Deep-LIFT: Deep Label Specific Feature Learning for Image Annotation

Junbing Li, Changqing Zhang, *Member, IEEE,* Joey Tianyi Zhou, *Member, IEEE,*
Huazhu Fu, *Senior Member, IEEE,* Shuyin Xia, *Member, IEEE,* and Qinghua Hu, *Senior Member, IEEE*

***Abstract*—Image annotation aims to jointly predict multiple tags for an image. Although significant progress has been achieved, existing approaches usually overlook aligning specific labels and their corresponding regions due to the weak supervised information (*i.e.*, "bag of labels" for regions), thus fail to explicitly exploit the discrimination from different classes. In this paper, we propose the Deep Label-Specific Feature (Deep-LIFT) learning model to build the explicit and exact correspondence between label and local visual region, which improves the effectiveness of feature learning and enhances the interpretability of the model itself. Deep-LIFT extracts features for each label by aligning each label and its region. Specifically, deep label-specific features are achieved through learning multiple correlation maps between image convolutional features and label embeddings. Moreover, we construct two variant graph convolutional networks (GCN) to further capture the inter-dependency among labels. Empirical studies on benchmark datasets validate that the proposed model achieves superior performance on multi-label classification over other existing state-of-the-art methods.**

***Index Terms*—Deep-LIFT, image annotation, label-specific, variant GCN.**

I. Introduction

IMAGE annotation aims to accurately predict multiple tags for an image reflecting its semantic content, which is a fundamental, practical and still very challenging task in computer vision. Multi-label image annotation can be applied in many real-world applications such as scene recognition [1], [2], [3], human attribute recognition [4], medical diagnosis recognition [5]. Compared to predicting one single class label for an image [6], [7], [8], multi-label annotation problem is more difficult due to the combinatorial nature of the output label space. Existing methods [9], [10], [11], [12] usually address the problem by simultaneously modeling the relationships between the input image and all categories and capturing the inter-dependency among different labels.

This work was supported in part by National Natural Science Foundation of China (No. 61976151, No. 61732011, No. 61876127 and No. 62076179), the Natural Science Foundation of Tianjin of China (No. 19JCYBJC15200), the Tianjin Science and Technology Plan Project under Grant (No. 19ZXZNGX00050).

J. Li, C. Zhang and Q. Hu are with the College of Intelligence and Computing, Tianjin Key Laboratory of Machine Learning, Tianjin University, Tianjin 300072, China. (e-mail: lijunbing@tju.edu.cn; zhangchangqing@tju.edu.cn; huqinghua@tju.edu.cn.) (Corresponding author: Changqing Zhang.)

J. T. Zhou is with the Institute of High Performance Computing, the Agency for Science, Technology and Research (A*STAR), Singapore (e-mail: joey.tianyi.zhou@gmail.com).

H. Fu is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: hzfu@ieee.org).

S. Xia is with the Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: xiasy@cqupt.edu.cn).

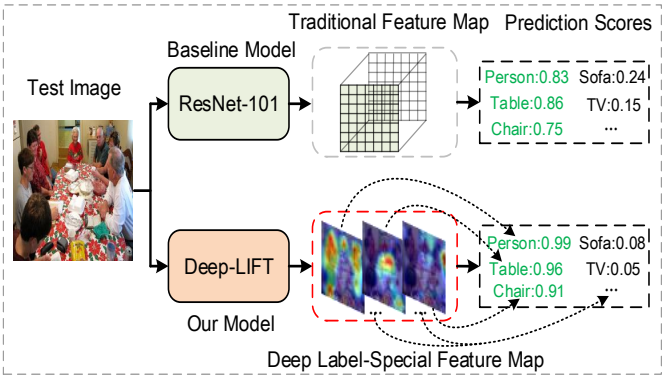


Fig. 1. Advantage of the proposed Deep-LIFT over baseline (ResNet-101). Traditional feature map (from the "conv5_x" layer of ResNet-101) can only express general characteristics for all labels, while our deep label-specific feature map explicitly exploits the discrimination from different classes and establishes the exact correspondence between visual region and semantic label.

Multi-label recognition problem can be transformed into a set of binary classification tasks and equipped with powerful feature representations learned with deep Convolutional Neural Networks (CNNs) [13], [14], [15], [16], [17] from raw images. However, this strategy ignores the useful label correlations which can improve the prediction accuracy in practice. Therefore, another line of researches on multi-label image annotation focus on effectively capturing the inter-dependency among labels, including probabilistic graphical models [18], [19], structured inference neural network [20], and Recurrent Neural Networks (RNNs) [21].

To improve the image representation, recent works [22], [23] introduce the attention mechanisms. Wang *et al.* [22] introduced a spatial transformer layer to locate important regions on the convolutional maps. Guo *et al.* [23] proposed the attention consistency assumption and designed a two-branch network with an original image and its transformed image as inputs. Different from the above methods, Chen *et al.* [24] firstly proposed a novel graph convolutional network based model (ML-GCN) to capture the label correlations for multi-label image recognition, which employs GCN to map label representations to inter-dependent object classifiers. The multi-label predicted scores of ML-GCN are directly obtained using the correlation score between the learned object classifier and global image features.

Previous attention-based methods have achieved significant improvement, however, they fail to consider the exact correspondence between class label and local visual region. The

main reason is these attention-based models cannot locate the semantic regions accurately due to the weak supervision, *i.e.*, "bag of labels" for "whole image" instead of "specific label" for "specific region". The label-specific features (LIFT) [25] have been recognized to be pertinent and discriminative features for each class label. Inspired by the label-specific features, we propose a Deep Label-Specific Feature (Deep-LIFT) learning model to promote the image annotation. Figure 1 illustrates the advantage of our proposed Deep-LIFT over the baseline (ResNet101), and we simultaneously extend graph convolutional network (GCN) to AGCN (Attention-GCN) and RGCN (Residual-GCN) to further capture the correlations among labels. The key of the proposed Deep-LIFT is to build a correlation map between image feature map and each label embedding, indicating the existence of content associated with a specific label. Furthermore, we introduce a channel attention module (CAM) [26] to strengthen the effect of the deep LIFT maps associated with positive labels, and reduce the effect of negative labels. The overall framework of our model is shown in Figure 2.

For clarification, the contributions of this work are summarized as follows:

- We propose a novel deep label-specific feature (Deep-LIFT) learning network by exploring the correlation between the semantic label and deep feature map, which effectively decomposes the traditional global features (for all labels) into label-specific features (for specific label), establishing the explicit and exact correspondence between visual region and semantic label.
- We develop two different types of GCN, *i.e.*, Attention-GCN and Residual-GCN, to explore the label semantic dependency in depth, where Attention-GCN can learn aggregation weights of neighbors automatically and Residual-GCN can reliably converge in training based on residual learning.

II. Related Work

Image classification has achieved great success under the deep convolutional networks (CNNs). Recently, researchers have proposed to tackle the problem of multi-label image classification, which is still challenging due to the complexity of correlations among multiple labels. The straightforward strategy is the label-by-label manner [1], [27], which learns an independent classifier for each label and is convenient to adopting existing single-label methods to multi-label task. Obviously, this strategy ignores the relationships among different labels, which is usually critical for multi-label learning. For this issue, there are plenty of approaches proposed to exploit the correlations among labels. Read *et al.* [28] proposed a chain of binary classifiers to extend the binary relevance method, where each classifier makes predictions based on the input and the predicted labels in last steps. Gong *et al.* [2] combines the deep convolutional neural networks with a ranking-based learning strategy for image annotation. Hu *et al.* [20] proposed a structured inference neural network to transfer multi-label prediction across multiple semantic concept layers. Wang *et al.* [21] converted multi-label image classification into a sequential prediction problem, and explored the semantic dependency among labels by Recurrent Neural Networks (RNN).

Graph is widely used in modeling label dependency and can capture complex relationships among labels. Traditional methods based on probabilistic graphical models include Conditional Random Field [29], Dependency Network [30], and co-occurrence matrix [31]. Recently, Li *et al.* [18] introduced the maximum spanning tree algorithm over mutual information matrix of labels to construct the label graph. Li *et al.* [19] employed the graphical Lasso framework to learn image-dependent conditional label structures. Lee *et al.* [32] advanced a label propagation mechanism by incorporating structured knowledge graphs. Recently, inspired by the rapid development of graph convolutional network, Chen *et al.* [24] employed GCN to map label representations (word embeddings) to inter-dependent object classifiers and capture the correlation between labels.

Moreover, attention mechanism has been proven to be beneficial for improving the performance of multi-label classification. Zhu *et al.* [33] proposed a spatial regularization network to exploit both semantic and spatial relations between labels with image-level supervision. Wang *et al.* [22] developed a recurrent memorized-attention module to effectively capture the label correlation, which consists of a spatial transformer layer and a long short-term memory (LSTM).

Different from previous methods, our proposed Deep-LIFT aims to establish the explicit and exact correspondence between visual region and semantic label. Specifically, The key of our model is to build a Deep-LIFT map between image feature maps and label re-embeddings with variant GCN, to decompose the traditional global features (for all labels) into label-specific features (for specific label). Moreover, we extend the GCN to AGCN (Attention-GCN) and RGCN (Residual-GCN) to effectively capture the correlations among labels. Thus, our Deep-LIFT is able to simultaneously learn label-specific feature representation and complex label correlations, leading to promising performance improvement in image annotation.

III. Our Method

In this section,we elaborate on our proposed Deep-LIFT for image annotation. Deep-LIFT consists of three modules: (1) Image Representation Learning, (2) Label Correlation Exploration and (3) Deep Attention LIFT Network. Firstly, we introduce the way for image feature extraction. Then, we describe our proposed two variant GCN (AGCN and RGCN) for label correlation exploration. Finally, the construction of Deep LIFT Maps and Deep Attention LIFT Maps will be described in detail.

A. Image Feature Extraction

There are different deep CNN architectures [13], [7], [34] in image feature extraction. Similar to the recently proposed multi-label learning approaches [24], [35], [33], we employ ResNet-101 [13] as backbone consisting of repetitive network modules with different output dimensions. Let $\mathbf{I}$ denotes an

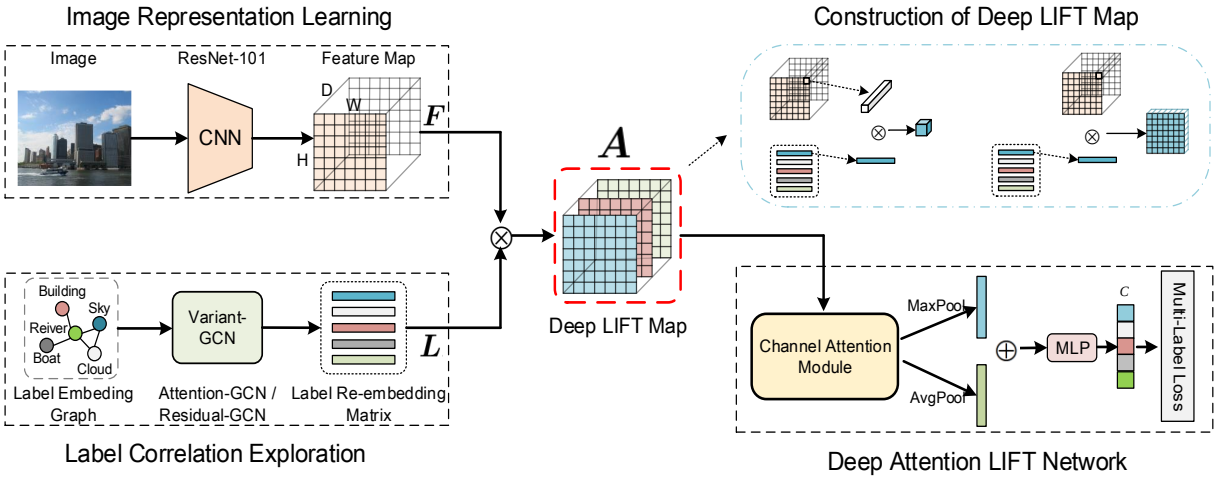


Fig. 2. Overview of our proposed Deep-LIFT. Deep-LIFT consists of three modules: (1) Image Representation Learning, (2) Label Correlation Exploration and (3) Deep Attention LIFT Network. The key of our model is to build a Deep-LIFT map between image feature map and each label re-embedding with variant GCN, to decompose the traditional global features (for all labels) into label-specific features (for specific label) and establish the explicit and exact correspondence between visual region and semantic label. Furthermore, we introduce a channel attention module (CAM) for the deep label-specific feature maps to automatically focus on the deep LIFT maps of positive labels. Specifically, the detailed construction of Deep LIFT Map is clearly shown in the upper right of the framework. Where $\otimes$ denotes multiplication of two tensors (matrixs).

input image (with the size of 448×448), the feature map (with the size of $2048 \times 14 \times 14$) from the "conv5_x" layer (the fully connected layer is on the top of it) of ResNet-101 is obtained as follows:

$$\boldsymbol{F} = f_{cnn}(\mathbf{I}; \theta_{cnn}), \tag{1}$$

where $\boldsymbol{F} \in \mathbb{R}^{D \times H \times W}$ can be considered as M spatial locations and each one corresponds to a D-dimensional visual feature vector, where $M = H \times W = 14 \times 14$ and $D = 2048$. The main reason for choosing the "conv5_x" layer is that higher layer has larger receptive fields and thus the corresponding high-level features tend to contain richer semantic information, which is more favorable to classification or detection task. Please refer to "Image Representation Learning" in Fig. 2

B. Encoding Label Correlation

Capturing the correlations among different labels can improve the multi-label image annotation performance, which has been well recognized. Inspired by the effectiveness of Graph Convolutional Network (GCN) [36] in exploring graph, we construct a graph to model the inter-dependencies among labels. Specifically, each node in the graph is associated with a semantic category, and there is also a word embedding for each label. To more effectively explore and utilize the correlation among labels, we propose two variants of GCN, AGCN (Attention-GCN) and RGCN (Residual-GCN). For clarification, we first introduce the preliminary knowledge of GCN, and then introduce AGCN and RGCN in detail.

1) Preliminaries of GCN: Graph Convolutional Network (GCN) [36] propagates link information on a graph to produce embedding for each node (semantic label in our task). A graph convolution layer takes a feature matrix $\boldsymbol{X} \in \mathbb{R}^{C \times d}$ with an adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{C \times C}$ as input, and outputs an updated feature matrix $\boldsymbol{Y} \in \mathbb{R}^{C \times d'}$, where C, d and d' are the number of classes, the dimensionality of the original feature space and the embedding with GCN, respectively. The GCN layer update the embeddings for nodes by

$$\boldsymbol{Y} = h(\widehat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}), \tag{2}$$

where $\widehat{\boldsymbol{A}} \in \mathbb{R}^{C \times C}$ is the normalized version of adjacency matrix $\boldsymbol{A}$ and $\boldsymbol{W} \in \mathbb{R}^{d \times d'}$ is a transformation matrix to be learned, and $h(\cdot)$ denotes a non-linear activation function.

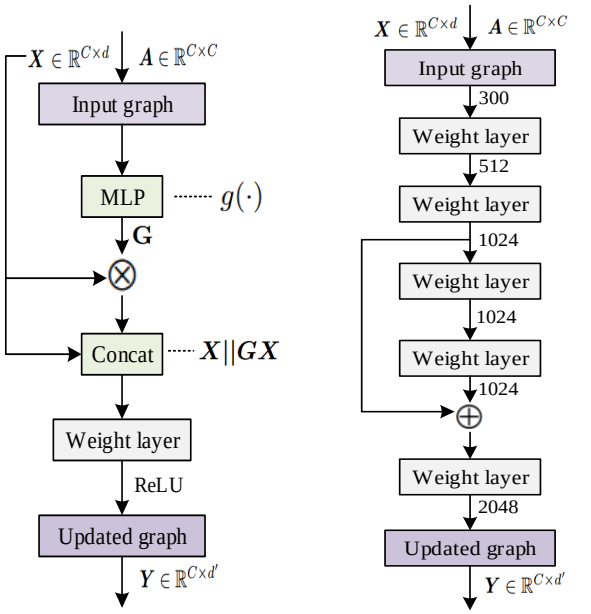
In our model, each node (label) is associated with a word embedding vector obtained by the GloVe [37] model trained on the Wikipedia dataset. The existence of links between label pairs are decided by the correlation of two labels. Specifically, following [24], we measure the degree of correlation between labels by computing their co-occurrence conditional probability. Furthermore, we set the threshold τ to filter possible noisy links, and produce our label adjacency matrix $\boldsymbol{A}$. Accordingly, the input feature matrix (in the first layer of GCN) is the original label word embedding feature matrix $\boldsymbol{X}$. For the last layer, the output representation matrix of GCN is $\boldsymbol{L} \in \mathbb{R}^{C \times D}$, where $D = 2048$ denotes the dimensionality of image visual feature $\boldsymbol{F}$ on each spatial location.

2) AGCN: Attention-GCN: Motivated by the idea of graph attention network [38], we try to *learn* the aggregation weights between node pairs instead of predefining manually. Specifically, one attention graph convolution layer in this case can be defined as

$$\boldsymbol{Y} = h([\boldsymbol{X}||\boldsymbol{G}\boldsymbol{X}]\boldsymbol{W}), \tag{3}$$

where d and d' are the dimensionality of input and output label embedding, respectively. $\mathbf{G} = g(\mathbf{X}, \mathbf{A})$ is an aggregation matrix of size $C \times C$ and each row is normalized with $\sum_j |G_{ij}| = 1$. $g(\cdot)$ is an attention aggregation function of $\boldsymbol{X}$ and $\boldsymbol{A}$. Operator $||$ denotes matrix concatenation along the feature direction. $\boldsymbol{W} \in \mathbb{R}^{2d \times d'}$ is the learnable weight matrix and $h(\cdot)$ is the non-linear activation function. In this way, our model can automatically learn attention weights of label pairs. Specifically, similar to the way in [38] to produce an attention

coefficient matrix, $g(\cdot)$ is implemented as a two-layer Multi-Layer Perceptron (MLP) with the features of a pair of label - neighbor nodes as inputs.



(a) A graph convolution layer of the Attention-GCN.

(b) The detailed structure of the Residual-GCN.

Fig. 3. The detailed structures of the proposed AGCN and RGCN.

3) RGCN: Residual-GCN: Inspired by the success of ResNet [13], we extend the deep GCN architecture with residual learning. Specifically, we propose a graph residual learning framework which learns an underlying mapping $\mathcal{H}$ based on the original graph convolution mapping $\mathcal{F}$. Assume that $\mathcal{G}_l$ is transformed by $\mathcal{F}$, we can obtain $\mathcal{G}_{l+1}$ with

$$\begin{aligned}\mathcal{G}_{l+1} &= \mathcal{H}(\mathcal{G}_l, \mathcal{W}_l) \\ &= \mathcal{F}(\mathcal{G}_l, \mathcal{W}_l) + \mathcal{G}_l = \mathcal{G}_{l+1}^{res} + \mathcal{G}_l,\end{aligned} \tag{4}$$

where the residual mapping $\mathcal{F}$ learns to take a graph as input and obtain a residual graph representation $\mathcal{G}_{l+1}^{res}$ for the next layer. $\mathcal{W}_l$ is the learnable parameters at layer l. A building block is shown in Fig. 4. With the introduced residual learning mechanism, RGCN enables reliable converge in training and achieves superior performance. Fig. 3 illustrates the detailed structure of our proposed Attention-GCN and Residual-GCN.

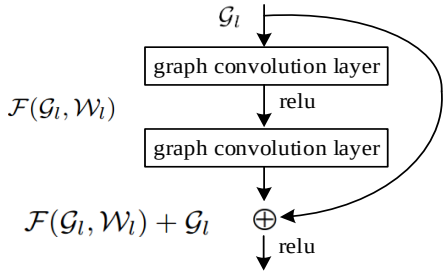


Fig. 4. A residual block of Residual-GCN.

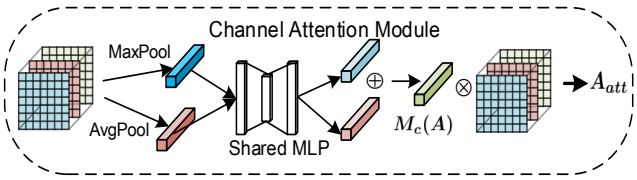


Fig. 5. Overview of the Channel Attention Module.

C. Deep LIFT Maps

Recently advanced image annotation methods [22], [33], [35] usually focus on locating attention regions in the convolutional feature map. The method in [24] employs GCN to transform label embedding into inter-dependent object classifiers. These approaches cannot decompose the traditional global features (for all labels) into label-specific features (for specific label), thus fail to establish the explicit and exact correspondence between visual region and semantic label. Our proposed model provide a route to explicitly extract label-specific features by flexibly exploiting the correlations between the local image region and each label embedding. Specifically, the multi-channel image features are extracted by ResNet-101 [13] while the label embeddings are obtained with AGCN and RGCN to explore the complex semantic dependency among labels.

The main goal of Deep-LIFT is to relate visual regions with specific label, thus provides fine-grained spatial attention and label-specific features. To this end, we try to learn a correlation matrix for each label. Specifically, this correlation matrix, named deep label-specific feature map $\boldsymbol{A} \in \mathbb{R}^{C\times M}$ in the following is obtained with

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{F}, \tag{5}$$

where $\boldsymbol{L} \in \mathbb{R}^{C\times D}$ denotes the label re-embedding matrix from GCN, and $\boldsymbol{F} \in \mathbb{R}^{D\times M}$ is the image visual feature map, as shown in Fig. 2. The deep label-specific feature map is obtained with the inner product of each label embedding and visual feature of each spatial location, thus, each value in deep LIFT map measures the similarity between one specific label and one specific local visual region (refer to "Construction of Deep LIFT Map" in Fig. 2).

D. Deep Attention LIFT Map

With the learned deep LIFT maps, the relationship between each label and the visual region is well explored. Furthermore, to reduce the effect from the Deep-LIFT maps of negative labels, we introduce a channel attention module (CAM) [26]. Since each convolutional feature map can be considered as a feature detector [39], our channel attention module aims to focus on the channels which correspond to positive labels for an image.

To obtain the channel attention efficiently, we first re-consider the deep LIFT maps $\boldsymbol{A} \in \mathbb{R}^{C\times M}$ as a order-3 tensor $\boldsymbol{A} \in \mathbb{R}^{C\times H\times W}$, where $M = H \times W = 14 \times 14$. Then average-pooling and max-pooling strategies are adopted for aggregating spatial information of the deep LIFT maps $\boldsymbol{A}$ from different views, generating two types of representations:

$\boldsymbol{A_{avg}}$ and $\boldsymbol{A_{max}}$. Then, these two types of representations are integrated to produce the final feature vectors by a shared network, which is composed of multi-layer perceptron (MLP) with one hidden layer. Accordingly, the channel attention map $\boldsymbol{M_c} \in \mathbb{R}^{C\times 1\times 1}$ is obtained by

$$\boldsymbol{M_c}(\boldsymbol{A}) = \sigma(MLP(\boldsymbol{A_{avg}}) + MLP(\boldsymbol{A_{max}})), \quad (6)$$

where σ denotes the sigmoid function. Once the channel attention map $\boldsymbol{M_c}$ obtained, we derive the deep-LIFT maps with channel attention with

$$\boldsymbol{A_{att}} = \boldsymbol{M_c}(\boldsymbol{A}) \odot \boldsymbol{A}, \quad (7)$$

where $\odot$ denotes element-wise multiplication and $\boldsymbol{A_{att}} \in \mathbb{R}^{C\times H\times W}$. Fig. 5 illustrates the structure of the introduced channel attention module (CAM).

E. Object Function

After obtaining the deep attention LIFT maps $\boldsymbol{A_{att}}$, average-pooling and max-pooling operations are performed to produce two types of vectors. Then we concatenate these two vectors as input to a multi-layer perceptron (MLP) with one hidden layer to obtain the final multi-label predictions $\hat{\mathbf{y}} \in \mathbb{R}^C$. Assume that a training image with ground-truth labels $\mathbf{y} \in \mathbb{R}^C$, where $y^i = 1(0)$ indicate the i^{th} label is (not) associated with this image. The model is trained in the end-to-end manner guiding by the the cross-entropy classification loss as follows:

$$\mathcal{L} = \sum_{c=1}^{C} y^c \log(\sigma(\hat{y}^c) + (1-y^c)\log(1-\sigma(\hat{y}^c)), \quad (8)$$

where σ is the sigmoid activation function.

IV. Experiments

We conduct experiments on two benchmark multi-label image datasets: MS-COCO [40] and VOC 2007 [41]. We first provide the implementation details and evaluation metrics, and then compare our experimental results with state-of-the-arts. Finally, ablation studies and visualization are presented to investigate the effectiveness of our model.

A. Implementation and Evaluation

We employ ResNet-101 [13] as the feature extractor, which is pre-trained on ImageNet [6]. To obtain the label word embedding as the input of GCN, the GloVe [37] model is introduced and trained on the Wikipedia, which produces a 300-dimension vector for each label. Moreover, we obtain the word embeddings of classes with multiple words by averaging their word embeddings. The AGCN consists of two graph convolution layers with output dimensionality of 1024 and 2048, and the RGCN contains one residual block and much deeper structure with the number of output channel as 512→1024→1024→1024→2048. We set the threshold by $\tau = 0.5$ to produce the label adjacency matrix $\boldsymbol{A}$. The input images are randomly cropped and resized into 448×448 with random horizontal flips for data augmentation. The batch size is set as 32 with the momentum being 0.9 and weight decay being 0.0001. The LeakyReLU [42] activation function is applied with the negative slope of 0.2, and the initial learning rate is set as 0.01, which decays by a factor of 10 for every 30 epochs. The total size of the trainable parameters of our proposed Deep-LIFT is 51.51M, where the parameters of the backbone network ResNet-101 is about 42.50M. We implement the model by PyTorch on 4 NVIDIA Geforce GTX TITAN Xp with 12GB GPU memory.

Following [21], [33], [43], [24], for fair comparison with these approaches, we also employ the overall precision, recall, F1 (OP, OR, OF1) and the per-class precision, recall, F1 (CP, CR, CF1) for performance evaluation, which are defined as below

$$\begin{aligned} \text{OP} &= \frac{\sum_i N_i^c}{\sum_i N_i^p}, & \text{CP} &= \frac{1}{C}\frac{\sum_i N_i^c}{\sum_i N_i^p} \\ \text{OR} &= \frac{\sum_i N_i^c}{\sum_i N_i^g}, & \text{CR} &= \frac{1}{C}\frac{\sum_i N_i^c}{\sum_i N_i^g} \\ \text{OF1} &= \frac{2\times \text{OP}\times \text{OR}}{\text{OP}+\text{OR}}, & \text{CF1} &= \frac{2\times \text{CP}\times \text{CR}}{\text{CP}+\text{CR}}, \end{aligned}$$

where C is the number of labels, N_i^c is the number of images that are correctly predicted for the i-th label, N_i^p is the number of predicted images for the i-th label, N_i^g is the number of ground truth images for the i-th label. For fair comparison with other methods [33], [43], [24], we also report the results of top-3 highest-ranked labels and the mean average precision (mAP) over all categories. For each image, if the predicted label confidence for one label is larger than 0.5, then the label is considered as positive. These metrics evaluate the performance of multi-label predictor from diverse aspects. Generally, mAP, OF1, and CF1 are relatively more important for performance evaluation.

B. Experimental Results

1) Results on MS-COCO: MS-COCO [40] is usually used for object detection, and recently, it is also widely applied for image annotation. MS-COCO contains 82,081 images as the training set and 40,137 images as the validation set, covering 80 common object categories with about 2.9 labels per image. The ground truth labels of the test set is unavailable since it is used for the annual visual challenge, so the performance of all the methods are evaluated on the validation set.

The results of MS-COCO are presented in Table I. We compare our method with the state-of-the-art approaches, including WARP [2], CNN-RNN [21], RLSD [44], RNN-Attention [22], Order-Free RNN [35], KD-WSD [45], SRN [33], ResNet-101 [13], Multi-Evidence [43], ACfs [23] and ML-GCN [24]. We try our best to tune the parameters of all the above compared methods to obtain the best performance according to the suggested ways in their literatures. For the proposed Deep-LIFT, we report the results based on AGCN ("Deep-LIFT (AGCN)") and RGCN ("Deep-LIFT(RGCN)"), respectively. It is observed that our Deep-LIFT (AGCN) and Deep-LIFT (RGCN) both obtain competitive performance against state-of-the-art methods. Furthermore, Deep-LIFT (AGCN) performs

TABLE I
COMPARISONS WITH STATE-OF-THE-ART METHODS ON MS-COCO DATASET. THE PERFORMANCE OF OUR DEEP-LIFT BASED ON TWO VARIANT GCN (AGCN AND RGCN) ARE REPORTED. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY.

Mehtods	All							Top-3					
	mAP	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
WARP [2]	58.8	61.2	46.3	55.6	65.6	52.9	58.3	59.3	52.5	55.7	59.8	61.4	60.7
CNN-RNN [21]	61.5	64.8	50.4	58.6	68.7	56.8	61.1	66.3	56.7	61.4	69.1	66.3	68.6
RLSD [44]	63.4	66.4	50.9	60.8	69.3	56.4	62.8	67.6	57.2	62.0	70.1	63.4	66.5
RNN-Attention [22]	74.3	76.6	61.6	69.3	79.2	67.8	73.6	79.1	58.7	67.4	84.0	63.0	72.0
Order-Free RNN [35]	72.8	75.9	60.5	68.2	77.6	66.5	71.9	73.8	54.5	61.7	78.3	62.0	68.7
KD-WSD [45]	74.6	77.3	61.8	69.1	79.5	67.6	73.4	80.9	55.8	64.8	84.2	61.6	72.2
SRN [33]	76.0	80.1	63.9	70.6	81.7	68.8	74.6	84.4	57.4	66.3	86.3	61.5	71.2
ResNet-101 [13]	76.1	79.4	64.9	71.5	82.3	69.4	75.3	83.1	57.9	68.6	88.5	61.4	72.6
Multi-Evidence [43]	78.1	80.1	69.5	73.5	84.2	71.7	77.6	83.2	61.3	69.6	88.4	63.5	73.8
ACfs [23]	78.6	80.3	69.2	73.2	80.9	74.1	77.4	86.2	60.4	69.3	87.6	64.3	74.3
ML-GCN(Binary) [24]	80.2	81.5	69.4	75.1	83.6	74.2	78.4	85.6	61.4	71.4	88.5	65.3	75.3
ML-GCN(Re-weighted) [24]	81.7	83.3	70.1	76.2	86.1	74.5	79.9	86.9	62.5	72.7	90.3	65.3	75.3
Deep-LIFT(RGCN)	81.8	83.6	70.0	76.6	85.5	74.3	79.9	86.0	63.0	73.1	90.1	65.4	75.8
Deep-LIFT(AGCN)	82.2	85.4	70.9	77.5	86.4	74.6	80.1	87.5	63.2	73.4	90.6	66.0	76.4

TABLE II
COMPARISONS WITH STATE-OF-THE-ART METHODS IN TERMS OF AP AND MAP ON VOC 2007 DATASET. THE PERFORMANCE OF OUR DEEP-LIFT BASED ON TWO VARIANT GCN (AGCN AND RGCN) ARE REPORTED. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
CNN-RNN [21]	96.8	85.5	94.4	92.4	62.3	82.3	89.2	94.4	64.5	83.7	70.2	92.2	91.4	84.3	93.2	59.4	93.4	75.7	99.5	78.8	84.2
RMIC [46]	97.0	90.6	94.3	93.8	71.4	90.2	93.6	93.5	73.2	86.6	73.5	94.2	91.8	92.1	93.3	62.7	91.0	76.5	96.8	79.7	86.8
RLSD [44]	96.8	92.8	93.4	94.3	71.4	92.4	94.3	95.3	74.5	90.2	74.4	95.6	96.4	92.2	97.7	66.4	93.2	73.2	97.2	87.2	88.4
VeryDeep [7]	98.7	94.6	96.7	95.5	69.8	90.5	93.7	95.4	73.8	85.8	87.8	96.2	96.1	93.2	97.4	70.1	92.3	80.1	98.2	87.2	89.7
ResNet-101 [13]	98.8	97.2	97.4	96.1	65.2	91.4	96.3	97.8	74.4	80.2	85.4	98.2	96.6	95.6	98.6	70.3	87.7	80.3	98.4	89.1	89.8
FeV+LV [47]	97.5	97.2	96.5	94.3	73.8	93.2	96.4	95.1	73.4	90.5	82.6	95.1	97.4	95.6	98.7	77.2	88.4	78.2	98.4	89.3	90.4
HCP [48]	98.6	97.3	98.1	94.8	74.7	94.2	95.1	96.9	73.2	90.2	80.1	96.6	96.2	94.4	96.2	78.4	94.5	76.4	97.7	91.2	90.7
RNN-Attention [22]	98.7	97.2	95.8	96.1	74.9	92.5	96.3	96.8	76.3	91.8	87.7	96.1	97.6	93.6	97.9	81.5	93.4	82.2	98.7	89.1	91.7
Atten-Reinforce [49]	98.7	97.3	96.8	94.5	75.8	92.4	95.9	97.1	77.9	92.3	87.1	96.2	95.9	93.1	97.9	81.7	93.2	83.1	98.1	89.5	91.7
ML-GCN(Binary) [24]	99.4	98.1	97.9	97.6	78.2	92.3	97.2	97.4	79.2	94.3	86.5	97.4	97.9	97.1	98.7	84.2	95.3	83.0	98.3	90.4	93.1
ML-GCN(Re-weighted) [24]	99.2	97.8	98.0	97.7	78.2	95.0	97.4	96.6	80.2	94.4	85.9	98.0	97.4	96.1	98.7	85.0	96.2	81.9	98.4	93.2	93.3
Deep-LIFT(RGCN)	99.7	98.3	98.2	97.8	80.7	94.5	97.4	97.6	80.6	94.2	83.4	97.7	97.5	95.6	98.8	84.4	95.8	82.1	99.0	94.0	93.4
Deep-LIFT(AGCN)	99.5	98.3	98.3	97.6	81.2	93.2	96.7	97.4	81.9	94.3	85.7	97.6	98.1	95.5	98.8	83.8	96.3	80.0	99.1	94.0	93.4

much better than other state-of-the-arts in terms of almost all metrics, validating the effectiveness and superiority of the proposed model.

TABLE III
COMPARISONS WITH DIFFERENT TYPES OF GCN IN OUR MODEL.

Architecture	MS-COCO					VOC
	All			Top-3		All
	mAP	CF1	OF1	CF1	OF1	mAP
Original-GCN	79.6	74.8	78.7	71.4	75.3	93.1
Residual-GCN	81.8	76.6	79.9	73.1	75.8	93.4
Weighted-GCN	80.3	75.0	78.6	71.7	75.4	93.2
Attention-GCN	**82.2**	**77.5**	**80.1**	**73.4**	**76.4**	**93.4**

2) Results on VOC 2007: PASCAL VOC 2007 [41] is the most widely used benchmark for image multi-label classification, which contains 9,963 images with 20 object categories. The data are divided into training, validation and test sets. Following [22], [49], [24], We train our model on the *trainval* sets (5011 images), and evaluate the performance on the test set (4952 images).

We compare our method with the following state-of-the-art approaches: CNN-RNN [21], RMIC [46], RLSD [44], VeryDeep [7], ResNet-101 [13], FeV+LV [47], HCP [48], RNN-Attention [22], Atten-Reinforce [49] and ML-GCN [24]. For comparison convenience, we also report the results in terms of average precision (AP) and mean average precision (mAP).

As shown in Table II, we report the quantitative experimental results of different methods on the VOC 2007 dataset. The previous competitive methods include RNN-Attention [22], Atten-Reinforce [49] and ML-GCN [24], which achieve 91.7%, 91.7% and 93.3% in terms of mAP, respectively. Similar to MS-COCO, we report the results based on Attention-GCN ("Deep-LIFT (AGCN)") and Residual-GCN ("Deep-LIFT (RGCN)"), respectively. We can clear find that the proposed Deep-LIFT (AGCN) and Deep-LIFT (RGCN) both outperform the compared state-of-the-arts. In terms of the

average precision (AP), our Deep-LIFT (AGCN) and Deep-LIFT (RGCN) alternately outperforms others on most classes.

C. Ablation Studies

1) Different types of GCN: To explore the correlation among different labels, Attention-GCN and Residual-GCN are developed. We evaluate the performance of our model with different types of GCN. Specifically, we compare Residual-GCN and Attention-GCN with the original GCN [36] and Weighted-GCN, respectively. For the Weighted-GCN, the cosine similarity is employed to replace each non-zero element in adjacency matrix $\boldsymbol{A}$. Table III shows the results using different types of GCN. It can be observed that the proposed model with Residual-GCN obtains better performance than the original GCN, and the proposed model with Attention-GCN obviously outperforms Weighted-GCN. This validates the superiority of learning attention weight automatically for our model.

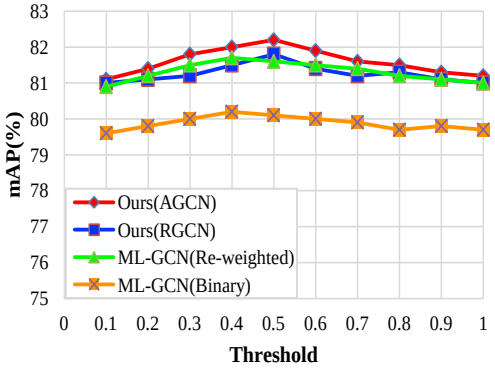
2) Channel Attention Module: In the channel attention module (CAM), average-pooling and max-pooling are adopted for aggregating spatial information of the Deep-LIFT maps. To investigate the effectiveness of the introduced channel attention module, we report the result of our Deep-LIFT (AGCN) without the channel attention module, and compare it with 3 variants of channel attention: average pooling, max pooling, and joint both these two poolings. The experimental results are shown in Table IV. The proposed models with three variants of channel attention enjoy better performance than that without CAM. Moreover, the Deep-LIFT equipped with both average-pooling and max-pooling CAM obtains the best performance. The results clearly demonstrate the effectiveness of channel attention module.

TABLE IV
COMPARISON OF DIFFERENT ARCHITECTURES OF THE CAM.

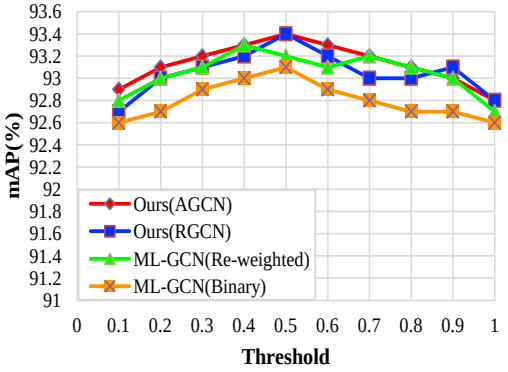
Architecture	MS-COCO					VOC
	All			Top-3		All
	mAP	CF1	OF1	CF1	OF1	mAP
Ours without CAM	80.5	75.1	78.7	71.8	75.4	92.9
Ours+AvgPool	81.0	75.6	79.2	72.2	75.7	93.1
Ours+MaxPool	80.6	75.2	78.8	72.1	75.4	93.0
Ours+AvgPool+MaxPool	**82.2**	**77.5**	**80.1**	**73.4**	**76.4**	**93.4**

TABLE V
COMPARISON WITH DIFFERENT NUMBER OF LAYERS FOR AGCN.

Architecture	MS-COCO					VOC
	All			Top-3		All
	mAP	CF1	OF1	CF1	OF1	mAP
2-layer	**82.2**	**77.5**	**80.1**	**73.4**	**76.4**	**93.4**
3-layer	81.7	76.2	79.9	72.7	76.3	93.0
4-layer	80.9	75.3	79	71.8	75.5	92.8
5-layer	80.1	74.5	78.3	71.3	74.9	92.0



(a) Comparisons on MS-COCO.



(b) Comparisons on VOC 2007.

Fig. 6. Comparisons under different values of τ on MS-COCO and VOC 2007.

3) The deeper layers of Attention-GCN and Residual-GCN: We investigate the effect of using different number of layers for Attention-GCN and Residual-GCN. For Attention-GCN, we conduct experiments on 2-layer ($1024 \rightarrow 2048$), 3-layer ($512 \rightarrow 1024 \rightarrow 2048$), 4-layer ($512 \rightarrow 1024 \rightarrow 1024 \rightarrow 2048$) and 5-layer ($512 \rightarrow 1024 \rightarrow 1024 \rightarrow 1024 \rightarrow 2048$), respectively. According to the results in Table V, the performance of our Deep-LIFT (AGCN) drops slowly as the number of graph convolution layers increases. For Residual-GCN, we compare the performance on 1, 2, 3 and 4 residual blocks. According to Table VI, the more residual blocks lead to a little worse performance. The underlying reason may be that the original node embeddings tend to be over-smoothed as the propagation between nodes is accumulated using more convolution layers [24].

4) Effect of different threshold values τ: Following [24], we construct the original label adjacent matrix for GCN in a data-driven way, which measures the degree of correlation between labels by computing their co-occurrence conditional probability within dataset and then sets the threshold τ to filter possible noisy links. For fair comparison with ML-GCN [24], we vary the values of the threshold τ from 0.1 to 1 and report the performance in terms of mAP. According to Fig. 6, our Deep-LIFT(AGCN) and Deep-LIFT(RGCN) perform slightly

Fig. 7. Visualization results of our model with Grad-CAM [50]. We compare the visualization results of Deep-Attention-LIFT (DA-LIFT) with Deep-LIFT (D-LIFT) and the baseline (ResNet-101) in the target layer of deep attention LIFT map, deep LIFT map and "conv5_x". The sigmoid scores for top-3 highest-ranked target classes are also shown for each test image. The labels in black and gray indicate positive and negative classes of the image respectively.

better than ML-GCN (Re-weighted) and ML-GCN (Binary). Moreover, we note that the optimal value of τ for our model is 0.5 on both MS-COCO and VOC 2007, which indicates that it is reasonable to filter out noisy edges with appropriate probability.

D. Visualization Analysis

For qualitative analysis, we visualize the learned deep LIFT maps and deep attention LIFT maps by using the Grad-CAM [50], which provides one way to investigate the capability in establishing the correspondence between visual regions and labels and improving the multi-label classification performance. Grad-CAM is a recently proposed visualization method which uses the class-specific gradient information flowing into a convolutional layer of CNN to calculate the importance of the different spatial locations in an image. For multi-label image classification, Grad-CAM calculates the gradients with respect to each class in an image and then shows the attention regions corresponding to each label. We compare the visualization results of Deep-Attention-LIFT (AGCN), Deep-LIFT (AGCN) and the baseline (ResNet-101) in the target layer of deep attention LIFT map, deep LIFT map and "conv5_x", where the deep attention LIFT map and deep LIFT map can be considered as an output map of a convolution layer.

TABLE VI
COMPARISON WITH DIFFERENT NUMBER OF BLOCK FOR RGCN.

Architecture	MS-COCO					VOC
	All			Top-3		All
	mAP	CF1	OF1	CF1	OF1	mAP
1-block	**81.8**	**76.6**	**79.9**	**73.1**	**75.8**	**93.4**
2-block	81.1	75.3	78.9	72.0	75.5	92.7
3-block	80.4	74.8	78.6	71.3	75.2	91.2
4-block	79.6	72.4	76.5	70.6	73.1	90.6

Fig. 7 presents the visualization results. We show some images containing 3 (the first four lines) or 2 (the last four line) object classes on VOC 2007 or MS-COCO. To compare the prediction performance, the sigmoid scores for the top-3 highest-ranked target classes are shown in the figure. In Fig. 7, we can clearly find that our Deep-Attention-LIFT (Ours(DA-LIFT)) describes the target object regions of each existing class more accurately than our Deep-LIFT (Ours(D-LIFT)) and the baseline (ResNet-101). Moreover, the predicted scores for the positive classes of Ours(DA-LIFT) and Ours(D-LIFT) are much higher than those of the baseline (ResNet-101). For the negative labels, our Deep-Attention-LIFT (Ours(DA-LIFT)) tends to obtain the lower predicted scores. Intuitively, the proposed Deep-LIFT and Deep-Attention-LIFT can accurately capture the correlation between visual regions and specific labels so as to significantly improve the performance on multi-label classification.

V. CONCLUSION

In this work, we propose a novel Deep Label-Specific Feature (Deep-LIFT) learning model for image annotation. Different from existing end-to-end image annotation methods, the proposed Deep-LIFT explicitly decomposes the global feature sets into label-specific features to better exploit discrimination information from different classes. With extended graph convolutional network (GCN), our model can further capture the correlations among labels. Extensive experiments, including quantitative, qualitative and ablation results on benchmark datasets, validate the effectiveness of the proposed model over existing state-of-the-art approaches.

REFERENCES

[1] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern recognition*, vol. 37, no. 9, pp. 1757–1771, 2004. 1, 2

[2] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," *arXiv preprint arXiv:1312.4894*, 2013. 1, 2, 5, 6

[3] J. Shao, C.-C. Loy, K. Kang, and X. Wang, "Slicing convolutional neural network for crowd video understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5620–5628. 1

[4] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *European Conference on Computer Vision*. Springer, 2016, pp. 684–700. 1

[5] Z. Ge, D. Mahapatra, S. Sedai, R. Garnavi, and R. Chakravorty, "Chest x-rays classification: A multi-label and fine-grained problem," *arXiv preprint arXiv:1807.07247*, 2018. 1

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. 1, 5

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 1, 2, 6

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708. 1

[9] H. Liu, X. Li, and S. Zhang, "Learning instance correlation functions for multilabel classification," *IEEE transactions on cybernetics*, vol. 47, no. 2, pp. 499–510, 2016. 1

[10] Y. Zhou, J. He, and H. Gu, "Partial label learning via gaussian processes," *IEEE transactions on cybernetics*, vol. 47, no. 12, pp. 4443–4450, 2016. 1

[11] C. Zhang, Z. Yu, H. Fu, P. Zhu, L. Chen, and Q. Hu, "Hybrid noise-oriented multilabel learning," *IEEE transactions on cybernetics*, 2019. 1

[12] J. Du and C.-M. Vong, "Robust online multilabel learning under dynamic changes in data distribution with labels," *IEEE transactions on cybernetics*, vol. 50, no. 1, pp. 374–385, 2019. 1

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 1, 2, 4, 5, 6

[14] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258. 1

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. 1

[16] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500. 1

[17] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856. 1

[18] X. Li, F. Zhao, and Y. Guo, "Multi-label image classification with a probabilistic label enhancement model." in *UAI*, vol. 1, 2014, p. 3. 1, 2

[19] Q. Li, M. Qiao, W. Bian, and D. Tao, "Conditional graphical lasso for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2977–2986. 1, 2

[20] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori, "Learning structured inference neural networks with label relations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2960–2968. 1, 2

[21] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2285–2294. 1, 2, 5, 6

[22] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 464–472. 1, 2, 4, 5, 6

[23] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang, "Visual attention consistency under image transforms for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 729–739. 1, 5, 6

[24] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5177–5186. 1, 2, 3, 4, 5, 6, 7

[25] M.-L. Zhang and L. Wu, "Lift: Multi-label learning with label-specific features," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 1, pp. 107–120, 2014. 2

[26] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19. 2, 4

[27] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007. 2

[28] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, p. 333, 2011. 2

[29] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005, pp. 195–200. 2

[30] Y. Guo and S. Gu, "Multi-label classification using conditional dependency networks," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011. 2

[31] X. Xue, W. Zhang, J. Zhang, B. Wu, J. Fan, and Y. Lu, "Correlative multi-label multi-instance image annotation," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 651–658. 2

[32] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C. Frank Wang, "Multi-label zero-shot learning with structured knowledge graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1576–1585. 2

[33] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5513–5522. 2, 4, 5, 6

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826. 2

[35] S.-F. Chen, Y.-C. Chen, C.-K. Yeh, and Y.-C. F. Wang, "Order-free rnn with visual attention for multi-label classification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2, 4, 5, 6

[36] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016. 3, 7

[37] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543. 3, 5

[38] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017. 3

[39] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833. 4

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755. 5

[41] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010. 5, 6

[42] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3. 5

[43] W. Ge, S. Yang, and Y. Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1277–1286. 5, 6

[44] J. Zhang, Q. Wu, C. Shen, J. Zhang, and J. Lu, "Multilabel image classification with regional latent semantic dependencies," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2801–2813, 2018. 5, 6

[45] Y. Liu, L. Sheng, J. Shao, J. Yan, S. Xiang, and C. Pan, "Multi-label image classification via knowledge distillation from weakly-supervised detection," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 700–708. 5, 6

[46] S. He, C. Xu, T. Guo, C. Xu, and D. Tao, "Reinforced multi-label image classification by exploring curriculum," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 6

[47] H. Yang, J. Tianyi Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai, "Exploit bounding box annotations for multi-label object recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 280–288. 6

[48] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "Hcp: A flexible cnn framework for multi-label image classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1901–1907, 2015. 6

[49] T. Chen, Z. Wang, G. Li, and L. Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 6

[50] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626. 8