MULTI-LABEL IMAGE RECOGNITION WITH JOINT CLASS-AWARE MAP DISENTANGLING AND LABEL CORRELATION EMBEDDING

Zhao-Min Chen[1,2] *Xiu-Shen Wei*[2,*] *Xin Jin*[2] *Yanwen Guo*[1,3,*]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Megvii Research Nanjing, Megvii Technology, China
[3]Science and Technology on Information Systems Engineering Laboraty, China
{chenzhaomin123, weixs.gm}@gmail.com, jinxin@megvii.com, ywguo@nju.edu.cn

ABSTRACT

Multi-label image recognition is a fundamental but challenging computer vision task. Great progress has been achieved by exploring the label correlation among these multiple labels which is the most crucial issue for multi-label recognition. In this paper, we propose a unified deep learning framework to jointly disentangle class-specific maps corresponding to discriminative category-wise information and then evaluate the label co-occurrence of these maps. Specifically, after obtaining the general deep image features and conducting multi-label classification, we employ the classification weights to reform the feature maps into class-aware disentangled maps (CADMs). Then, based on CADMs, we first transfer them into label vectors and then formulate the label correlation dependency from an embedding perspective. The whole model is driven by both the classification loss and the label correlation embedding loss, which is end-to-end trainable with only image-level supervisions. Extensive quantitative results of two benchmark multi-label image datasets show our model consistently outperforms other competing methods by a large margin. Meanwhile, qualitative analyses also demonstrate our model can effectively capture relatively pure class-aware maps and model label correlation dependency as well.

Index Terms— Multi-label image recognition, label correlation, CNNs, class-aware disentangled maps (CADMs).

1. INTRODUCTION

Recognizing multiple labels of image is an important and practical problem in the computer vision and multimedia field, as real-world images always contain rich and diverse semantic information. Multi-label image recognition is general and has various applications, such as scene recognition [1], human attribute recognition [2], face alignment [3], retail checkout recognition [4], etc. An important but challenging issue for multi-label recognition is to identify and recover the co-occurrence of multiple labels, such that satisfactory prediction accuracy can be expected.

A simple and straightforward method for multi-label recognition is to train one binary deep classifier for each label. However, the major challenge of learning from multi-label data lies in the potentially tremendous-sized output space. To deal with the challenge of such a huge output space, a common practice is to explore the label correlation to facilitate the learning process [5, 6, 7]. In the literature, Gong *et al.* [8] evaluated various loss functions and found that weighted approximate ranking loss worked best with deep CNNs. Additionally, Hu *et al.* [9] proposed to employ structured inference neural network to model the label correlation of multiple labels. Li *et al.* [10] leveraged probabilistic graphical models to capture the label correlation dependency.

Recently, researchers attempted to apply attention mechanisms to discover the label correlation among different attentional regions, *e.g.*, [11, 12]. In [12], the authors developed the spatial regularization net to focus on the objectiveness regions, and further learned label correlation of these regions by self-attention. While, Wang *et al.* [11] proposed the spatial transformer to first capture the objectiveness regions and then use LSTMs to handle the label correlation. Despite the good improvements obtained, existing methods still have limitations on identifying and recovering the co-occurrence of multiple labels, more concretely: 1) disentangling class-specific image regions and 2) further evaluating their corresponding label co-occurrence jointly. If these two processes can be performed well, it will significantly boost multi-label recognition performance.

In this paper, we propose a unified multi-label image recognition framework, which consists of two crucial modules aiming at the two aforementioned processes. The architecture of our model is illustrated in Fig. 1. After obtaining the general and holistic image feature, the first module can disentangle the

Z.-M. Chen's contribution was made when he was an intern in Megvii Research Nanjing. This research was supported by National Key R&D Program of China (No. 2017YFA0700800), the National Natural Science Foundation of China under Grants 61772257 and 61672279.

*Corresponding authors

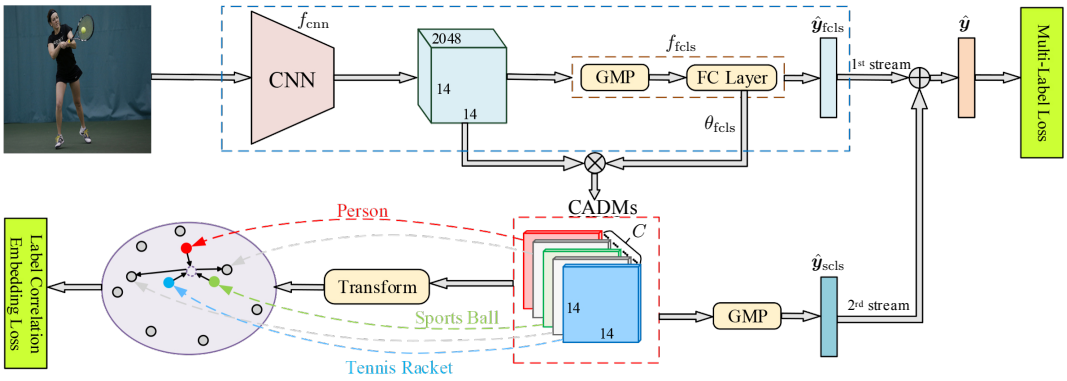

Fig. 1. Overall framework of our proposed model for multi-label image recognition. The input image is firstly fed to conventional CNNs for learning the image representations (*i.e.*, $\mathbf{X}$) of the final convolutional layer. After that, we utilize global max-pooling to obtain the image-level features, and then conduct multi-label classification (*i.e.*, f_{fcls}) based on these features. In the following, we employ the classification weights (*i.e.*, θ_{fcls}) on $\mathbf{X}$ for generating the class-aware disentangled maps (CADMs), which could disentangle class-specific image regions/maps corresponding to these multiple image labels. Based on CADMs, label correlation information is embedded in the label vector space. It is the key and could be benefit to multi-label image recognition.

class-aware maps from global image-level representation in a simple but effective way, *i.e.*, using the classification weights to reform the feature maps into class-aware disentangled maps. Each map of the so called class-aware disentangled maps corresponds to one specific class/label meaning of the multiple labels. It can relatively purely reflect the semantic information of its specific label and meanwhile associate with spatial contexts (cf. Fig. 2). Particularly, the number of class-aware disentangled maps (CADMs) is equal to the number of labels. The second module of our model is based on the obtained CADMs, which is designed for modeling the multiple label co-occurrence in a more explicit way, as shown in Fig. 3. We first transform the CADMs into a label vector. In the label vector space, for multi-label recognition, it assumes that the relevant labels (label vectors) should be closed to each other and form a dummy cluster. While, the irrelevant/negative labels should be apart from the "positive" dummy cluster. Please note that the discriminative ability of CADMs will not be destroyed since the embedding operation is performed in the label vector space. Then, we formulate it as the label correlation embedding loss function. Driven by both traditional multi-label recognition loss and our label correlation embedding loss, our model can be trained in an end-to-end fashion with only image-level supervisions, which does not require any additional annotations. In a nutshell, our method explores the spatial correlation of labels and utilizes it as an additional cue for classification. Particularly, the label correlation embedding layers can essentially learn the correlation among class-aware disentangled maps (*e.g.*, the activations of "`snowboard`" will be boosted if there is a "`person`" on top of it).

The main contributions of this paper are three-fold: (1) We propose a unified multi-label image recognition framework for jointly identifying and recovering the label co-occurrence of multiple labels. The proposed model is end-to-end trainable with only image-level supervisions. (2) We devise two functional modules of the proposed model, *i.e.*, the class-aware map disentangling and label correlation embedding modules, for capturing the class-specific information with spatial contexts and modeling the label co-occurrence, respectively. (3) We conduct comprehensive experiments on two popular multi-label image recognition datasets, and our proposed model consistently achieves superior performance over competing state-of-the-arts methods on these datasets.

2. PROPOSED METHOD

2.1. Preliminary

Let $\mathbf{I}$ denote an input image with ground-truth labels $\boldsymbol{y} = \left[y^1, y^2, \ldots, y^C\right]^\top$, where y^c is a binary indicator and C is the number of all possible labels in the dataset. $y^c = 1$ presents image $\mathbf{I}$ is tagged with label c and $y^c = 0$ otherwise. For multi-label image recognition, the goal is to predict the multi-label vector $\hat{\boldsymbol{y}}$ for a test input $\hat{\mathbf{I}}$.

2.2. Model overview

As shown in Fig. 1, for an input multi-label image, traditional convolutional neural networks are employed to learn a holistic image representation. The obtained deep image representations will be processed by the following modules. Concretely, the activations of a convolution layer can be formulated as an order-3 tensor $\mathbf{X}$ with $d \times h \times w$ elements, which includes a set of 2-D feature maps. These feature maps are embedded with rich spatial information, and are also known to obtain

mid- and high-level information [13]. In experiments, following [12, 14], we use ResNet-101 [15] as our base model. Thus, if an image with the 448×448 resolution is the input, we can obtain $2048 \times 14 \times 14$ feature maps from the "`conv5_x`" layer by

$$\mathbf{X} = f_{\text{cnn}}(\mathbf{I}; \theta_{\text{cnn}}) \in \mathbb{R}^{d \times h \times w}, \tag{1}$$

where $\mathbf{X}$ is the aforementioned feature maps, and θ_{cnn} indicates the parameters of CNNs. Specifically, here $h = 14$, $w = 14$, and $d = 2048$.

After that, for measuring the final prediction errors, we combine two predicted label confidences as the bi-stream aggregation. As illustrated in Fig. 1, for the label confidences of the first stream, we employ global max-pooling on $\mathbf{X}$ to obtain the image-level features, following with binary classification for each of the C labels:

$$\hat{\boldsymbol{y}}_{\text{fcls}} = f_{\text{fcls}}(\mathbf{X}; \theta_{\text{fcls}}) \in \mathbb{R}^{C}, \tag{2}$$

where $\hat{\boldsymbol{y}}_{\text{fcls}} = \left[\hat{y}^1_{\text{fcls}}, \hat{y}^2_{\text{fcls}}, \ldots, \hat{y}^C_{\text{fcls}}\right]^\top$, and each element of $\hat{\boldsymbol{y}}_{\text{fcls}}$ is a confidence score. For the second stream, we obtain its label confidences $\hat{\boldsymbol{y}}_{\text{scls}}$ via directly depth-wise global max-pooling on the class-aware disentangled maps (CADMs). These class-aware maps not only contain the local-level spatial contexts information (*i.e.*, activations), but also have the global-level class-specific semantic meaning. The detailed CADMs generation processing will be described in the next sub-section.

In the following, we aggregate these two label confidences as the final label prediction confidences by

$$\hat{\boldsymbol{y}} = \frac{1}{2}(\hat{\boldsymbol{y}}_{\text{fcls}} + \hat{\boldsymbol{y}}_{\text{scls}}) \in \mathbb{R}^{C}. \tag{3}$$

Finally, $\hat{\boldsymbol{y}}$ will be used to measure the prediction errors w.r.t. the ground-truth labels $\boldsymbol{y}$ as

$$\mathcal{L}_{\text{cls}} = \sum_{c=1}^{C} y^c \log(\sigma(\hat{y}^c)) + (1 - y^c) \log(1 - \sigma(\hat{y}^c)), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function.

In addition, beyond $\mathcal{L}_{\text{cls}}$, our model is also driven by another loss function, *i.e.*, $\mathcal{L}_{\text{lce}}$, for explicitly modeling the label co-occurrence cues (which will be elaborated in the following sub-section). Our final loss function is presented as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{lce}}. \tag{5}$$

Here, λ is a trade-off parameter, which is set to 0.5 in all the experiments.

2.3. Class-aware map disentangling

In this section, we elaborate our class-aware map disentangling module. This proposed module is designed to disentangle the class-specific maps from the deep representations. The disentangled maps could benefit to evaluate the label correlation and embed the label correlation information into the whole multi-label learning system.

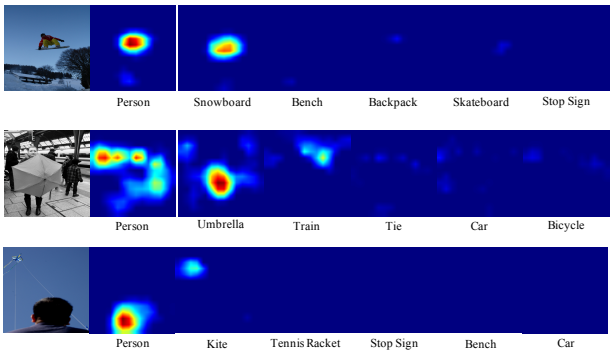

Fig. 2. Exampled images from the *MS-COCO* dataset with the CADMs. For each image, we first sort the summation activation values of every CADM in the descending order, and then present the class-aware maps in the same order. It is clear that positive labels correspond to strong activations in their own CADM, while negative labels almost activate nothing by comparison. (Best viewed in color.)

As aforementioned, based on $\mathbf{X}$, we globally max-pool the image representations into an image-level feature and then conduct one fully-connected layer $\theta_{\text{fcls}} \in \mathbb{R}^{d \times C}$ for classification. Inspired by [16], we can use θ_{fcls} to disentangle C class-specific maps from these distributed representations of $\mathbf{X}$ [17, 18]. However, different from the global average-pooling used in [16], here we employ global max-pooling for keeping the highlighted activations of small-scale objects which is usually emerged in multi-label images.

Concretely, for a given image, $\hat{\boldsymbol{y}}_{\text{fcls}}$ is the predicted label confidence via $f_{\text{fcls}}(\mathbf{X}; \theta_{\text{fcls}})$. θ^c_{fcls} denotes the classification weights w.r.t. the c-th label. From another perspective, θ^c_{fcls} can be treated as the filter to filter out class-specific discriminative information for the c-th label from $\mathbf{X}$. Here we omit the bias term since it has little to no impact on the classification performance.

We denote $\boldsymbol{A}_c$ as the corresponding class-aware disentangled map for class c, which can be obtained by

$$\boldsymbol{A}_c = {\theta^c_{\text{fcls}}}^\top \cdot \mathbf{X} \in \mathbb{R}^{h \times w}. \tag{6}$$

Each $\boldsymbol{A}_c$ is disentangled for its corresponding c-th label. Thus, by collecting all C disentangled maps, we obtain

$$\mathbf{A} = {\theta_{\text{fcls}}}^\top \cdot \mathbf{X} \in \mathbb{R}^{C \times h \times w}. \tag{7}$$

In fact, the class-aware disentangled maps (CADMs) $\mathbf{A}$ is simply a weighted linear sum of the presence of these visual patterns at different spatial locations. In Fig. 2, several qualitative results of CADMs for multi-label images are provided. As shown in that figure, each CADM corresponds to one specific and independent label meaning. Moreover, it is apparent to see that the positive label has more strong activations in its class-aware map, and the negative labels has much weaker

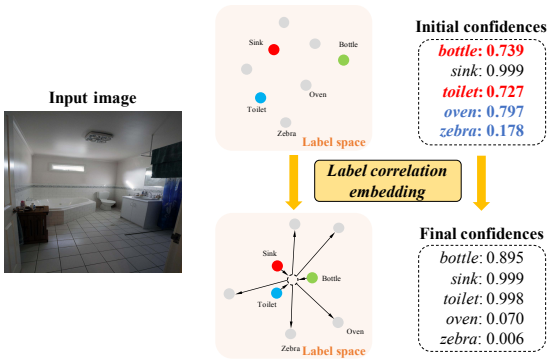

Fig. 3. Illustration of our proposed label correlation embedding for improving multi-label image recognition performance.

and even none activations. The observations verify that the class-aware map disentangling approach can both decouple label semantic information and localize class-specific regions at the same time.

2.4. Label correlation embedding

After disentangling, the obtained class-aware maps contain both the original image appearance cues and specific label semantic information. In order to capture label correlation, which is the most crucial thing for multi-label image recognition, we propose to explicitly model it via label correlation embedding in a metric learning fashion.

Metric learning is popular in face recognition [19], person re-identification [20] and vehicle re-identification [21]. In these previous metric learning work, they almost embedded the objective images into feature vectors in the *feature space*. However, different from them, our model embeds the class-aware region maps associating with an image $\mathbf{I}$ into a multi-dimensional *label space*, where each label corresponds to its fixed size label vector $\boldsymbol{a}_c$. Therefore, the co-occurrence of two related labels (*i.e.*, label vectors) can be measured by their distance in this label space. More intuitively, in the multi-label scenario, these correlated labels (*i.e.*, label vectors) could be clustered, while these uncorrelated labels should be apart from the dummy cluster, cf. Fig. 3.

Specifically, for obtaining the label vectors, we first flatten the class-aware disentangled map $\boldsymbol{A}_c$ into a single vector $f_{\text{flat}}(\boldsymbol{A}_c) \in \mathbb{R}^{1\times(h\times w)}$. Then, we introduce a non-linear transformation $f_{\text{embed}}(\cdot;\theta_{\text{embed}})$ on $f_{\text{flat}}(\boldsymbol{A}_c)$ for embedding it into $\boldsymbol{a}_c$ in the aforementioned label space:

$$\boldsymbol{a}_c = f_{\text{embed}}(f_{\text{flat}}(\boldsymbol{A}_c);\theta_{\text{embed}})\,, \tag{8}$$

where θ_{embed} is the embedding parameters. In experiments, $f_{\text{embed}}(\cdot;\theta_{\text{embed}})$ is a two-layer fully connected network with ReLU as its activation function.

Thus, the objective of label correlation embedding becomes minimizing the summation of the pair-wise Euclidean distances of correlated label vectors:

$$\min_{\theta_{\text{embed}}} \sum_{j\in S} \sum_{(k<j,k\in S)} (\boldsymbol{a}_j - \boldsymbol{a}_k)^2\,, \tag{9}$$

where the set $S = \{j \mid y^j = 1\}$. However, for a large-scale number of labels, Eq. (9) is computational redundancy. By performing some transformations on the term in Eq. (9), we have

$$\sum_{j\in S} \sum_{(k<j,k\in S)} (\boldsymbol{a}_j - \boldsymbol{a}_k)^2 \propto \sum_{j\in S} (\boldsymbol{a}_j - \bar{\boldsymbol{a}})^2\,. \tag{10}$$

Thus, the optimization problem in Eq. (9) can be written as

$$\min_{\theta_{\text{embed}}} \sum_{j\in S} (\boldsymbol{a}_j - \bar{\boldsymbol{a}})^2\,, \tag{11}$$

where $\bar{\boldsymbol{a}} = \frac{1}{|S|}\sum_{j\in S}\boldsymbol{a}_j$ is the mean label vector of all the correlated labels. Compared with Eq. (9), Eq. (11) is computationally efficient and could contribute to fast model convergence. Furthermore, considering the uncorrelated label vectors should be apart from the label mean, the final label correlation embedding loss function becomes

$$\mathcal{L}_{\text{lce}} = \sum_{j\in S} (\boldsymbol{a}_j - \bar{\boldsymbol{a}})^2 + \sum_{k\in \overline{S}} \left[1 - (\boldsymbol{a}_k - \bar{\boldsymbol{a}})^2\right]_+\,, \tag{12}$$

where the $[\cdot]_+$ operation indicates the hinge function $\max(0,\cdot)$, and $\overline{S} = \{k \mid y^k = 0\}$. By introducing the second term $\left[1 - \sum_{k\in\overline{S}}(\boldsymbol{a}_k - \bar{\boldsymbol{a}})^2\right]_+$ into Eq. (12), it can consider the relationship of correlated labels and uncorrelated labels at the same time, which can better capture the label co-occurrence from these two different perspectives. Furthermore, it can also prevent to obtain the trivial solution [22], *i.e.*, $\boldsymbol{a}_c = f_{\text{embed}}(f_{\text{flat}}(\boldsymbol{A}_c);\theta_{\text{embed}}) = \mathbf{0}$.

3. EXPERIMENTS

3.1. Evaluation metrics

Following conventional settings [12, 14], we report the average per-class precision, recall, F1 (CP, CR, CF1) and the average overall precision, recall, F1 (OP, OR, OF1) for performance evaluation. For each image, we assign labels with confidence greater than 0.5 as positive, and compare with the ground-truth labels. These measures do not need a fixed number of labels per image. Particularly, to fairly compare with exiting state-of-the-art methods, we also report the results of top-3 labels with highest confidences.

Additionally, the average precision (AP) for each label and the mean average precision (mAP) are also important for evaluating multi-label image recognition accuracy, which are also employed for performance comparisons. In general, average overall F1 (**OF1**), average per-class F1 (**CF1**) and **mAP** are relatively more important for evaluation.

Table 1. Comparisons with state-of-the-art methods on the *MS-COCO* dataset.

Methods	All							top-3					
	mAP	CP	CR	**CF1**	OP	OR	**OF1**	CP	CR	**CF1**	OP	OR	**OF1**
CNN-RNN [23]	61.2	–	–	–	–	–	–	66.0	55.6	60.4	69.2	66.4	67.8
Order-Free RNN [24]	–	–	–	–	–	–	–	71.6	54.8	62.1	74.2	62.2	67.7
ML-ZSL [25]	–	–	–	–	–	–	–	74.1	64.5	69.0	–	–	–
SRN [12]	77.1	81.6	65.4	71.2	82.7	69.9	75.8	85.2	58.8	67.4	87.4	62.5	72.9
Multi-Evidence [14]	–	80.4	70.2	74.9	**85.2**	72.5	78.4	84.5	62.2	70.6	89.1	64.3	74.7
ResNet-101 (Baseline)	78.3	80.2	66.7	72.8	83.9	70.8	76.8	84.1	59.4	69.7	89.1	62.8	73.6
Ours ($\lambda = 0$)	79.9	81.6	68.5	74.5	84.4	72.5	78.0	83.2	62.0	71.1	89.4	64.2	74.7
Ours ($\lambda = 0.5$)	**82.3**	**82.5**	**72.2**	**77.0**	84.0	**75.6**	**79.6**	**87.1**	**63.6**	**73.5**	**89.4**	**66.0**	**76.0**

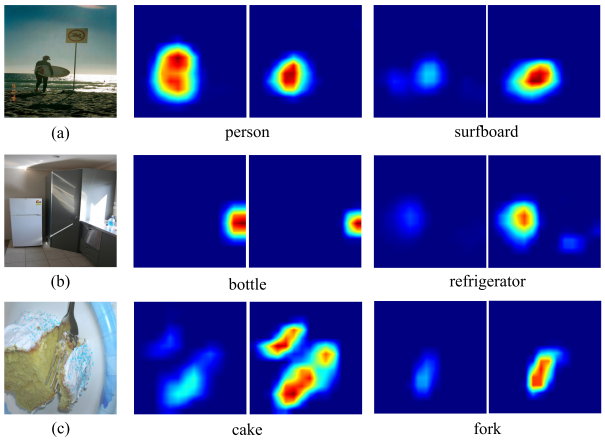

Fig. 4. Visualization and comparisons of class-aware disentangled maps generated *with* (on the right) vs. *without* (on the left) our label correlation embedding module.

3.2. Implementation details

In our experiments, the resolution of input images is set to 448×448, and we use a common pre-processing strategy, *i.e.*, random horizontal flip, for data augmentation. Following [12, 14, 24, 25], ResNet-101[15] is chosen as the base model of our proposed method. We utilize the pre-trained model based on ImageNet for model parameter initializations. For optimization, SGD with momentum of 0.9 is selected as the network optimizer. The weight decay is set to 10^{-4}. Initial learning rate is 0.01, and it is divided by 10 for every 20 epochs until 60 as the total training epochs.

3.3. Comparison with state-of-the-art methods

3.3.1. Performance on the MS-COCO *dataset*

Microsoft COCO [26] is a wildly used dataset for multi-label image recognition. The training set is composed of $82,081$ images and the validation set consists of $40,504$ images. The dataset covers 80 common object categories, and each image contains about 3.5 labels on average. As the ground-truth labels of the test set are not available, we evaluate the performance of all the methods on the validation set instead. On *MS-COCO*, we compare our proposed model with recent state-of-the-art methods, *e.g.*, CNN-RNN [23], SRN [12], Order-Free RNN [24], ML-ZSL [25] and Multi-Evidence [14], etc. The comparison results are reported in Table 1. It is clear that our method outperforms the previous state-of-the-arts by a sizable margin, especially $+5.2\%$, $+2.1\%$, $+1.2\%$, $+2.9\%$, $+1.3\%$ improvements on the evaluation of mAP, CF1 (All), OF1 (All), CF1 (top-3) and OF1 (top-3), respectively.

In addition, we also conduct an ablation study of our model, *i.e.*, directly setting the trade-off parameter λ in Eq. (5) to 0, for validating the effectiveness of the two proposed crucial modules. Compared with the results of $\lambda = 0.5$, the model without label correlation embedding has a significant performance drop, *i.e.*, 3.2% mAP lower than our proposal.

3.3.2. Performance on the NUS-WIDE *dataset*

The *NUS-WIDE* dataset [27] is another benchmark dataset for multi-label recognition, which contains $269,648$ images with associated tags from `Flickr`. This dataset is manually annotated by 81 concepts, with 2.4 concept labels per image on average. Official train/test splits are utilized, *i.e.*, $161,789$ images for training and $107,859$ images for test.

Empirical results on this dataset are shown in Table 2. Our method achieves the best multi-label recognition performance competing with the previous state-of-the-arts, especially on the evaluation of mAP, CF1 (All), OF1 (All), CF1 (top-3) and OF1 (top-3). Moreover, the ablation study validates the effectiveness of our proposed modules: We obtain about 1% improvement comparing with the result of $\lambda = 0$, which is consistent with the observations on *MS-COCO*.

3.4. Visualization and analyses

In this section, we validate the effectiveness of our proposed key modules (especially for label correlation embedding) by visualization results from the qualitative perspective. We show the class-aware disentangled maps (CADMs) in Fig. 4 for comparisons. Three sampled input images with the corresponding CADMs are presented in each sub-figure. We select two of the multiple image labels which are apparent to be observed in the input image. For each label, the CADMs generated with our label correlation embedding module (*i.e.*, $\lambda = 0.5$) are shown on the right, while the CADMs generated without label cor-

Table 2. Comparisons with state-of-the-art methods on the *NUS-WIDE* dataset.

Methods	All							top-3					
	mAP	CP	CR	**CF1**	OP	OR	**OF1**	CP	CR	**CF1**	OP	OR	**OF1**
CNN-RNN [23]	–	–	–	–	–	–	–	40.5	30.4	34.7	49.9	61.7	55.2
Order-Free RNN [24]	–	–	–	–	–	–	–	59.4	50.7	54.7	69.0	**71.4**	70.2
ML-ZSL [25]	–	–	–	–	–	–	–	43.4	48.2	45.7	–	–	–
SRN [12]	62.0	65.2	55.8	58.5	75.5	71.5	73.4	48.2	**58.8**	48.9	56.2	69.6	62.2
ResNet-101 (Baseline)	60.4	63.1	55.5	59.1	74.3	71.7	72.9	64.9	48.3	55.3	76.8	62.1	68.7
Ours ($\lambda = 0$)	61.6	63.7	56.1	59.7	75.7	70.5	73.0	**65.4**	48.8	55.9	78.3	61.4	68.8
Ours ($\lambda = 0.5$)	**62.8**	**63.8**	**57.8**	**60.7**	**75.8**	**72.5**	**74.1**	64.3	57.7	**56.3**	**78.8**	63.9	**70.6**

relation embedding (*i.e.*, $\lambda = 0$) are shown on the left. From these figures, it is clear to observe that when utilizing our label correlation embedding, it could significantly strengthen the activations of these relevant labels' CADMs, *e.g.*, "`surfboard`" of Fig. 4 (a), "`refrigerator`" of Fig. 4 (b), "`cake`" and "`fork`" of Fig. 4 (c), etc. It is reasonable to benefit the recognition of the labels whose original CADM is weak. Therefore, it could give an intuitive and straightforward explanation on why our model achieves outperforming multi-label image recognition accuracy on two aforementioned benchmark datasets.

4. CONCLUSION

In this paper, we proposed a unified framework for multi-label image recognition. Our model consisted of two key modules, *i.e.*, class-aware map disentangling and label correlation embedding. With only image-level supervision, our model can be trained in an end-to-end manner. Experimental results and visualization analyses validated the effectiveness of the proposed method from both quantitative and qualitative perspectives. In the future, developing novel label correlation embedding loss is promising for further boosting the performance.

5. REFERENCES

[1] J. Shao, C.-C. Loy, K. Kang, and X. Wang, "Slicing convolutional neural network for crowd video understanding," in *CVPR*, 2016, pp. 5620–5628.

[2] Y. Li, C. huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *ECCV*, 2016, pp. 684–700.

[3] T. Yang, S. Qin, J. Yan, and W. Zhang, "Multi-label dilated recurrent network for sequential face alignment," in *ICME*, 2018, pp. 1–6.

[4] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "RPC: A large-scale retail product checkout dataset," *arXiv preprint arXiv:1901.07249*, pp. 1–9, 2019.

[5] Y. Zhu, J. Kwok, and Z.-H. Zhou, "Multi-label learning with global and local correlation," *TKDE*, vol. 30, no. 6, pp. 1081–1094, 2018.

[6] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *TKDE*, vol. 26, no. 8, pp. 1819–1837, 2014.

[7] C. Liu, P. Zhao, S.-J. Huang, Y. Jiang, and Z.-H. Zhou, "Dual set multi-label learning," in *AAAI*, 2018, pp. 3635–3642.

[8] Y. Gong, Y. Jia, L. Thomas, T. Alexander, and I. Sergey, "Deep convolutional ranking for multi-label image annotation," *arXiv preprint arXiv:1312.4894*, pp. 1–9, 2013.

[9] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori, "Learning structural inference neural networks with label relations," in *CVPR*, 2016, pp. 2960–2968.

[10] Q. Li, M. Qiao, W. Bian, and D. Tao, "Conditional graphical LASSO for multi-label image classification," in *CVPR*, 2016, pp. 2977–2986.

[11] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *ICCV*, 2017, pp. 464–472.

[12] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *CVPR*, 2017, pp. 5513–5522.

[13] L. Liu, C. Shen, and A. van den Hengel, "Cross-convolutional-layer pooling for image recognition," *TPAMI*, vol. 39, no. 11, pp. 2305–2313, 2016.

[14] 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 *CVPR*, 2018, pp. 1277–1286.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[16] B. Zhou, K. Aditya, L. Agata, O. Aude, and T. Antonio, "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.

[17] G. E. Hinton, "Learning distributed representations of concepts," in *CogSci*, 1986, pp. 1–12.

[18] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *TPAMI*, vol. 35, no. 8, pp. 1798–1828, 2013.

[19] J. Yang, C Qi, Y. Li, and J. Li, "Face recognition using extended generalized rayleigh quotient," in *ICME*, 2017, pp. 1–6.

[20] H. Yao, S. Zhang, D. Zhang, Y. Zhang, J. Li, Y. Wang, and Q. Tian, "Large-scale person re-identification as retrieval," in *ICME*, 2017, pp. 1–6.

[21] X.-S. Wei, C.-L. Zhang, L. Liu, C. Shen, and J. Wu., "Coarse-to-fine: A RNN-based hierarchical attention model for vehicle re-identification," in *ACCV*, 2018, pp. 1–16.

[22] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *NIPS*, 2016, pp. 1857–1865.

[23] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *CVPR*, 2016, pp. 2285–2294.

[24] S.-F. Chen, Y.-C. Chen, and Y.-C. Frank C.-K. Yeh, C.-K. Wang, "Order-free RNN with visual attention for multi-label classification," in *AAAI*, 2018, pp. 6714–6721.

[25] C.-W. Lee, F. Wei, C.-K. Yeh, and Y.-C. Wang, "Multi-label zero-shot learning with structured knowledge graphs," in *CVPR*, 2018, pp. 1576–1585.

[26] T.-Y. Lin, M. Michael, B. Serge, H. James, P. Pietro, R. Deva, D. Piotr, and Z. C Lawrence, "Microsoft COCO: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[27] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *CIVR*, 2009, pp. 1–9.