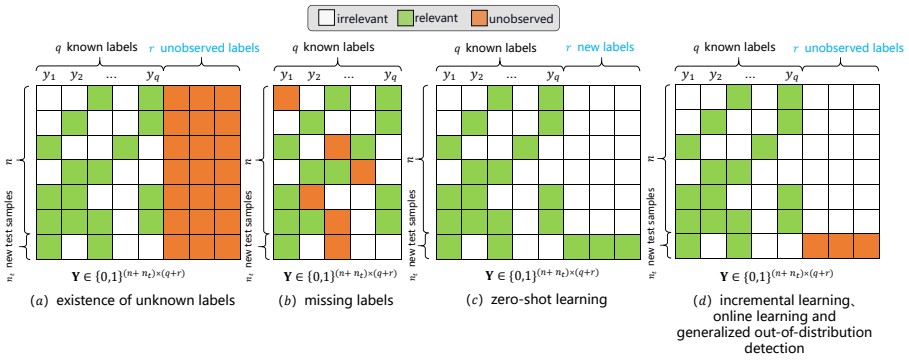

Fig. 2 The label matrices of four problems are as follows. (a) The problem related to the existence of unknown labels. There exist r kinds of completely unknown labels. (b) The problem related to the missing labels. The labels are partly missing rather than completely missing. (c) The problem related to zero-shot learning. There are new labels in the test set, and the semantic description or attribute matrix is known. (d) The problem related to generalized out-of-distribution detection. There are new labels in the test set.

constructed based on the restored labelling result. However, the restored label can not effectively guide the classification model in such a two-stage recovery strategy. According to the previous researches, some methods can reduce the impact of missing labels by creating an observation matrix [5, 7–13] to indicate the missing values. In practice, however, it is difficult to acquire this prior knowledge in advance. In addition, some other methods [14–18] try to recover label matrix by matrix completion, matrix decomposition, and label reconstruction. The existing missing label learning algorithms can be feasible when each label has at least one positive data sample. If the labelling results of one class label are completely missing for all the samples, we will not obtain a satisfactory result. However, in our problem, some of the labels are missing entirely, and they have no labelled samples. Therefore, the missing labels methods can not be directly applied to our problem.

Zero-shot learning: Zero-shot learning can mainly solve the existence of unknown new labels during the test stage. In recent years, zero-shot learning has attracted considerable attention from researchers, at the same time, some effective learning methods [30–33] have been proposed. Zero-shot learning often depends on attribute matrix or label semantic descriptions, and learns the similarity between semantic description and image features to classify the unknown labels. Furthermore, in zero-shot learning, the attribute matrix or the unknown label semantic descriptions are taken as a prior knowledge.

Generalized out-of-distribution detection: Yang et al. [34] proposed a concept called generalized out-of-distribution detection. It contains five problems which are anomaly detection (AD) [36–39], novelty detection (ND) [40–42], open set recognition (OSR) [43–45] and outlier detection (OD) [35]. These problems have similar motivation and methodology. Anomaly detection (AD) aims to detect any anomalous sample which is deviated from the predefined normality during testing [34]. It just judge whether a test sample is

"normal" or "abnormal", and there is only one class in the train set (data in this class is regarded as "normal"). ND is similar to AD. It classify a test samples as "known class" or "novel class", but there may be multiple classes in train set (all classes in train set is regarded as "normal" in test set). OSR need to classify a test samples as "known class" or "novel class", and it also need to construct multi-class classifier to judge what known classes accurately. OD has similar purpose to the above problems. It aims to detect test samples with non-overlapping labels w.r.t training data. OD is a super-category that includes semantic AD, one-class ND, multi-class ND, and open-set recognition [34]. Outlier detection aims to detect samples that are labelled different from the others in the given observation set. We don't considered OD in this paper, because OD does not follow a train-test scheme. Except for OD, the above problems are all trained on labelled train data, known as in-distribution. And they are tested on unlabelled data that some samples contain unknown labels, known as out-of-distribution. In our study, not only do unknown labels appear in the test stage, but also appear in the train stage. So if the above methods are directly used in our problem, unknown labels in train set cannot be fully utilized.

Online learning and class-incremental learning: Online learning is defined as follows. The training data is input in a sequence, and the model parameters are updated online to provide better results for new test samples after each iteration [21–25]. Class-incremental learning can be used to segment training data into sequences according to tasks, process the tasks in order, incrementally learn a classification model, and then effectively classify any new labels in task sequence [26–29]. In class-incremental learning and online learning tasks, new samples might continually introduce new labels. These two problems are similar to novelty detection and open set recognition. And They need to use novel samples for training after detect them. Similarly, regarding class-incremental learning and online learning tasks, if there are unknown labels in training samples at the initial stage of training, the current methods cannot classify for unknown labels in the subsequent training process. Therefore, the methods of online learning and class-incremental learning can not be directly applied to our problem.

In our research, the problem of the existence of unknown labels in MLL does not contain any additional prior knowledge of unknown labels, and both training and test samples may exist unknown labels. Therefore, Current learning methods based on online learning, class-incremental learning, zero-shot learning and generalized OOD detection can't be directly used to solve the unknown labels problems in our study.

In this paper, we propose a deep model to deal with the existence of unknown labels, and it can predict unknown and known labels simultaneously. Finally, we use attention maps to obtain the semantic information of unknown labels.

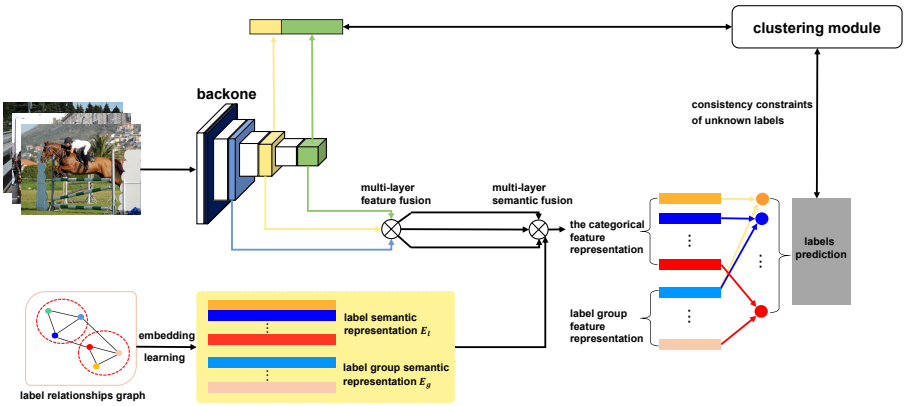

Fig. 3 Overall architecture of our model. Given an image, a CNN network outputs the image features from different layers to different branches. We fuse the features of different branches, and then fuse the semantic information obtained by the graph neural network with the image features as the final feature representation. At the same time, we cluster the image features to obtain approximate unknown label information to guide the classification network to learn unknown labels, and use classifier to classify known and unknown labels. In addition, we use unknown label information to complete the label relationships.

3 Proposed method

3.1 Problem Definition

The image dataset is $\mathbf{I} \in \mathcal{R}^{n\times c\times w\times h}$, where n indicates the number of samples, numbers c, w, and h are the sizes of channels, width, and height, respectively. The complete label matrix is $\widehat{\mathbf{Y}} = \left[\mathbf{Y}, \overline{\mathbf{Y}}\right] \in \{0,1\}^{n\times l}$, the known label matrix is $\mathbf{Y} \in \{0,1\}^{n\times q}$, and the unknown label matrix is $\overline{\mathbf{Y}} \in \{0,1\}^{n\times r}$. Here, the number of unknown labels is set to r. Accordingly, the number of complete labels is $l = q+r$. If label j appears in the ith image, then $y_{ij} = 1$, and $y_{ij} = 0$ otherwise. The complete label relationship graph is expressed as $\widehat{\mathbf{O}} = \left\{\widehat{\mathbf{D}}, \widehat{\mathbf{S}}\right\}$, where $\widehat{\mathbf{D}}$ represents the vertex (label) set and $\widehat{\mathbf{S}}$ is the edge set. The vertex feature is the semantic description of complete labels $\widehat{\mathbf{V}} = \left[\mathbf{V}, \overline{\mathbf{V}}\right] \in \mathcal{R}^{l\times d}$, where $\mathbf{V} = [V_1,...,V_q] \in \mathcal{R}^{q\times d}$ is the semantic description of known labels and $\overline{\mathbf{V}} = \left[\overline{V}_1,...,\overline{V}_r\right] \in \mathcal{R}^{r\times d}$ is the semantic description of unknown labels. Furthermore, the weight of each edge is the co-occurrence probability of complete labels $\widehat{\mathbf{A}} \in [0,1]^{l\times l}$. Similarly, $\mathbf{A} \in [0,1]^{q\times q}$ is the co-occurrence probability matrix of known labels.

The task of this paper is to build a deep model f_θ which can map all the data samples in the image set $\mathbf{I}$ into the l complete label space, i.e., $f_\theta\left(\mathbf{I}, \widehat{\mathbf{O}}\right) = \widehat{\mathbf{Y}}$, where θ is the model parameters. Figure 3 shows the overall framework of our proposed method which is mainly composed of three parts, i.e., initialization of the complete label relationship graph, constructing the approximate unknown label matrix, and building a multi-label image classification model. In the following sections, we will introduce them in detail.

3.2 Initialization of the complete label relationship graph

For the proposed method, we first need to get a complete label graph $\widehat{\mathbf{O}}$ which contains the complete label semantic descriptions $\widehat{\mathbf{V}}$ and the label co-occurrence probability matrix $\widehat{\mathbf{A}}$. In addition to expanding the size of $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{A}}$, we also need to initialize the contents of the expanded part of $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{A}}$. In our problem, no information is provided for the unknown labels. However, there are more or less certain correlations between the unknown and known labels. Therefore, we can use the known label relationships to initialize the unknown labels relationships, i.e. unknown label semantic descriptions and unknown label co-occurrence probability.

First, we use known label semantic descriptions to initialize the unknown labels semantic descriptions. The process is conducted as

$$\overline{V}_k = \frac{1}{q}\sum_{i=1}^{q} V_i, \tag{1}$$

where $V_i \in \mathcal{R}^{1\times d}$ represents the ith known label semantic description and $\overline{V}_k \in \mathcal{R}^{1\times d}$ represents the kth unknown label semantic description.

Second, in order to obtain $\widehat{\mathbf{A}}$, we need to initialize the co-occurrence times between the ith known label and the kth unknown label, i.e., L_{ki} and L_{ik}, and the occurrence times of kth unknown labels, i.e., N_k, and $1 \le i \le q$, $q+1 \le k \le l$. Specifically, we calculate the average co-occurrence times between all known labels with a random value as the co-occurrence times between the kth unknown label and ith known label. The process is conducted as follows:

$$L_{ki} = \max\left(\frac{1}{q}\sum_{j=1}^{q} L_{ji} + randint(-\alpha \times \frac{1}{q}\sum_{j=1}^{q} L_{ji}, \alpha \times \frac{1}{q}\sum_{j=1}^{q} L_{ji}), 0\right) \tag{2}$$

$$N_k = \max\left(\frac{1}{q}\sum_{j=1}^{q} N_j + randint(-\alpha \times \frac{1}{q}\sum_{j=1}^{q} N_j, \alpha \times \frac{1}{q}\sum_{j=1}^{q} N_j), 0\right) \tag{3}$$

where α is a constant, the function $randint(a, b)$ randomly generate a integer between a and b, L_{ji} represents the co-occurrence times between the ith and jth known labels, and the function $max(a, b)$ returns the maximum value between a and b.

After the initialization of L_{ki}, we can initialize the co-occurrence times between the kth unknown label and the hth unknown label., $q+1 \le k \le l$, $q+1 \le h < k$. The process is as follow:

$$L_{kh} = \max\left(\frac{1}{q}\sum_{j=1}^{q} L_{kj} + randint(-\alpha \times \frac{1}{q}\sum_{j=1}^{q} L_{kj}, \alpha \times \frac{1}{q}\sum_{j=1}^{q} L_{kj}), 0\right) \tag{4}$$

$$L_{kk} = 0 \tag{5}$$

In order to facilitate a better optimization, we symmetrize the co-occurrence times L_{ij} by $L_{ij} = L_{ji}$, where $1 \leq i \leq l$, $1 \leq j \leq l$.

Following chen et al. [50], we can estimate $\widehat{\mathbf{A}}_{ki}$ with the label co-occurrence times L_{ki} and label occurrence times N_k, i.e., $\widehat{\mathbf{A}}_{ki} = L_{ki}/N_k$, where $q+1 \leq k \leq l$ and $1 \leq i \leq l$. And the chen et al. use the threshold τ to filter noisy edges, the process is as follow:

$$\widehat{A}_{ki} = \begin{cases} 0, \widehat{A}_{ki} < \gamma \\ 1, \widehat{A}_{ki} \geq \gamma \end{cases} \tag{6}$$

where γ is a threhold, and it is set to 0.4 following chen et al. To alleviate over-smoothing problem, Chen et al. propose the following re-weighted scheme, the process is as follow:

$$\widehat{A}_{ki} = \begin{cases} p/\sum_{j=1,i\neq j}^{l} \widehat{A}_{ki}, i \neq j \\ 1-p, i = j \end{cases} \tag{7}$$

where p is a hyper-parameter, and it is set to 0.2 following chen et al. Similarly, the solution for $\widehat{\mathbf{A}}_{ik}$ is the same as $\widehat{\mathbf{A}}_{ki}$.

3.3 Constructing the approximate unknown label matrix

After obtaining the complete label relationship graph, we need to construct an approximate unknown label matrix, and then use it to construct the classification model. First, we fuse the feature maps from different layers of a deep convolution neural network. Then, we adopt the deep embedding clustering (DEC) [51] method to obtain approximate unknown label cluster centers and an approximate unknown label matrix.

3.3.1 Clustering Feature Fusion

In a deep convolution neural network, feature map from the low layers focus more on image texture features. It has a high resolution, clear image details, and unknown label information, but it also has a lot noise. On the contrary, the feature map from deep layers can abstractly represent the whole image with less noise but low resolution and less unknown label information. To discover unknown labels clearly, we fuse feature maps from multiple layers.

Taking ResNet-101, which consists of 4 residual blocks, as the backbone. The last convolution layers of the latter two residual blocks of ResNet-101 are used to extract the feature maps: $F = \{f_s\}_{s=1}^{B}$, $f_s \in \mathcal{R}^{c_s \times w_s \times h_s}$, where $B = 3$, c_s, w_s and h_s denote the number of channels, width, and height of the feature map in different layers. f_2 represents the feature map of penultimate block. We use a convolution layer with 1×1 kernel and interpolation with a billinear mode to project feature maps from f_3 to a new $\bar{f}_3$. By concatenating $\bar{f}_3$ and f_2,

we can obtain the feature vector $f_c \in \mathcal{R}^{c_2+c_3}$ through a global average pooling layer, the processes are defined as

$$\bar{f}_3 = \text{interpolate}\left(f_3, size = (w_2, h_2)\right) \in \mathcal{R}^{c_3 \times w_2 \times h_2} \tag{8}$$

$$f_c = \text{avg\_pool}(\text{concat}(f_2, \bar{f}_3)) \in \mathcal{R}^{c_2+c_3} \tag{9}$$

where interpolate$(\cdot, size)$ function up samples the input to the given size. The concat$(\cdot, \cdot)$ function concatenate different feature maps on the channel dimension. The avg_pool$(\cdot)$ function applies a adaptive average pooling over input feature maps.

The feature vectors f_c of all samples constitute the feature matrix. $\mathbf{X} \in \mathcal{R}^{n \times (c_1+c_2)}$. Based on the feature matrix $\mathbf{X}$, we can adopt the DEC [51] method to obtain an approximate label matrix of all training samples.

3.3.2 Approximate Unknown Label Cluster Centers

To use DEC, we first need to initialize the cluster centers for it. Here we adopt the k-means algorithm, and $\widehat{g} \in \mathcal{R}^{l \times (c_2+c_3)}$ are the cluster centers for the labels, including the known and unknown ones. $\widehat{\mathbf{Q}} \in [0,1]^{n \times l}$ is the soft cluster assignments, and each element $\widehat{Q}_{ij}$ indicates the probability that the ith sample belongs to the jth cluster center. In this paper, following the student's t-distribution [52], each element $\widehat{Q}_{ij}$ is calculated according to

$$\widehat{Q}_{ij} = \frac{\left(1 + \left\|X_i - \widehat{g}_j\right\|^2 / \eta\right)^{-\frac{\eta+1}{2}}}{\sum_{j'} \left(1 + \left\|X_i - \widehat{g}_{j'}\right\|^2 / \eta\right)^{-\frac{\eta+1}{2}}}, \tag{10}$$

where X_i is the feature vector for the ith image, and $\widehat{g}_j$ is cluster center vector for the jth class, and η is the degrees of freedom of the student's t-distribution. Following DEC [51], η is set to 1 in this paper for all experiments.

In MLL, each image may have multiple class labels. Based on $\widehat{\mathbf{Q}}$, we can obtain a hard assignment matrix $\widehat{\mathbf{G}} \in \{0,1\}^{n \times l}$ with a threshold $\tau \in [0,1]$. $\widehat{\mathbf{G}}$ can be taken as an approximate complete label matrix, including the known and unknown labels, and it is more reasonable to MLL problem.

$$\widehat{G}_{i,j} = \begin{cases} 1, \widehat{Q}_{i,j} > \tau \\ 0, \widehat{Q}_{i,j} \leq \tau \end{cases} \tag{11}$$

Then we split the cluster centers into two parts for known and unknown labels respectively. Based on the previous research [47], we figure out the best match between the columns of assignments $\widehat{\mathbf{G}}$ and the known label matrix $\mathbf{Y}$ by maximizing the F1 measure according to

$$\arg\max_j \left\{ F\left(\widehat{G}_{:,i}\ ,\ Y_{:,j}\right), i \in \{1, ..., l\}, j \in \{1, ..., q\} \right\}, \tag{12}$$

where $\widehat{G}_{:,i}$ represents the ith column of $\widehat{\mathbf{G}}$, and $Y_{:,j}$ represents the jth column of known label matrix $\mathbf{Y}$. According equation (12), we can obtain the q best matched cluster centers $g \in \mathcal{R}^{q\times(c_2+c_3)}$ for known labels, and the rest ones $\overline{g} \in \mathcal{R}^{r\times(c_2+c_3)}$ are taken as the cluster centers for r unknown labels, i.e., $\widehat{g} = [g, \overline{g}] \in \mathcal{R}^{l\times(c_2+c_3)}$. Correspondingly, $\widehat{\mathbf{Q}} = \left[\mathbf{Q}, \overline{\mathbf{Q}}\right] \in [0,1]^{n\times l}$, where $\mathbf{Q} \in [0,1]^{n\times q}$ is soft assignments for the known labels and $\overline{\mathbf{Q}} \in [0,1]^{n\times r}$ is soft assignments for the unknown labels.

3.3.3 Deep Embedding Clustering

After the initialization for the cluster centers, we adopt DEC [51] to optimize the cluster centers $\widehat{g}$ from current high confidence assignments $\widehat{\mathbf{Q}}$ using an auxiliary target distribution $\widehat{\mathbf{P}} = \left[\mathbf{P}, \overline{\mathbf{P}}\right] \in [0,1]^{n\times l}$, where $\mathbf{P} \in [0,1]^{n\times q}$ is target distribution for the known labels and $\overline{\mathbf{P}} \in [0,1]^{n\times r}$ is target distribution for the unknown labels. The target distribution $\widehat{\mathbf{P}}$ is defined as

$$\bar{P}_{ij} = \frac{\frac{\widehat{Q}_{ij}^2}{Z_j}}{\sum_{j'} \frac{\widehat{Q}_{ij'}^2}{Z_{j'}}}, \tag{13}$$

where $Z_j = \sum_i \widehat{Q}_{ij}$. It is used to normalize the soft assignments and prevent large clusters. $\widehat{Q}_{ij}^2$ is used to put more emphasis on data points assigned with higher confidence. DEC uses KL divergence to constrain the consistency between the soft assignments and target distribution.

$$\ell_{DEC} = KL\left(\widehat{\mathbf{P}} \| \widehat{\mathbf{Q}}\right) = \sum_i \sum_j \widehat{P}_{ij} \log \frac{\widehat{P}_{ij}}{\widehat{Q}_{ij}} \tag{14}$$

After each iteration of DEC, we update the soft assignment $\widehat{\mathbf{Q}}$ and the target assignments $\widehat{\mathbf{P}}$ according to equations (10) and (13) based on the new optimized image features and the cluster centers $\widehat{g}$. Finally, we take $\bar{\mathbf{P}}$ as the approximate unknown label matrix.

3.4 Establish Complete Label Classification Model

After obtaining the approximate unknown label matrix, we can construct a classification for multi-label image classification. Specifically, we first fuse the feature maps from last three blocks of backbone Resnet-101. Then we adopt an attention mechanism to learn label-specific and group-specific features. Finally, we construct a label related classifier to classify both known and unknown labels.

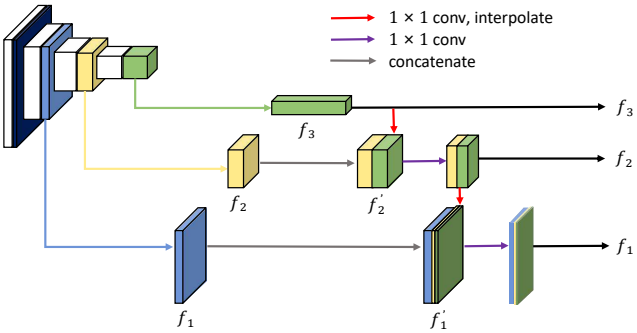

Fig. 4 After up-sampling and dimension reduction, deep and shallow features are concatenated in channel dimensions. Shallow features fuse information about themselves and deep features.

3.4.1 Multi-layer feature fusion

In order to obtain more meaningful and discriminative features to improve the classification performance, similar to FPN [53], we fuse the features from multi-layers. This method can produce a multi-scale feature representation in which all layers are semantically strong, including the high-resolution layers. Our feature fusion method is shown in Figure 4.

Firstly, through a 1x1 conv layer, the feature map f_3 is interpolated with a bilinear operation and concatenated with f_2 to obtain $f_2^{'}$. Then we change its dimension to obtain new f_2 with a 1x1 conv layer. The processes are as follows:

$$f_3 = \text{map}(f_3, c_3, c_2),$$

$$f_2^{'} = \text{concat}(\text{interpolate}\,(f_3, size = (w_2, h_2)), f_2),$$
$$f_2 = \text{map}(f_2^{'}, 2c_2, c_2),$$

where $f_2^{'} \in \mathcal{R}^{2c_2 \times w_2 \times h_2}$ and $f_2 \in \mathcal{R}^{c_2 \times w_2 \times h_2}$. $\text{map}(\cdot, inchannels, outchannels)$ function applies a convolution layer with 1×1 kernel over input feature maps, "in channels" is a parameter that represents the number of input channels of feature maps and "out channels" is a parameter that represents the number of channels produced by the convolution.

Secondly, we fuse the feature maps f_2 and f_1 by the same strategy. The processes are as follows:

$$f_1^{'} = \text{concat}\,(\text{interpolate}\,(\text{map}(f_2, c_2, c_1), size = (w_1, h_1)), f_1)\,,$$

$$f_1 = \text{map}(f_1^{'}, 2c_1, c_1).$$

After the feature fusion, we can obtain the new feature maps: $F = \{f_s\}_{s=1}^{B}$, where $f_s \in \mathcal{R}^{c_s \times w_s \times h_s}$ and $B = 3$.

3.4.2 Label Related Classifier

Motivated by the work MSRN [54], we use an attention mechanism to learn label-specific and group-specific features, and then construct the label related classifiers.

First, following MSRN [54], with the graph attention networks (GAT) [55] and the differentiable graph pooling (Diffpool) [56], we can obtain the semantic label embeddings E_l and semantic group embeddings E_g from the complete label relationship graph $\widehat{A}$ by

$$E_l = \text{GAT}\left(\widehat{V}, \widehat{A}\right), E_g = \text{Diffpool}\left(E_l, \widehat{A}\right), \tag{15}$$

where $E_l = \left\{e_l^i\right\}_{i=1}^{l} \in \mathbb{R}^{l\times d}$ and $E_g = \left\{e_g^j\right\}_{i=1}^{z} \in \mathbb{R}^{z\times d}$, e_l^i represents the semantic label embedding for the ith label, and e_g^j represents the semantic group embedding for the jth label group, and there are z groups.

Second, we use the the semantic label embeddings E_l and semantic group embeddings E_g to guide the network to learn label-specific and group-specific features by an attention mechanism. Specifically, we first map the feature maps from different blocks to the same dimension with the label embeddings, i.e., $f_s = \text{map}(f_s, c_s, d)$ and finally we get the feature maps $F = \{f_s\}_{s=1}^{B}$, where $f_s \in \mathcal{R}^{d\times w_s\times h_s}$. Then, we adopt the same attention mechanism in MSRN [54]. The attention scores of each position of the feature maps are calculated as

$$sl^i_{s_{w,h}} = f_b^{w,h} \odot e_l^i, \quad sg^j_{s_{w,h}} = f_s^{w,h} \odot e_g^j, \tag{16}$$

where the $\odot$ is Hadamard product, $sl_{s_{w,h}} \in \mathbb{R}^{1\times1\times l\times d}$ and $sg_{s_{w,h}} \in \mathbb{R}^{1\times1\times z\times d}$. Then we can obtain the normalized scores $al_s \in \mathbb{R}^{w_s\times h_s\times l\times d}$ and $ag_s \in \mathbb{R}^{w_s\times h_s\times z\times d}$ according to

$$al_s^{w,h} = \frac{\exp\left(sl_{s_{w,h}}\right)}{\sum_{x,y}\exp\left(sl_{s_{x,y}}\right)}, ag_s^{w,h} = \frac{\exp\left(sg_{s_{w,h}}\right)}{\sum_{x,y}\exp\left(sg_{s_{x,y}}\right)} \tag{17}$$

Then we apply the second Hadamard product to generate position-wise attention maps according to

$$o_s = \sum_{w,h} al_s^{w,h} \odot f_s^{w,h}, \quad q_s = \sum_{w,h} ag_s^{w,h} \odot f_s^{w,h}. \tag{18}$$

Consequently, for each image, we can obtain the label-specific features $K = \{o_s\}_{s=1}^{B} \in \mathbb{R}^{l\times(Bd)}$ and the group-specific features $J = \{q_s\}_{s=1}^{B} \in \mathbb{R}^{z\times(Bd)}$.

Finally, we construct the classifier for labbel prediction. Different from MSRN [54], for each label y_i, the classifier is constructed based on the feature representation which is composed of its label-specific feature $K^i \in \mathcal{R}^{1\times Bd}$ $(1 \le i \le l)$ and the corresponding group features $J^k \in \mathcal{R}^{1\times Bd}$ $(1 \le k \le z)$[1].

[1] The ith label belongs the k group.

Therefore, the classifier for the ith label can be defined as

$$u^i = sum\left(T^i \odot tanh(S^i)\right) + Bias^i, \tag{19}$$

where $u^i \in \mathcal{R}^{1\times 1}$, and $S^i = concat(K^i, J^k) \in \mathcal{R}^{1\times 2Bd}$. $T^i \in \mathcal{R}^{1\times 2Bd}$ and $Bias^i \in \mathcal{R}^{1\times 1}$ are learnable model parameters.

3.4.3 Trainning Loss

In this section, we will introduce the training loss function used in this paper. Firstly, for the known labels, we adopt the BCE loss

$$\ell_{known} = -\frac{1}{q}\sum_{i=1}^{q}\left[y^i \times \ln W^i + (1-y^i)\times \ln\left(1-W^i\right)\right], \tag{20}$$

where $W^i = \frac{1}{1+e^{-u^i}}$. y^i is the ground truth of the ith known label.

Same to MSRN, we add the following loss function to constrain the model to learn more compact group embeddings.

$$\ell_{Diff} = \sum_{k=1}^{z}\sum_{e_k^i \in C_k} \left\|e_l^i - e_g^k\right\|_2^2, \tag{21}$$

where C_k represents the set of label semantic representations for the kth label group.

For the unknown labels, we also adopt the BCE loss and constrain the consistency between the prediction of the classifier and the output of DEC.

$$\ell_{unknown} = -\frac{1}{r}\sum_{i=1}^{r}\left[\overline{P}^i \times \ln W^{q+i} + (1-\overline{P}^i)\times \ln\left(1-W^{q+i}\right)\right], \tag{22}$$

where $W^{q+i} = \frac{1}{1+e^{-u^{q+i}}}$, $1 \le i \le r$. u^{q+i} and $\overline{P}^i$ represent the prediction of classifier and the output of DEC for the ith unknown label respectively.

Therefore, the total loss function of the proposed method is defined as

$$loss = \ell_{known} + \ell_{unknown} + \ell_{DEC} + \ell_{Diff}. \tag{23}$$

According to the above analyses, the training processes of our proposed method can be summarized in Algorithm 1. $\widehat{\mathbf{A}}$ is complete label relationship graph, $\widehat{g}$ is cluster centers of DEC, $\widehat{\mathbf{V}}$ is complete label semantic description, and θ is the model parameters. Give a set of test images $\mathbf{I}_{test}$, the predictions can be obtained by $\mathbf{U} = h(\mathbf{I}_{test}, \theta)$.

Algorithm 1 Training.

Input: $\mathbf{I}, \mathbf{A}, \mathbf{V}$
Output: the classifier $h(\theta)$
Initialize $\widehat{\mathbf{A}}, \overline{\mathbf{V}}$
Repeat:
  Update θ via SGD and ℓ_{known}
  Until convergence or the maximum iteration is reached.
Initialize $\widehat{g}$.
Repeat:
  Update $\theta, \widehat{g}$ via SGD and $loss$
  Until convergence or the maximum iteration is reached.

4 Experiments

4.1 Experiment Setting and Evaluation Metrics

To verify the effectiveness of our model on known labels, we compared it with four advanced models. Detailed configurations of models are summarized as: 1) MCAR [57]: A multi-class attention classification model via combining global and local regions. 2) ADD-GCN [58]: A classification model with dynamic graph. 3) MSRN [54]: A classification model with multi-layered semantic representation. We take known label relationship graph as input to GCN. 4) ResNet-101 [59]: Residual structure. 5) our model: We use kmeans as the initialize method of DEC, and learning rate of DEC is set to 0.0001. The threshold τ is set to 0.3, α is set to 0.3. The backbone ResNet-101 is pretrained on ImageNet for accelerating training process. We remove the last average pooling layer and classifier of backone and apply the MaxPooling with kernel size 2×2 and stride 2 to obtain image features. The initial learning rate of our model is set to 0.1, and the learning rate decay by 10% each 15 epochs in total 90 epochs. The output dimension of first four models is set to q, and to l for the proposed method. For all models, we train all models on GeForce RTX 2080Ti-11GB GPU and set the batch size to 8. The input image size of all models is 448×448.

In the experiment, we evaluate the performance of our proposed model on known and unknown labels respectively. The evaluation metrics we used to evaluate the performance of known labels include mean average precision (mAP) over all categories, precision (CP, OP), recall (CR, OR), and F1 score (CF1, OF1). We adopt the following metric proposed in Zhu et al. [35] to evaluate the performance of our model on predicted unknown labels.

$$F_{novel} = \max\left(\{F\left(\ \mathbf{U}_{:,q+i}\ ,\ \mathbf{Y}_{:,j}\right), j \in \{1, ..., l\}\}\right) \tag{24}$$

Where $\mathbf{U} \in \mathcal{R}^{n \times l}$ is the label prediction of all samples from classifier. $\mathbf{U}_{:,q+i}$ is the prediction of the $(q+i)$th label of all samples. $\mathbf{Y}_{:,j}$ is the jth column of $\mathbf{Y}$.

4.2 Experiment Results of Known Labels

Our experiments are conducted on three multi-label image classification benchmark datasets, i.e., VOC2007 [60], VOC2012 [61] and Apparel[2]. Following the settings in previous work [47], the first 70%l, 80%l and 90%l labels are set be to known labels and the rest are taken as unknown ones respectively. All the comparing models have the same known and unknown label split. The average results of each comparing algorithm on the known labels are shown in Tables 1-3.

VOC2007: The comparison results on VOC2007 dataset are shown in Table 1. There are 5011 images for training and 4952 images for testing. Our model perform better than state-of-the-art methods on most metrics. When the numbers of unkonwn labels are 10%l and 20%l, the mAP of ours is 92.979% and 92.694%, and it is slightly higher than the second-place model by 0.18% and 0.17%. When the numbers of unkonwn labels is 30%l, our model achieves 95.140% mAP. It is significantly higher than the others. It can be seen that our model can have better performance when there are more unknown labels. It indicates that our model can effectively discover unknown label information and use discovered unknown labels' information to improve the performance of known labels.

VOC2012: It consists of 11,540 images as training and validation set, we use VOC2012 as the training set and VOC2007 as the test set. When the numbers of unkonwn labels are 30%l, the comparison results are shown in Table 2, Our method can achieve 96.007% mAP score. Our model achieves comparable performance with the state-of-the-art methods, and wins the second place in terms of CP, CR, CR-3 and OR-3.

Apparel: The comparison results on Apparel dataset are shown in Table 3. In our experiment, we randomly select 50% images from the dataset for training, and other 50% images for testing. Our model performs better than Resnet on all metrics.

4.3 Experiment Results of Unknown Labels

Experiments of The Best Matched Labels. The prediction of unknown labels are matched with the complete label matrix, and F_{novel} is used to measure the compatibility of distributions. In the experiment, the unknown labels are deleted from the complete labels. Table 4 shows the results of the five best matched labels for the unknown labels over the VOC2007 dataset with $q = 80\%l$ and $r = 20\%l$. The left column lists the top-5 matched labels for a predicted unknown label, and the right column indicates the best matched name of unknown labels. It can be seen that the first label is matched with a very higher F1 score than other labels. For example, the unknown label "aeroplane" is matched with the ground-truth with a very high F1 score 0.856. These experimental results verify that the proposed method can discover these unknown labels.

[2]www.kaggle.com/kaiska/apparel-dataset

Table 1 Experimental results on VOC2007

$q = 90\%l, r = 10\%l$	mAP	CP	CR	CF1	OP	OF1	CP-3	CR-3	CF1-3	OP-3	OR-3	OF1-3
ResNet-101	91.354	0.848	0.833	0.840	0.868	0.861	0.814	0.857	0.835	0.831	0.874	0.852
ADD-GCN	92.65	0.845	0.875	0.86	0.871	**0.883**	0.847	0.871	0.859	0.874	0.892	**0.883**
MCAR	92.799	0.761	**0.915**	0.831	0.781	0.846	0.766	**0.909**	0.831	0.7873	**0.918**	0.848
MSRN	91.662	0.857	0.854	0.856	0.872	0.875	0.836	0.869	0.852	0.8448	0.891	0.867
Ours	**92.979**	**0.887**	0.848	**0.867**	**0.90**	0.886	**0.871**	0.860	**0.866**	**0.883**	0.883	**0.883**

$q = 80\%l, r = 20\%l$	mAP	CP	CR	CF1	OP	OF1	CP-3	CR-3	CF1-3	OP-3	OR-3	OF1-3
ResNet-101	91.324	0.859	0.811	0.834	0.827	0.820	0.822	0.846	0.833	0.766	0.849	0.805
ADD-GCN	92.44	0.856	0.868	0.862	0.829	0.852	0.856	0.868	0.862	0.829	0.876	0.852
MCAR	92.694	0.777	**0.914**	0.840	0.751	0.825	0.777	**0.914**	0.840	0.752	**0.915**	0.825
MSRN	92.245	0.862	0.864	0.863	0.847	**0.856**	0.839	0.879	0.858	0.820	0.881	0.849
Ours	**92.863**	**0.885**	0.844	**0.864**	**0.880**	**0.856**	**0.874**	0.854	**0.864**	**0.866**	0.845	**0.855**

$q = 70\%l, r = 30\%l$	mAP	CP	CR	CF1	OP	OF1	CP-3	CR-3	CF1-3	OP-3	OR-3	OF1-3
ResNet-101	93.020	0.8547	0.868	0.861	0.829	0.850	0.824	0.889	0.855	0.790	0.8927	0.838
ADD-GCN	94.73	0.888	0.902	**0.895**	0.879	0.893	**0.889**	0.902	**0.895**	0.880	0.908	**0.894**
MCAR	94.630	0.8090	**0.928**	0.864	0.822	0.871	0.810	**0.927**	0.864	0.822	**0.926**	0.871
MSRN	92.756	0.8637	0.847	0.855	0.822	0.832	0.870	0.836	0.853	0.832	0.833	0.833
Ours	**95.140**	**0.908**	0.880	0.894	**0.907**	**0.894**	0.897	0.890	0.893	**0.894**	0.890	0.892

Table 2 Experimental results on VOC2012

$q = 70\%l, r = 30\%l$	mAP	CP	CR	CF1	OP	OF1	CP-3	CR-3	CF1-3	OP-3	OR-3	OF1-3
ResNet-101	94.294	0.893	0.880	0.886	0.878	0.881	0.872	0.893	0.882	0.854	0.896	0.875
MCAR	95.492	**0.947**	**0.925**	0.884	0.852	0.880	0.862	**0.904**	0.882	0.862	**0.910**	0.880
MSRN	93.744	0.870	0.843	0.856	0.858	0.853	0.845	0.866	0.856	0.823	0.872	0.847
Ours	**96.007**	0.909	0.893	**0.901**	**0.901**	**0.898**	**0.904**	0.899	**0.901**	**0.894**	0.902	**0.898**

Table 3 Experimental results on Apparel

$q = 70\%l, r = 30\%l$	mAP	CP	CR	CF1	OP	OF1	CP-3	CR-3	CF1-3	OP-3	OR-3	OF1-3
ResNet-101	99.281	0.964	0.970	0.967	0.969	0.968	0.970	0.962	0.966	0.974	0.959	0.966
Ours	**99.506**	**0.983**	**0.974**	**0.978**	**0.986**	**0.981**	**0.977**	**0.980**	**0.978**	**0.980**	**0.981**	**0.981**

Table 4 Experimental results of the best matched labels on voc2007 for the three predicted unknown labels

Top 5 Matched Labels (i.e., label name (F1 score))	Unknown Label
aeroplane(**0.856**), person(0.039), car(0.028), motorbike(0.004), pottedplant(0.004)	aeroplane
bicycle(**0.882**), person(0.120), motorbike(0.053), car(0.014), bottle(0.013)	bicycle
bird(**0.717**), sheep(0.012), dog(0.006), pottedplant(0.004), person(0.003)	bird

Table 5 Experimental results on voc2007 with different clustering algorithms in terms of F_{novel}

	cluster algorithm	bird	train	cat	chair	sofa	tvmonitor	aeroplane	dog	horse	motorbike
$q = 90\%l$	NMF	0.650	0.5								
$r = 10\%l$	DEC	**0.873**	**0.902**								
$q = 80\%l$	NMF			0.544	0.248	**0.241**	0.330				
$r = 20\%l$	DEC			**0.739**	**0.490**	0	**0.637**				
$q = 70\%l$	NMF	0.293	0.402					0.318	**0.050**	0.0428	0.3169
$r = 30\%l$	DEC	**0.792**	**0.830**					**0.904**	0	**0.786**	**0.770**

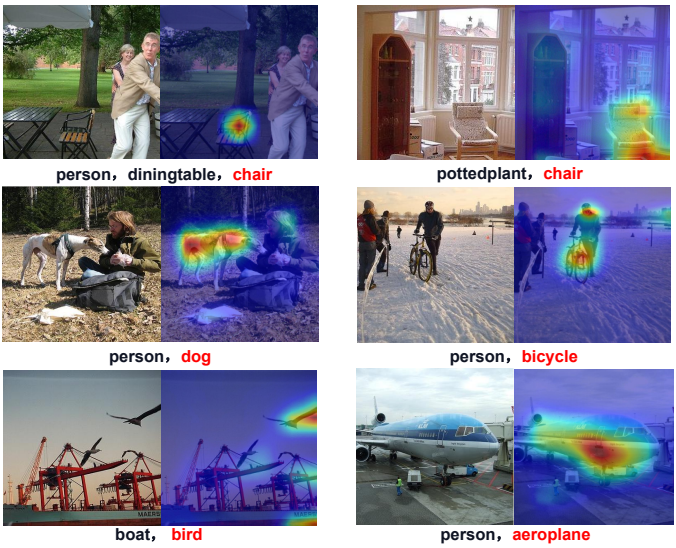

Fig. 5 In the six images, the left side of each image shows the original image of unknown labels, and the right side of each image displays the attention map of unknown labels after discovering by our model. The words below each image are the labels of the image. The black words are known labels, while the red words are unknown labels.

Experiments with DLCL by setting different clustering algorithms. Since DLCL [48] solves the same problem as we do, we set up an experiment to compare with DLCL. However, DLCL is a shallow model and and uses NMF to discover unknown label information, and its feature extraction ability is pool as that of a deep model, so we replace the DEC component with NMF in our model. We selected different number of unknown labels to conduct the experiment. Table 5 shows the results of using different clustering algorithms over the VOC2007 dataset. It can be seen from first line, unknown labels are "bird" and "train". The scores of F_{novel} of our model with NMF are 0.650 and 0.5. They are much lower than our model with DEC. However, it can be seen from second line, when unknown label is "sofa", F_{novel} of DEC is 0, and that of NMF is 0.241. It is suggest that our model is better than DLCL at most of the time, and cannot discover some unknown labels in some cases.

4.4 Attention Map of Unknown Labels

Our model learns an attention map for each label. After learning the unknown labels, we output the attention maps of unknown labels. Figure 5 shows six examples of attention maps for the unknown labels (i.e., chair, dog, bicycle, bird, and aeroplane). We can observe that the attention maps focus on the unknown labels, and we can get the semantic information for the unknown labels from these results. These experimental results clearly demonstrate that our proposed model can effectively discover unknown labels.

Table 6 Ablation Study

	mAP
Basemodel	89.632
Basemodel + multi-layer feature fusion	90.004
Basemodel + label related classifier	91.374
Ours	**93.42**

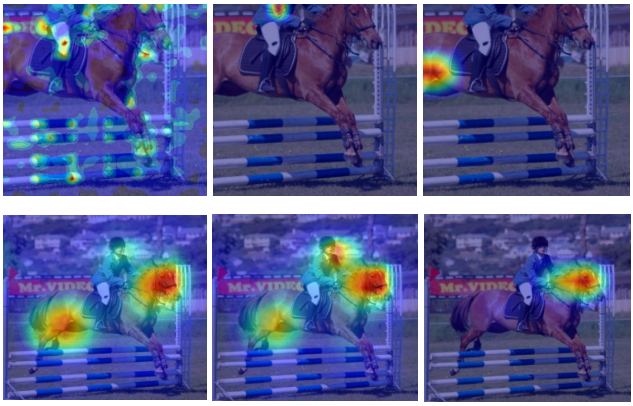

Fig. 6 The attention maps of label "horse". The attention maps for three images in the first line are generated by the basemodel, and the following three ones are generated by our model.

4.5 Ablation Study

We used basemodel as a benchmark to verify the effectiveness of each component of our model. The basemodel uses ResNet-101 to extract features, GAT and Diffpool to extract semantic embeddings, and a semantically guided attention mechanism is used to fuse semantics and features. And it does not take into account unknown labels. On the basis of basemodel, our model added feature fusion, label related classifier, and module for discovering unknown labels in our model. At the same time, we use complete label relationship graph as GAT's input, including known and unknown labels. We conduct some ablation studies on VOC2007 dataset, as shown in table 6. From the results, we can see the mAP of basemodel is 89.632. The mAP of basemodel with multi-layer feature fusion is 90.004 and that of basemodel with label related classifier is 91.374. These are all slightly higher than basemodel. Ours containing both the two modules and discovering unknown labels is much higher than the basemodel. These results clearly verify the effectiveness of each component of our model.

We also visualize the attention maps of the basemodel and our model. It is shown in Figure 6. The following three images are generated by our model with the multi-layer feature fusion module. It can be seen that attention maps of our model focus on more meaningful regions. It also can verify the effective of our model.

5 Conclusion

In this paper, we proposed a new deep model that can predict unknown labels in MLL for image classification. At the same time, we proposed a method to update graph neural network with the information of unknown labels and improved the performance of known labels. Finally, we can obtain the semantic information of unknown labels by attention maps. Extensive experiments on various datasets demonstrate the effectiveness of the proposed framework on discovering and predicting unknown labels for multi-label image classification.

Acknowledgement

This work is supported by NSFC: 61806005, The University Synergy Innovation Program of Anhui Province:GXXT-2020-012 and GXXT-2019-025, and the Natural Science Foundation of the Educational Commission of Anhui Province of China: KJ2021A0373.

References

[1] Xing, Y., Yu, G., Domeniconi, C., Wang, J., Zhang, Z.: Multi-label co-training. In: Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18 (2018)

[2] Wu, X., Chen, Q.-G., Hu, Y., Wang, D., Chang, X., Wang, X., Zhang, M.-L.: Multi-view multi-label learning with view-specific information extraction. In: IJCAI, pp. 3884–3890 (2019)

[3] Wang, H., Liu, W., Zhao, Y., Zhang, C., Hu, T., Chen, G.: Discriminative and correlative partial multi-label learning. In: IJCAI, pp. 3691–3697 (2019)

[4] Xu, N., Lv, J., Geng, X.: Partial label learning via label enhancement. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 5557–5564 (2019)

[5] Sun, Y.-Y., Zhang, Y., Zhou, Z.-H.: Multi-label learning with weak label. In: Twenty-fourth AAAI Conference on Artificial Intelligence (2010)

[6] Zhao, F., Guo, Y.: Semi-supervised multi-label learning with incomplete labels. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)

[7] Bi, W., Kwok, J.: Multilabel classification with label correlations and missing labels. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 28 (2014)

[8] Yu, H.-F., Jain, P., Kar, P., Dhillon, I.: Large-scale multi-label learning with missing labels. In: International Conference on Machine Learning, pp. 593–601 (2014). PMLR

[9] Wu, B., Jia, F., Liu, W., Ghanem, B., Lyu, S.: Multi-label learning with missing labels using mixed dependency graphs. International Journal of Computer Vision **126**(8), 875–896 (2018)

[10] Wu, B., Lyu, S., Ghanem, B.: Constrained submodular minimization for missing labels and class imbalance in multi-label learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30 (2016)

[11] Liu, Y., Wen, K., Gao, Q., Gao, X., Nie, F.: Svm based multi-label learning with missing labels for image annotation. Pattern Recognition **78**, 307–317 (2018)

[12] Ma, Z., Chen, S.: Expand globally, shrink locally: Discriminant multi-label learning with missing labels. Pattern Recognition **111**, 107675 (2021)

[13] Yang, H., Zhou, J.T., Cai, J.: Improving multi-label learning with missing labels by structured semantic correlations. In: European Conference on Computer Vision, pp. 835–851 (2016). Springer

[14] Xu, L., Wang, Z., Shen, Z., Wang, Y., Chen, E.: Learning low-rank label correlations for multi-label classification with missing labels. In: 2014 IEEE International Conference on Data Mining, pp. 1067–1072 (2014). IEEE

[15] Huang, J., Qin, F., Zheng, X., Cheng, Z., Yuan, Z., Zhang, W., Huang, Q.: Improving multi-label classification with missing labels by learning label-specific features. Information Sciences **492**, 124–146 (2019)

[16] Guo, B., Hou, C., Shan, J., Yi, D.: Low rank multi-label classification with missing labels. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 417–422 (2018). IEEE

[17] Zhu, P., Xu, Q., Hu, Q., Zhang, C., Zhao, H.: Multi-label feature selection with missing labels. Pattern Recognition **74**, 488–502 (2018)

[18] Zhao, D., Gao, Q., Lu, Y., Sun, D.: Two-step multi-view and multi-label learning with missing label via subspace learning. Applied Soft Computing **102**, 107120 (2021)

[19] Pham, A., Raich, R., Fern, X., Arriaga, J.P.: Multi-instance multi-label learning in the presence of novel class instances. In: International Conference on Machine Learning, pp. 2427–2435 (2015). PMLR

[20] Huang, J., Xu, L., Qian, K., Wang, J., Yamanishi, K.: Multi-label learning with missing and completely unobserved labels. Data Mining and Knowledge Discovery **35**, 1061–1086 (2021)

[21] Li, P., Wang, H., Böhm, C., Shao, J.: Online semi-supervised multi-label classification with label compression and local smooth regression. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 1359–1365 (2021)

[22] Zhang, X., Graepel, T., Herbrich, R.: Bayesian online learning for multi-label and multi-variate performance measures. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 956–963 (2010). JMLR Workshop and Conference Proceedings

[23] Zhang, Y.-J., Zhao, P., Zhou, Z.-H.: A simple online algorithm for competing with dynamic comparators. In: Conference on Uncertainty in Artificial Intelligence, pp. 390–399 (2020). PMLR

[24] Boulbazine, S., Cabanes, G., Matei, B., Bennani, Y.: Online semi-supervised growing neural gas for multi-label data classification. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2018). IEEE

[25] Wu, F., Liu, Q., Hao, T., Chen, X., Wu, Q.: Online multi-instance multi-label learning for protein function prediction. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 780–785 (2016). IEEE

[26] Wang, Q., Cheng, J.-Z., Zhou, Y., Zhuang, H., Li, C., Chen, B., Liu, Z., Huang, J., Wang, C., Zhou, X.: Low-shot multi-label incremental learning for thoracic diseases diagnosis. In: International Conference on Neural Information Processing, pp. 420–432 (2018). Springer

[27] Tao, X., Hong, X., Chang, X., Dong, S., Wei, X., Gong, Y.: Few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12183–12192 (2020)

[28] Kuzborskij, I., Orabona, F., Caputo, B.: From n to n+ 1: Multiclass transfer incremental learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3358–3365 (2013)

[29] Dong, Q., Gong, S., Zhu, X.: Imbalanced deep learning by minority class incremental rectification. IEEE transactions on pattern analysis and machine intelligence **41**(6), 1367–1381 (2018)

[30] Wang, Q., Chen, K.: Multi-label zero-shot human action recognition via joint latent ranking embedding. Neural Networks **122**, 1–23 (2020)

[31] Lee, C.-W., Fang, W., Yeh, C.-K., Wang, Y.-C.F.: Multi-label zero-shot learning with structured knowledge graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1576–1585 (2018)

[32] Huynh, D., Elhamifar, E.: A shared multi-attention framework for multi-label zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8776–8786 (2020)

[33] Zhang, L., Wang, P., Liu, L., Shen, C., Wei, W., Zhang, Y., Van Den Hengel, A.: Towards effective deep embedding for zero-shot learning. IEEE Transactions on Circuits and Systems for Video Technology **30**(9), 2843–2852 (2020)

[34] Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey (2021)

[35] Wang, H., Bah, M.J., Hammad, M.: Progress in outlier detection techniques: A survey. Ieee Access **7**, 107964–108000 (2019)

[36] Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Müller, K.: A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE **PP**(99), 1–40 (2021)

[37] Pang, G., Shen, C., Cao, L., Hengel, A.: Deep learning for anomaly detection: A review. ACM Computing Surveys **54**(2), 1–38 (2021)

[38] Bulusu, S., Kailkhura, B., Li, B., Varshney, P.K., Song, D.: Anomalous example detection in deep learning: A survey. IEEE Access **PP**(99), 1–1 (2020)

[39] Chalapathy, R., Chawla, S.: Deep learning for anomaly detection: A survey (2019)

[40] Abati, D., Porrello, A., Calderara, S., Cucchiara, R.: Latent space autoregression for novelty detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 481–490 (2019)

[41] Perera, P., Nallapati, R., Xiang, B.: Ocgan: One-class novelty detection using gans with constrained latent representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2898–2906 (2019)

[42] Perera, P., Patel, V.M.: Deep transfer learning for multiple class novelty detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11544–11552 (2019)

[43] Kong, S., Ramanan, D.: Opengan: Open-set recognition via open data generation. (2021)

[44] Geng, C., Huang, S.-j., Chen, S.: Recent advances in open set recognition: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3614–3631 (2020)

[45] Perera, P., Morariu, V.I., Jain, R., Manjunatha, V., Wigington, C., Ordonez, V., Patel, V.M.: Generative-discriminative feature representations for open-set recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11814–11823 (2020)

[46] Zhang, Y.-J., Zhao, P., Zhou, Z.-H.: Exploratory machine learning with unknown unknowns. arXiv preprint arXiv:2002.01605 (2020)

[47] Zhu, Y., Ting, K.M., Zhou, Z.-H.: Discover multiple novel labels in multi-instance multi-label learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)

[48] Huang, J., Xu, L., Wang, J., Feng, L., Yamanishi, K.: Discovering latent class labels for multi-label learning. (2020). International Joint Conferences on Artificial Intelligence Organization (IJCAI)

[49] Lee, D., Seung, H.S.: Algorithms for non-negative matrix factorization. Advances in neural information processing systems **13** (2000)

[50] Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[51] Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning, pp. 478–487 (2016). PMLR

[52] Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

[53] Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

[54] Qu, X., Che, H., Huang, J., Xu, L., Zheng, X.: Multi-layered semantic representation network for multi-label image classification. arXiv preprint arXiv:2106.11596 (2021)

[55] Velikovi, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.:

Graph attention networks. (2017)

[56] Ying, R., You, J., Morris, C., Ren, X., Hamilton, W.L., Leskovec, J.: Hierarchical graph representation learning with differentiable pooling. (2018)

[57] Qu, X., Che, H., Huang, J., Xu, L., Zheng, X.: Multi-layered semantic representation network for multi-label image classification. arXiv preprint arXiv:2106.11596 (2021)

[58] Ye, J., He, J., Peng, X., Wu, W., Qiao, Y.: Attention-driven dynamic graph convolutional network for multi-label image recognition. In: European Conference on Computer Vision, pp. 649–665 (2020). Springer

[59] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[60] Everingham, M., Zisserman, A., Williams, C.K., Van Gool, L., Allan, M., Bishop, C.M., Chapelle, O., Dalal, N., Deselaers, T., Dorkó, G., et al.: The pascal visual object classes challenge 2007 (voc2007) results (2008)

[61] Everingham, M., Winn, J.: The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep **8**, 5 (2011)