Adaptive Attention Annotation Model: Optimizing the Prediction Path through Dependency Fusion

Fangxin Wang[1, 2], Jie Liu[1*], Shuwu Zhang[1,3], Guixuan Zhang[1], Yang Zheng[1], Xiaoqian Li[1, 2], Wei Liang[1] and Yuejun Li[1, 2]

[1] Institute of Automation, Chinese Academy of Sciences
Beijing, 100190 - P. R. China
[2] University of Chinese Academy of Sciences
Beijing, 100049 - P. R. China
[3] AICFVE, Beijing Film Academy
Beijing, 100088 – P.R. China
[e-mail: wangfangxin2014@ia.ac.cn, jie.liu@ia.ac.cn]
*Corresponding author: Jie Liu

Received September 17, 2018; revised November 7, 2018; revised March 2, 2019; accepted March 23, 2019; published September 30, 2019

Abstract

Previous methods build image annotation model by leveraging three basic dependencies: relations between image and label (image/label), between images (image/image) and between labels (label/label). Even though plenty of researches show that multiple dependencies can work jointly to improve annotation performance, different dependencies actually do not "work jointly" in their diagram, whose performance is largely depending on the result predicted by image/label section. To address this problem, we propose the adaptive attention annotation model (AAAM) to associate these dependencies with the prediction path, which is composed of a series of labels (tags) in the order they are detected. In particular, we optimize the prediction path by detecting the relevant labels from the easy-to-detect to the hard-to-detect, which are found using Binary Cross-Entropy (BCE) and Triplet Margin (TM) losses, respectively. Besides, in order to capture the inforamtion of each label, instead of explicitly extracting regional featutres, we propose the self-attention machanism to implicitly enhance the relevant region and restrain those irrelevant. To validate the effective of the model, we conduct experiments on three well-known public datasets, COCO 2014, IAPR TC-12 and NUSWIDE, and achieve better performance than the state-of-the-art methods.

Keywords: image annotation, multiple dependencies, self-attention, prediction path, Triplet Margin loss

1. Introduction

Image annotation refers to the process assigning any image with its relevant labels from predefined list of keywords, which is the key to semantic keyword based image retrieval and understanding. However, it can be very costly and subjective to annotate an large scale of image manually. Therefore, automatic image annotation (AIA) is receiving more attention in the field.

Previously methods build image annotation model based on three basic dependencies: relations between image and label (image/label), between images (image/image) and between labels (label/label). Assuming targets are independent to each other, they managed to establish the relations between images and labels [1-8]. Despite its powerful discriminative ability, it cannot detect those visually hard-to-detect targets, such as small and blured objects. In fact, we can further improve the performance by using multiple dependencies, since most of the targets are correlated with each other, and similar images share common features. Inspired by these conclusions, subsequent researches tried to obtain extra clues utilizing image/image and label/label dependencies. One of the most common paradigm is refining the annotation predicted by image/label dependency, using the other two dependencies. Even though this methods can, to some extent, keep discriminative and optimize generalization ability, its performance is largely affected by the initial annotation. In order to obtain better performance, we need to make them work jointly as a whole. Wang et al. [9] proposed prediction path, a series of labels in the predefined order to be recognized, to integrate image/label with label/label dependencies. However, different prediction path may produce very different results, and it is still a great chanllege to this method since it can be very costly to find the best one. Zhu et al. [44] proposed Spatial Regularization Network (SRN) to model the spatial relations between labels.

Another bottleneck of the current methods comes from the lack of regional features, as popular deep networks tend to extract global features that output from fully-connected layer. Similar to image detection, there are multiple targets in an image for annotation, global features usually lead to the loss of the information of some labels, especially for those targets with low occurrence. One of the feasiable solution is to devide the whole image into patches, and extract regional features of them. Zhang et al. [10] suggest finding accurate features for every object through Object Patch Net. Though this method is effective in find regional featutres, it fails to fuse the regional features with the global features.

In this paper, we put forward the adaptive attention annotation model (AAAM, **Fig. 1**) to address the above two problems. On one hand, we use image/label dependency to find those easy-to-detect targets where we apply Binary Cross-Entropy (BCE) loss to model image/label dependency, and at the same time, we apply Triplet Margin (TM) loss to make the undetected relevant labels close to detected relevant ones, and make the detected irrelevant labels away from detected relevant ones. In this way, TM can adaptively adjust the relations among labels every time BCE detects new targets, and annotation results can thus take both dependencies into account. Experimental results show that the proposed prediction path from the easy-to-detect to the hard-to-detect labels can obtain better performance than previous methods. On the other hand, different from previous methods that try to directly obtain regional featutres, we assume that the regional featutres are implicitly contained in the feature maps, and can be extracted by enhancing the relevant and restraining the irrelevant ones.

Therefore, we present an self-attention layer, acting on feature maps, to implicitly extract regional featutres. The key advantage of our method is that we can not only keep accurate regional featutres, but also incorporate them into the global features through forward propagation .

The main contributions of the paper are as follows: (1) Instead of traditional annotation paradigm, we propose an end-to-end automatic image annotation model AAAM, making image/label and label/label work jointly, to improve the performance of visually hard-to-detect targets. (2) Contrary to previous methods explicitly extracting regional featutres, we put forward an Self-Attention layer, implicitly extracting and incorporating them into global features, to further improve the discriminative ability. The experimental results on the COCO 2014 [11], IAPR TC-12 [12] and NUSWIDE [13] benchmark show that our method outperforms those methods using traditional paradigm.

The rest of the paper is organized as follows: we introduces the related work in image annotation in Section 2, present our motivation in Section 3, elaborate the proposed annotation model AAAM in Section 4, analyze our experiments in detail in Section 5, and make a conclusion in Section 6.

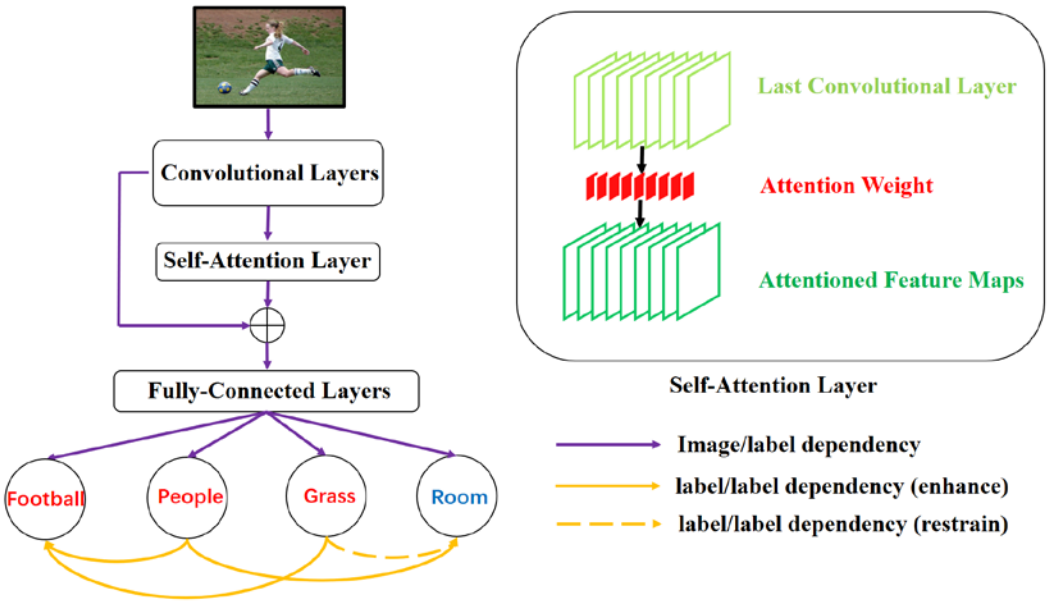

Fig. 1. The structure of the proposed AAAM. The red labels are relevant to the query image while the blue one is irrelevant. Those esay-to-detect labels, such as *people* and *grass*, are first detected by BCE (image/label dependency), then based on these previous found labels, we apply TM (label/label) to detect the hard-to-detect label *football*, and restrain the irrelevant label *room*. The Self-Attention layer helps us find more useful regional features of the targets.

2. Related Work

Earlier works built image annotation model mainly by leveraging three basic dependencies: image/label, image/image and label/label. Inspired by the topic models in natural language processing, some literatures applied LSA [14], pLSA [15] and LDA [16] to model the joint distribution over images and labels. Zheng et al. [17] proposed Supervised

Document Neural Autoregressive Distribution Estimator (SupDocNADE) to learn the joint representation from images and labels, and obtained a better performance than previous topic models. Park et al. [18] trained a model to build a shared feature space of both media by max-margin embedding method. Gong et al. [19] combined DNN with different top-k ranking loss functions to improve performance. Rezatofighi et al. [42] transfromed the multi-label classification problem into sets prediction problem, it used Convolutional Neural Network (CNN) to model the image/label dependency, and applied the cardinality as an extra supervised informantion to improve the annotation performance. Usunier et al. [29] proposed the Weighted Approximate-Rank Pairwise (WARP) loss to optimize the label ranking problem in image annotation. In order to separate the irrelevant labels from the target image, He et al. [5] took a triple consisting of images, relevant labels and irrelevant labels as input, to train a deep neural network (DNN) by pairwise hinge loss. Ghamrawi et al. [20] used Conditional Random Field to model the label/label dependency. Wu et al. [21] took image and tags as two instance sets, and construct a weakly supervised learning framework using deep multiple instance learning.

Even though models based on either of these dependencies can obtain fair performance, they still left much to be desired. Models based on image/label often failed to detect visually hard-to-detect targets, such as small objects and abstract objects; those based on label/label can extract less discriminative features; methods based on image/image often needed large amount of samples to get joint distribution. Plenty of researches demonstrated that multiple dependencies can work jointly to improve annotation performance. The common practice is first to train the model , based on image/label, to predict an initial annotation, and then refine them utilizing extra knowledge from the other two dependencies. Jin et al. [22] employed a cascading structure with CNN and Recurrent Neural Network (RNN) to predict arbitrary length image tag recurrently, where the dependency between labels and dependency between image and label are modeled using RNN. Wang et al. [23] put forward the CNN-RNN framework for multi-label classification problem. In this model, it transforms a multi-label prediction to an order prediction problem. The CNN part extracts image features and the RNN part captures the information of the previously predicted labels, followed by the projection layer computing the output label probabilities. The biggest innovation of this paper is using RNN to model the high-order dependency between labels and dependency between image and label. Liu [43] proposed to use a semantically regularised embedding layer as the interface between the CNN and RNN, to improve the co-training between CNN and RNN. Murthy et al. [24] and Uricchio et al. [25] adopted Canonical Correlation Analysis (CCA) and Kernel CCA (KCCA) to refine the annotation based on the features obtained through previous deep network based on image/label respectively. Wu et al. [41] put forward the diverse and distinct image annotation (D2IA) to produce the tags like humans, where it applied weighted semantic path to model the label/label dependency, and used CNN to model the image/label dependency.

As current deep networks tend to extract global features, those annotation model based on deep networks will lose discriminative ability to some extent. Therefore, some researches managed to explicitly extract regional featutres. One of the classical methods transformed multi-label classification into multiple binary classification problem, i.e. building a binary classifier for every label and picking out the most likely labels by output scores. However, it often failed to deal with large numbers of images, due to its high cost and low efficiency. To improve its computation cost, Tsoumakas et al. [26] divided the multiple label set into a few small random multiple label sets called RAndom *k* labELsets (RA*k*EL), and trained classifiers for every label powerset, which greatly reduced the computation complexity. Wei et al. [27]

proposed a novel framework called Hypotheses-CNN-Pooling, which took a few hypotheses of the image as input, and assembled these CNN features using max-pooling to output the predictions. Zhang et al. [10] presented the Object Patch Net, which divided each image into patches to build the dense patch group set, then incorporated these groups into a graph, and searched for the most similar nodes for every patches of the test image, finally predicted its annotations by voting. Wang et al. [28] proposed a patch-level and end-to-end architecture to model the appearance of local patches, called PatchNet. PatchNet was essentially a customized network trained in a weakly supervised manner, which used the image-level supervision to guide the patch-level feature extraction.

3. Motivation

As some recent researches has proven that the fusion of multiple dependencies can further improve the annotation performance, subsequent annotation methods began taking multiple dependencies into account. Most of them used the similar learning paradigm: first trained the model, based on image/label, to predict an initial annotation, and then refined them utilizing extra knowledge from the other two dependencies. However, in these models, different dependencies actually did not work jointly, and thus can be further improved. Wang et al. [23] treated all the labels as a sentence and each label as a word in the sentence, they asserted that the annotation is actually a matter of orders, and the order of labels are pre-determined that is called prediction path. However, it was often time-consuming to search for the best prediction path. Therefore, finding the prediction path adaptively should be a optimal solution. In order to confirm the relationship between dependencies and prediction paths, we conduct experiments using image/label (WARP, the 2nd row) and label/label (BCE, the 3rd row) dependencies respectively, and the results are shown in **Fig. 2**.

According to our experimental results, the features of some labels are learned in the early stage of training, while the other ones in the later stage, and they have different prediction path for different dependencies. Therefore, the prediction paths are actually the inherent attributes of different dependencies, and the optimal prediction paths can be found by the fusion of multiple dependencies. Here, we simply call the targets that are relevant to an image and deteced in the early stage the easy-to-detect targets, and in the later stage the hard-to-detect ones. In our paradigm, we assume that by the fusion of multiple dependencies, those easy-to-detect targets can help to find the hard-to-detect and filter out the irrelevant ones.

Similar to the image detection, image annotation needs to find out multiple targets relevant to an image, where local and global features are of equal importance. However, the deep network tend to extract global features and neglect regional ones. Attention mechanism [30-34] is widely applied in machine translation, text classification and etc., which addresses the problem that different parts do not make an equal contribution to the task. In the context of image annotation, most of the labels represent an object, which is only corresponding some part of the image. In order to detect all the relevant labels, we should assign different attentions to different parts. Instead of explicitly extracting regional featutres from large quantity of proposals, we assume that they are implicitly contained in the feature maps. Hence, we add an Self-Attention layer assigning different weights to those feature maps, to implicitly extract regional featutres.

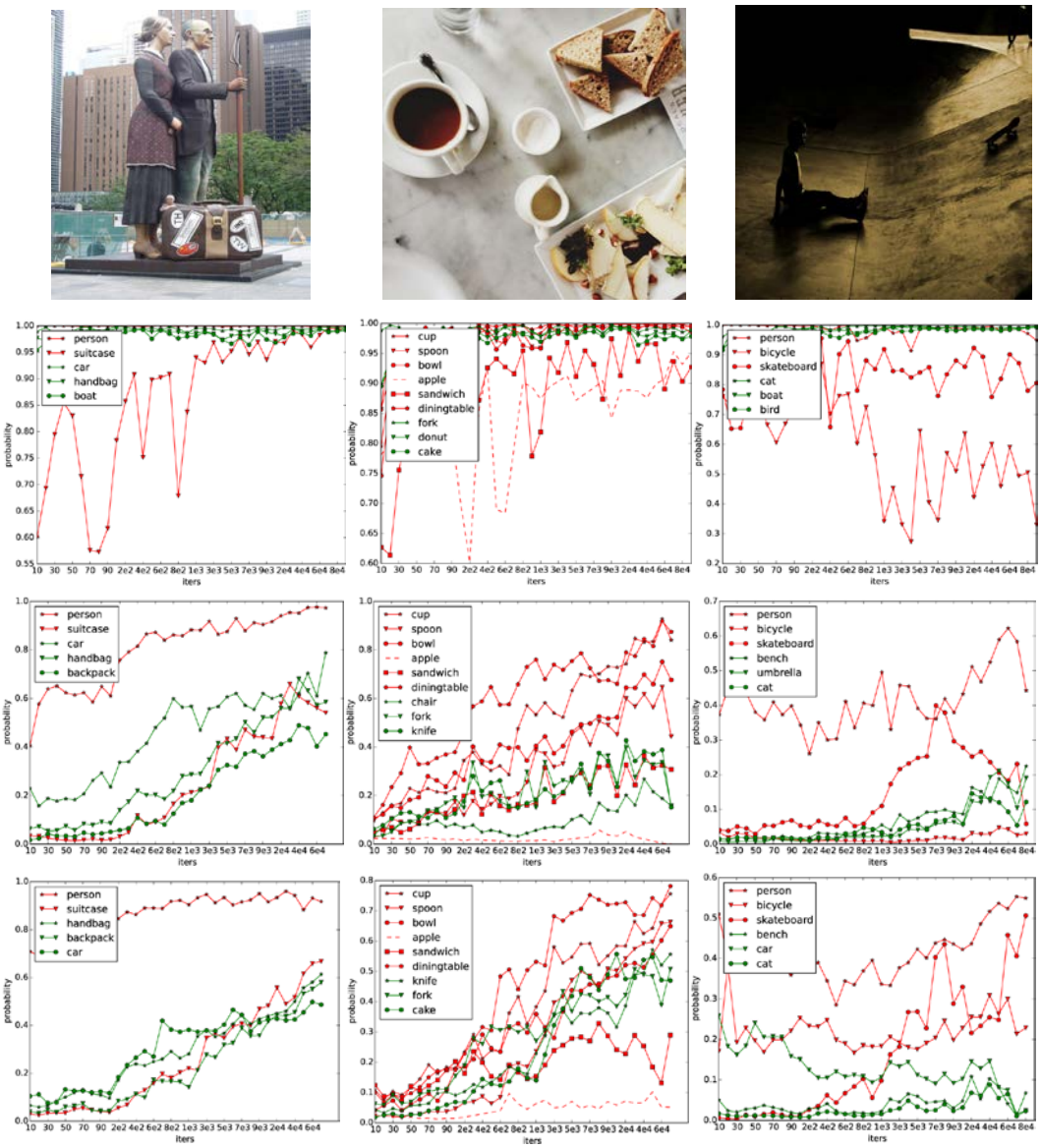

Fig. 2. Scores of labels for different models on COCO 2014: the 1st row shows the target images, the 2nd, 3rd and 4th rows show the results of WARP, BCE and AAAM respectively. The red texts represent the relevant labels, and the green ones represent the irrelevant labels.

4. Adaptive Attention Annotation Model

At the beginning, we will formulate the image annotation task. Given an image I, the set of all the possible labels $Y = \{y_1, y_2, \cdots, y_n\}$, where n is the vocabulary size, i.e. the number of labels used for annotation. we use a binary vector $Z_I = [z_1, z_2, \cdots, z_n]$ $(z_i \in \{0,1\})$, to represent relevant/irrelevant relations between an image and all the labels, where $z_i = 1$

means the image has been annotated with y_i and 0 otherwise. The goal of the image annotation is building a model $\hat{Z}_I = F(I;W)$, to predict if a label is relevant/irrelevant to an image.

4.1 Cross-Entropy Loss

In our paradigm, we make two dependencies, image/label and label/label, work jointly to adaptively find better prediction path. Especially, we apply Binary Cross-entropy (BCE) loss to model the image/label dependency as our baseline:

$$L_{BCE} = -\sum_{i=1}^{n} z_i \log \hat{z}_i + (1 - z_i) \log(1 - \hat{z}_i) \tag{1}$$

where the gradient is the difference between each pair of predicted scrore $\hat{z}_i$ and ground-truth z_i:

$$\frac{\partial L_{BCE}}{\partial \hat{z}_i} = \hat{z}_i - z_i \tag{2}$$

since $\hat{z}_i$ is a function of I, we can rewrite (2) as follows:

$$\frac{\partial L_{BCE}}{\partial \hat{z}_i} = F(I,W)_i - z_i \tag{3}$$

The Cross-Entropy loss optimizes the relationship between image I and label z_i as we desired, and bring more stable but faster convergence comparing with quadratic function such as Mean Square Error (MSE). However, the gradient with respect to $\hat{z}_i$ is not associate with other predicted label $\hat{z}_j$ ($j \neq i$), which means we can hardly detect those visually hard-to-detect targets.

4.2 Triplet Margin Loss

Although some researches have demonstrated that it can help to predict more precisely results to employ label/label dependency, which consider it as a ranking problem and give punishments to those cases that the irrelevant labels rank ahead of the relevant ones. However, from the view of the prediction path, early detected relevant objects should not only help model to detect the other hard-to-detect targets, but also suppress those irrelevant targets on the other hand. These intuitions indicate that not only relation between relevant and irrelevant objects, that between relevant objects also need to be considered. Therefore, we propose to use TM loss to address this problem.

$$L_{TM} = \frac{1}{2} \sum_{i,j \in R^+, k \in R^-} \max(0, m + \Delta\hat{z}_{i,j} - \Delta\hat{z}_{i,k}) \tag{4}$$

where $\Delta\hat{z}_{i,j} = |\hat{z}_i - \hat{z}_j|^2$ is the score discrepancy between two relevant labels y_i and y_j, and $\Delta\hat{z}_{i,k} = |\hat{z}_i - \hat{z}_k|^2$ between irrelevant labels y_i and y_k, respectively. The gradient with respect to $\hat{z}_i$ is:

$$\frac{\partial L_{TM}}{\partial \hat{z}_i} = I(m + \Delta\hat{z}_{i,j} - \Delta\hat{z}_{i,k})(|\hat{z}_i - \hat{z}_j| - |\hat{z}_i - \hat{z}_k|) \tag{5}$$

where $I(\bullet)$ is an indicator function. Based on the above analysis, we can observe that, comparing with absolute discrepancy in pairwise ranking losses, the TM gives consideration to both relations at the same time, by enlarging relative discrepancy between the relevant and irrelevant pairs.

4.3 Joint Training

In order to make above dependencies work jointly in our paradigm, we design a triplet sampler that choose a top-ranking relevant label y_i, a lower-ranking relevant label y_j and a top-ranking irrelevant tag y_k, respectively at each time, so that the model can adaptively changes its prediction path to detect more relevant objects. Then, we perform joint training on these two losses:

$$L = L_{BCE} + \lambda L_{TM} \tag{6}$$

where λ is designed to make a trade-off between the two losses. With the integration of two dependencies, the performance of these later-predict objects can be boosted by those previous-predict targets. We conclude the training process as **Algorithm 1**.

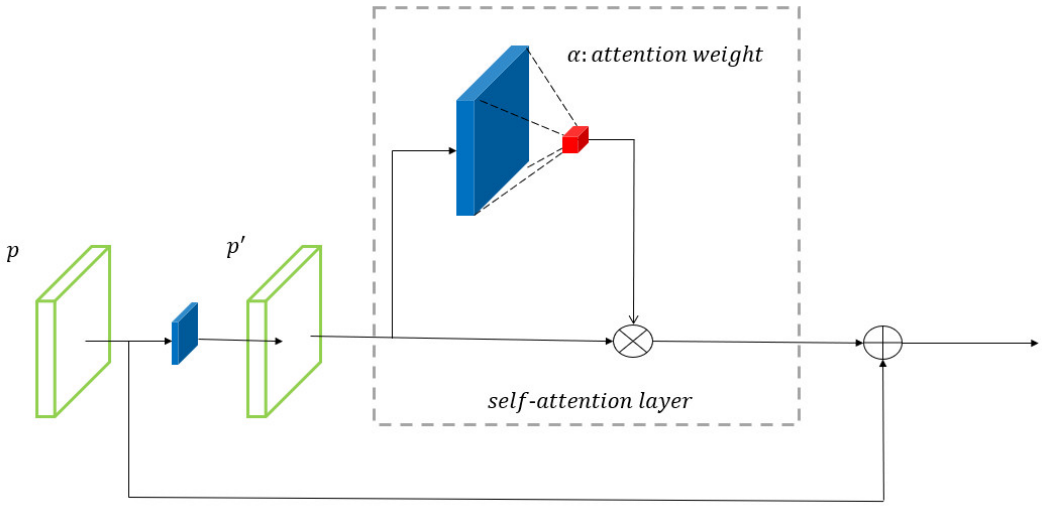

Fig. 3. The architecture of the Self-Attention layer. p is the feature map of the last convolution layer of VGG16, we first use a convolutional layer to produce feature map $'p$, then we use the filters whose kernel size is the same as input p', and output $n\times1\times1$ attention weights α, where n is the number of the feature maps. And we strengthen or weaken p' according to the value of α. At last, we adopt a skip-connections to add p and the output of the Self-Attention layer.

4.4 Self-Attention Layer

In the field of image annotation, an image corresponds to multiple labels, where the features for each target (regional featutres) are expected for accurate annotation. Instead of explicitly extracting regional featutres such as dividing an image into patches [28], or generating proposals [35-37], we insist that all the regional featutres are implicitly contained in the feature maps of convolutional layers, and propose an Self-Attention layer to enhance those relevant regions of targets and weaken those irrelevant regions. Besides, these enhanced feature maps are integrated to generate global features (features of FC layers), making more local information be included. The architecture of Self-Attention layer is shown in **Fig. 3**.

In fact, generating proposals can be equal to assigning weight 1to those relevant and 0 to those irrelevant proposals to some extent. Hence, in order to extract the features of those relevant proposals, we can give different weights to the features of all the proposals. However, it can be very costly to do it, since it can be thousands proposals for each image. In this paper, we simplify this process by assigning different weights to different feature maps. Intutively, we always give more attention to those partion that closely connected to the target tags. Therefore, we apply Self-Attention machanism to address this problem. In our method, at first, we use a convolutional layer to produce feature map $'p$ which has the same number of channels with p , then use a series of filters to convolve with $'p$ to obtain the attention weights, where these filters have the same size with the input feature maps. And, we can get attentional feature maps by multiplying them together.

Algorithm 1: *Joint Training with BCE and TM Loss*

Input: annotated data $(x^{(m)}, y^{(m)})$, $y^{(m)} \in Y$

Repeat

pick a random annotated labeled example $(x^{(m)}, y^{(m)})$

obtain predicted score through $\hat{Z}^{(m)} = F(x^{(m)}; W)$

rank all the elements of $\hat{Z}^{(m)}$ in descend order

Repeat

random choose the top-ranking relevant label $y_i^{(m)}$, lower-ranking relevant label $y_j^{(m)}$ and top-ranking irrelevant label $y_k^{(m)}$, get their predicted score $z_i^{(m)}$, $z_j^{(m)}$ and $z_k^{(m)}$ respectively.

compute the joint loss according to (6) and perform backpropagation

Until all the top-ranking relevant labels are scaned

Until validation error does not improve

Assuming the feature maps in the last convolutional layer are $p = \{p_1, p_2, \cdots, p_l, \cdots\}$, the input feeding into the first fully-connected layer is:

$$p_{attention-l} = p_l + \alpha_l p_l' \tag{7}$$

where α_l is trained to enchance or weaken the current feature maps, $\alpha_l P_l'$ is the output of the Self-Attention layer. In order to make the network easier to be optimized, we adop residual

structure to combine the input P_l and $\alpha_l P_l'$ to feed into the fully-connected layer, $P_{attention} = \{p_{attention-1}, p_{attention-2}, \cdots, p_{attention-l}, \cdots\}$. Comparing with the original feature maps, the proposed Self-Attention layer can tune the attention of the network to target local region precisely without accurate bounding box provieded.

Table 1. Configuration of evaluation datasets

Dataset	COCO 2014	IAPR TC-12	NUSWIDE
No. of Images	122585	19627	209347
No. of Labels	80	291	81
No. of Train Images	82081	17665	125449
No. of Test Images	40504	1962	83898
No. of Average Labels per Image	2.9, [1, 18]	5.7, [5, 23]	2.4, [1, 12]

5. Experiments

5.1 Datasets

To make a comprehensive evaluation, we conduct our experiments on three popular datasets, COCO 2014, IAPR TC-12, and NUSWIDE, respectively, which are widely used in the image annotation domain. Their configurations are shown in **Table 1**. Here, we list the No. of Images and Labels, Train and Test Images, and Average Labels per Images of above datasets. Specially, in order to better understand the ditributions of labels, we also show the minimum and maximum number of labels per images [minimum, maximum].

5.2 Metrics

The performances of automatic image annotation on above datasets have been measured by different metrics. Therefore, following the previous works, we assign a fixed number of labels to each image, and report the precision and recall of the predictions. **Table 1** shows that these datasets have an uneven distribution in labels per image, i.e. for each image, even though it has at least 1 label and at most 18 labels in the ground-truth, we only take top *k* labels (for COCO, *k*=3) as our final results. Consequently, it will bring a paradox that even though both model A and model B correctly predict different *k* labels in ground-truth, their performance measured by recall and precision can be very different. Therefore, in this paper, we also adopt mean Average Precision (mAP) as an important evaluation metric.

5.3 Related Methods

Give consideration to different configurations and preconditions of different approaches, we carefully pick some of them, according to the different dependencies they use, to compare with our model:

1. Softmax [23]: it used regression to model the image/label dependency

2. Logistic [25]: it transformed the multi-label problem to multiple one-vs-all classification problem, which only modeled the image/label relationship.

3. KCCA [25]: it mapped the image and label into a shared features using kernel based Canonical Correlation Analysis method, and used image/label dependency.

4. SVM [25]: for each label, a binary linear SVM classifier was trained using L_2 regularized least square regression, it modeled the image/label dependency.

5. BCE [23]: our baseline, it possesses discriminative ability to model the image/label dependency.

6. WARP [23]: this model mainly cared about the pairwise relationships between labels, i.e. label/label dependency. Based on the previous Pairwise loss, it came up with Weighted Approximately Ranked Pairwise loss, which gave different weights to labels according to their positions in the ranked list.

7. kNN [25]: this approach first computed the L_2 distance between CNN features of each test image and the training images, and then used voting strategy according to its neighborhoods. It utilized both image/label and image/image dependencies.

8. 2PkNN [25]: it used two-phase strategy to search for the neighborhoods of the query image based on the visual features, using image/image dependency.

9. TagRel [25]: it proposed a relevance measure based on the consideration that if several people label visually similar images using the same labels, and these labels were more likely to reflect objective aspects of the visual content.

10. Tagprop [25]: this model automatically found the optimal metric that maximize the likelihood of a probablistic model, which modeled the image/image dependency.

11. CNN-RNN [23]: this model utilized CNN to model the image/label relationships and RNN to model high-order label/label relationship.

5.4 Results

In our implementation, we adopt VGG16 [38] network pre-trained on ImageNet [39] as the image extractor, which is also used in related methods. At the beginning of training, we use only BCE loss to train the model, so as to get these visually easy-to-detect targets, i.e. the top-ranking relevant tags for each image. In the next phase, we build the triplet sampler that will randomly choose a tag y_i from those top-ranking relevant tags, e.g. person in the 3rd column of **Fig. 2**; select a lower-ranking relevant tag y_j, which has not yet been detected, e.g. skateboard or bicycle; select a top-ranking irrelevant tag y_k, e.g. car, for each image every time. After the selection is done, we perform joint training according to (6). As the process goes on, the scores of those visually hard-to-find relevant tags begin to increase, and those relevant tags begin to decrease. But we find that not all the relevant tags could be detected in this way, knife, fork and cake, in the 2nd colomn in **Fig. 2**, can not be depressed, since they are closely correlated with the picture, and need more samples to distinguish them from those relevant tags. Besides, from the **Fig. 2**, we can observe that the scores of the some relevant tags would decrease in the final stage, which is due to the random selection of the triplets. So, in the final stage, we adopt the hard sample detection method that choose the triplets maximizing the TM loss. For the training processes, we use ADAM to optimize our model, and set the learning rate in feature layer 1×10^{-3} and classifier layer 1×10^{-4}, weight decay rate 1×10^{-5} and dropout rate 0.5. All the parameters involved are obtained through cross-validation. For the missing results on some datasets, we reimplement BCE and WARP methods.

We show the image annotation results on COCO 2014, IAPR TC-12 and NUSWIDE datasets in **Table 2**, **3** and **4**, respectively. In order to make an overall analysis on different

dependencies, we carefully choose some models to compare the with proposed method: Softmax, Logistic, KCCA and BCE consider the image/label dependency; WARP gives consideration to the label/label dependency; kNN, 2PkNN, TagRel and TagProp addresses the image/image dependency. In order to confirm the advantages of dependeny fusion, we also compare the performance of fusion of above methods: KCCA + kNN, KCCA + 2PkNN, KCCA + TagRel and KCCA + TagProp combine image/label and image/image dependencies. CNN-RNN takes both image/label and label/label dependency into account. Since it is more powerful to extract discriminative features, which is very important for the majority of the targets, methods based on image/label dependency get better performance than those based on image/image and label/label. BCE outperforms 3% on COCO 2014 and IAPR TC-12, and 12% on NUSWIDE in mAP. For those giving consideration to multiple dependencies, their performances get extra promotion: on IAPR TC-12 and NUSWIDE, combing KCCA with *k*NN, 2P*k*NN, TagRel, Tagprop and SVM can obtain 2% - 7% promotion, which proves that mutiple dependencies can help to boost the detection performances.

Table 2. Image annotation results on COCO 2014 ($k = 3$)

Method	Precision@*k*	Recall@*k*	mAP@10	mAP
Softmax	59.00%	57.00%	47.40%	50.65%
BCE	59.30%	58.60%	55.73%	57.90%
WARP	59.30%	52.50%	49.20%	54.80%
CNN-RNN	66.00%	55.60%	61.2%	-
BCE + Self-Attention	63.50%	53.30%	59.45%	61.47%
BCE + TM	67.35%	53.56%	62.54%	64.33%
AAAM (BCE + TM + Self-Attention)	**68.62%**	**55.00%**	**63.88%**	**65.35%**

Table 3. Image annotation results on IAPR TC-12 ($k = 5$)

Method	Precision@*k*	Recall@*k*	mAP@10	mAP
SVM	31.00%	29.00%	-	34.00%
BCE	40.18%	32.95%	35.58%	37.36%
WARP	36.72%	27.78%	32.78%	34.99%
kNN	39.00%	29.00%	-	36.00%
2PkNN	41.00%	39.00%	-	41.00%
TagRel	34.00%	35.00%	-	35.00%
TagProp	40.00%	32.00%	-	38.00%
KCCA + kNN	44.00%	34.00%	-	40.00%
KCCA + 2PkNN	49.00%	38.00%	-	43.00%
KCCA + TagRel	41.00%	37.00%	-	40.00%
KCCA + TagProp	44.00%	37.00%	-	41.00%
BCE + Self-Attention	45.42%	36.30%	37.36%	39.74%
BCE + TM	49.35%	37.56%	39.74%	42.55%
AAAM (BCE + TM + Self-Attention)	**52.33%**	**40.66%**	**42.87%**	**44.98%**

As for our method, we train our model using image/label and label/label dependency at the same time, the BCE loss can quickly detect those visually easy-to-detect targets, the TM loss can find those visually hard-to-detect but related to the easy-to-detect ones. BCE + TM increases the mAP by 7%, 5% and 2% in three datasets respectively, compared to the baseline BCE.

Table 4. Image annotation results on NUSWIDE (k = 3)

Method	Precision@k	Recall@k	mAP@10	mAP
Logistic	40.90%	43.12%	-	45.78%
SVM	34.60%	**60.60%**	-	50.20%
BCE	41.14%	42.87%	49.40%	51.03%
WARP	31.65%	35.60%	37.66%	39.21%
kNN	39.60%	44.00%	-	49.30%
2PkNN	39.70%	52.20%	-	48.00%
TagRel	32.10%	50.30%	-	49.20%
TagProp	41.30%	44.60%	-	50.90%
KCCA + kNN	40.20%	50.50%	-	51.70%
KCCA + 2PkNN	53.00%	47.00%	-	50.70%
KCCA + TagRel	34.40%	57.20%	-	51.40%
KCCA + TagProp	45.20%	49.20%	-	52.20%
CNN-RNN	40.50%	30.40%	-	-
BCE + Self-Attention	45.34%	42.12%	50.66%	52.20%
BCE + TM	47.10%	44.56%	51.52%	53.66%
AAAM (BCE + TM + Self-Attention)	**54.33%**	47.25%	**52.68%**	**53.98%**

In order to illustrate the effect of two dependencies to the prediction path, we show how the predictions change over time under 3 cases: (1) label/label (2) image/label and (3) image/label + label/label. **Fig. 2** depicts the prediction path of WARP, BCE and our proposed BCE+TM models, respectively. Specially, after obaining the predictions of each method for the query images (the 1st row), we show how these predictions change over time: the red lines represent the ground-truth labels and the green lines the top-ranking irrelevant ones. The predictions of WARP (the 2nd row) focus more on the relationship between labels, its inferences often tend to be the semantic similar labels, such as *sandwitch*, *cake* and *donut* (in the 2nd query image), which leads to incorrect results. In contrast, BCE (the 3rd row) can distinguish simantic similar objects, but its inferences often contain visually similar objects such as *suitcase* , *backpack* and *handbag* (in the 1st query image), and often cannot detect those visually hard-to-detect objects such as *bicycle* (in the 3rd query image). Due to the double regularizations between relevant and irrelevant tags imposed by BCE and TM, our model is more discriminative, where all the probabilities of irrelevant labels are suppressed in a lower level (the green lines) and the distances between those relevant labels (the red lines) are much farther. Besides, as TM considers the relations between relevant tags, some visually

hard-to-detect objects such as *bicycle*, can also be successfully detected. **Fig. 4** depicts the probability distribution of some of our predictions, we can see clearly the discriminative power and effective generalization, our method can further increase the prediction probability of the relevant labels (red texts)and decrease the irrelevant ones (green texts). Therefore, comparing with baseline, our method can significantly improve the annotation performance.

Apart from the advantage of the dependency fusion, we also make an analysis of the impact of the proposed Self-Attention layer on the performance. **Fig. 5** shows the activation maps [40] of the baseline and proposed model. We compare their performance by their class activation maps of the ground truth labels, which are readily comprehensible to us, and the results of the same label are shown in the same column for each group. In the 1st group, the high-response regions of *people* and *police* are concentrated on the human body for our model (the 2nd row). While for the baseline (the 1st row), they expand their region to the *road* by mistake. In the 2nd group, the high-response area of *bird* and *animal* locate in the same region

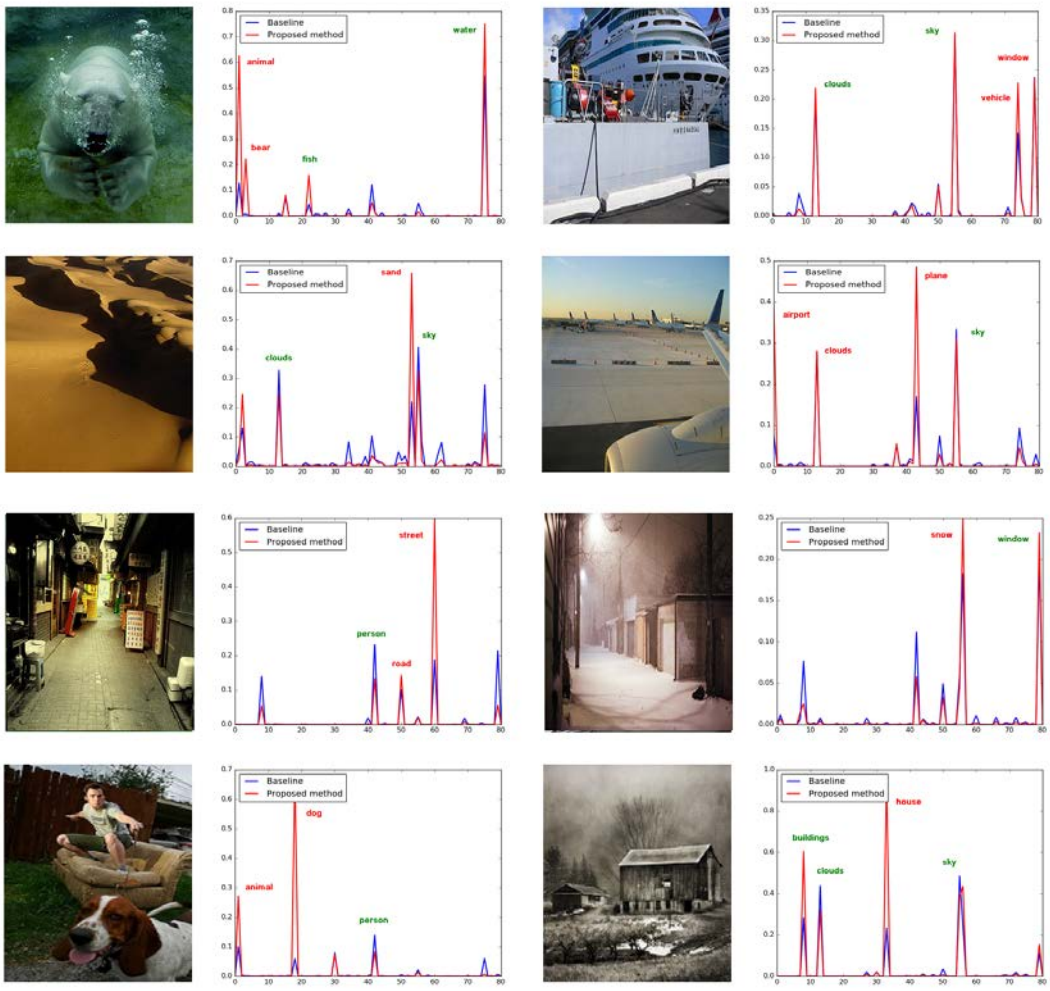

Fig. 4. Comparison between the baseline and AAAM on image annotation, where red texts represent the relevant labels and green ones the irrelevant labels.

for our model (the 2nd row), since they have the same semantic meaning in this image. While for the baseline (the 1st row), they locate at different areas. As far as the mAP results, with the Self-Attention layer, AAAM outperforms the state-of-the-art method by 2 - 3%. The above results show that the proposed Self-Attention layer can remarkably improve the regional features ability of the baseline, and achieve the semantic consistency of the labels.

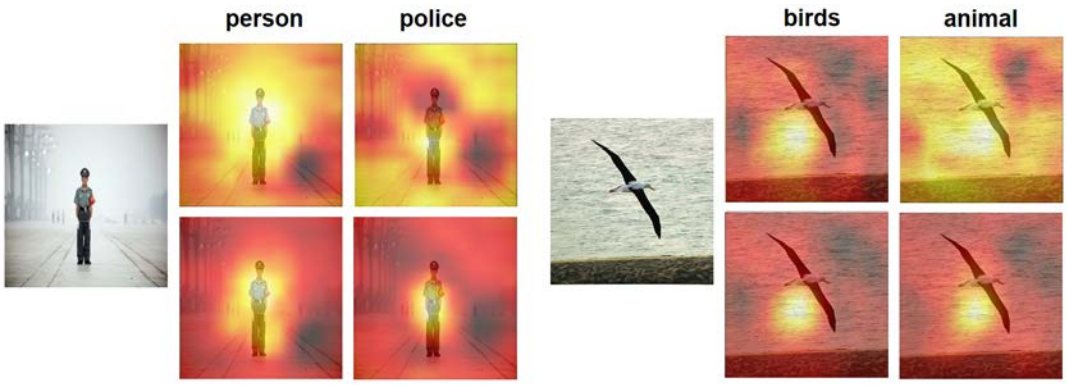

Fig. 5. Comparison between the baseline and proposed method on regional features. The first row of each group show the result using baseline and the second row proposed method.

6 Conclusion

In this paper, we propose the AAAM model for image annotation. To detect those visually hard-to-detect targets, we fuse image/label and label/label dependencies, where apply BCE and TM loss to model two dependencies respectively. Specially, we use BCE to find these visually easy-to-detect targets, and then find those hard-to-detect ones based on the label/label dependencies between them. This procedure forms an adaptive prediction path. Besides, in order to improve the ability to extract regional feature representations, we propose the self-attention layer, which enhances the relevant regions and restrains those irrelevant ones. Experimental results on the three datasets demonstrate that the proposed approach achieves superior performance to the state-of-the-art methods. However, predicting abstract annotation is still challenging due to the great chasm between visual and semantic information. We will investigate that in our future work.

Acknowledgement

This work has been supported by the National Key R&D Program of China under Grant NO.2017YFB1401000 and the Key Laboratory of Digital Rights Services, which is one of the National Science and Standardization Key Labs for Press and Publication Industry.

References

[1] YongHeng Chen, Fuquan Zhang and WanLi Zuo, “Deep Image Annotation and Classification by Fusing Multi-Modal Semantic Topics,” *KSII Transactions on Internet and Information Systems*, vol. 12, no. 1, pp. 392-412, 2018. Article (CrossRef Link).

[2] Minxian Li, Jinhui Tang and Chunxia Zhao, "Active Learning on Sparse Graph for Image Annotation," *KSII Transactions on Internet and Information Systems*, vol. 6, no. 10, pp. 2650-2662, 2012. Article (CrossRef Link).
[3] Bin Wang and Yuncai Liu, "Collaborative Similarity Metric Learning for Semantic Image Annotation and Retrieval," *KSII Transactions on Internet and Information Systems*, vol. 7, no. 5, pp. 1252-1271, 2013. Article (CrossRef Link).
[4] Wang M, Xia X, Le J, et al., "Effective automatic image annotation via integrated discriminative and generative models," *Information Sciences*, vol. 262, pp. 159-171, 2014. Article (CrossRef Link).
[5] Yonghao He, Jian Wang, Cuicui Kang, et al., "Large scale image annotation via deep representation learning and tag embedding learning," in *Proc. of the 5th ACM on International Conference on Multimedia Retrieval*, pp. 523-526, June 23-26, 2015. Article (CrossRef Link).
[6] Venkatesh N Murthy, Subhransu Maji, and R Manmatha, "Automatic image annotation using deep learning representations," in *Proc. of the 5th ACM on International Conference on Multimedia Retrieval*, pp. 603-606, June 23-26, 2015. Article (CrossRef Link).
[7] Changhu Wang, Shuicheng Yan, Lei Zhang, et al., "Multi-label sparse coding for automatic image annotation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1643-1650, June 20-25, 2009. Article (CrossRef Link).
[8] S. Hamid Amiri, Mansour Jamzad, "Automatic image annotation using semi-supervised generative modeling," *Pattern Recognition*, vol. 48, no. 1, pp. 174-188, 2015. Article (CrossRef Link).
[9] Jiang Wang, Yi Yang, Junhua Mao, et al., "CNN-RNN: A unified framework for multi-label image classification," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2285-2294, June 27-30, 2016. Article (CrossRef Link).
[10] Shiliang Zhang, Qi Tian, Guang Hua et al., "ObjectPatchNet: Towards scalable and semantic image annotation and retrieval," *Computer Vision and Image Understanding*, vol. 118, pp. 16-29, 2014. Article (CrossRef Link).
[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al., "Microsoft coco: Common objects in context," in *Proc. of European Conference on Computer Vision*, pp. 740-755, September 6-12, 2014. Article (CrossRef Link).
[12] Michael Grubinger, Paul Clough, Henning Muller, et al., "The IAPR TC-12 benchmark: A new evaluation resource for visual information systems," in *Proc. of International Workshop OntoImage*, pp. 13-23, May 22-23, 2006.
[13] Tat-Seng Chua, Jinhui Tang, Richang Hong, et al., "NUS-WIDE: A Real-World Web Image Database from National University of Singapore," in *Proc. of ACM International Conference on Image and Video Retrieval*, pp. 48, July 8-10, 2009. Article (CrossRef Link).
[14] Scott Deerwester, "Improving information retrieval with latent semantic indexing," *Information Sciences*, vol. 100, no. 1-4, pp. 105-137, 1988.
[15] Thomas Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1-2, pp. 177-196, 2001. Article (CrossRef Link).
[16] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003. Article (CrossRef Link).
[17] Yin Zheng, Yu-Jin Zhang, and Hugo Larochelle, "Topic modeling of multimodal data: an autoregressive approach," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1370-1377, June 23-28, 2014. Article (CrossRef Link).
[18] Sunho Park and Seungjin Choi, "Max-margin embedding for multi-label learning," *Pattern Recognition Letter*, vol. 34, no. 3, pp. 292-298, 2013. Article (CrossRef Link).
[19] Yunchao Gong, Yangqing Jia, Thomas Leung, et al., "Deep convolutional ranking for multilabel image annotation," *arXiv: 1312.4894 [cs]*, 2014. Article (CrossRef Link).
[20] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, et al., "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. of IEEE International Conference on Computer Vision*, pp. 309-316, September 29 - October 2, 2009. Article (CrossRef Link).

[21] Jiajun Wu, Yinan Yu, Chang Huang, "Deep multiple instance learning for image classification and auto-annotation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3460-3469, December 13-16, 2015. Article (CrossRef Link).
[22] Jiren Jin and Hideki Nakayama, "Annotation order matters: Recurrent image annotator for arbitrary length image tagging," in *Proc. of the IEEE International Conference on Pattern Recognition*, pp. 2452-2457, December 4-8, 2016. Article (CrossRef Link).
[23] Jiang Wang, Yi Yang, Junhua Mao, et al., "CNN-RNN: A unified framework for multi-label image classification," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2285-2294, June 27-30, 2016. Article (CrossRef Link).
[24] Venkatesh N Murthy, Subhransu Maji, and R Manmatha, "Automatic image annotation using deep learning representations," in *Proc. of the 5th ACM on International Conference on Multimedia Retrieval*, pp. 603-606, June 23-26, 2015. Article (CrossRef Link).
[25] Tiberio Uricchio, Lamberto Ballan, Lorenzo Seidenari, et al., "Automatic Image Annotation via Label Transfer in the Semantic Space," *Pattern Recognition*, vol. 71, pp. 144-157, 2017. Article (CrossRef Link).
[26] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas, "Random k-Labelsets for Multi-Label Classification," *IEEE Transactions on Knowledge and Data Enginerring*, vol. 23, pp. 1079 - 1089, 2011. Article (CrossRef Link).
[27] Yunchao Wei, Wei Xia, Junshi Huang, "CNN: Single-label to multi-label," *arXiv: 1406.5726 [cs]*, 2014.
[28] Zhe Wang, Limin Wang, Yali Wang, et al., "Weakly Supervised PatchNets: Describing and Aggregating Local Patches for Scene Recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 2028-2041, 2017. Article (CrossRef Link).
[29] Nicolas Usunier, David Bu [illegible] gford, and Patrick Gallinari, "Rankin classification," in *Proc. of the 26th International Conference on Machine Learning*, pp. 1057-1064, June 14-18, 2009. Article (CrossRef Link).
[30] Itti, Laurent, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998. Article (CrossRef Link).
[31] Desimone, Robert, and John Duncan, "Neural mechanisms of selective visual attention," *Annual Review of Neuroscience*, vol. 18, no. 1, pp. 193-222, 1995. Article (CrossRef Link).
[32] Raffel, Colin, and Daniel P.W. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," *arXiv: 1512.08756 [cs]*, 2015. Article (CrossRef Link).
[33] Vaswani, A., Shazeer, N., Parmar, N, et al., "Attention Is All You Need," *arXiv: 1706.03762 [cs]*, 2017. Article (CrossRef Link).
[34] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, et al., "Show, attend and tell: neural image caption generation with visual attention," in *Proc. of the 32nd International Conference on Machine Learning*, pp. 2048-2057, July 06-11, 2015. Article (CrossRef Link).
[35] Ning Sun, Feng Jiang, Hengchao Yan, et al., "Proposal generation method for object detection in infrared image", *Infrared Physics & Technology*, vol. 81, pp. 117-127, 2017. Article (CrossRef Link).
[36] Jing Liu, Tongwei Ren, Yuantian Wang, et al., "Object proposal on RGB-D images via elastic edge boxes," *Neurocomputing*, vol. 236, pp. 134-146, 2017. Article (CrossRef Link).
[37] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, et al., "Selective Search for Object Recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154-171, 2013. Article (CrossRef Link).
[38] Karen Simonyan, Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv: 1409.1556 [cs]*, 2015. Article (CrossRef Link).
[39] J. Deng, W. Dong, R. Socher, et al., "ImageNet: A large-scale hierarchical image database," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, June 20-25, 2009. Article (CrossRef Link).

[40] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proc. of the IEEE International Conference on Computer Vision*, pp. 618-626, October 22-29, 2017. Article (CrossRef Link).

[41] Baoyuan Wu , Weidong Chen, Peng Sun, Wei Liu, Bernard Ghanem, and Siwei Lyu, "Tagging like Humans: Diverse and Distinct Image Annotation," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7967-7975, June 18-22, 2018. Article (CrossRef Link).

[42] S. Hamid Rezatofighi, Vijay Kumar B G, Anton Milan, Ehsan Abbasnejad, Anthony Dick and Ian Reid, "DeepSetNet: Predicting Sets with Deep Neural Networks," in *Proc. of the IEEE International Conference on Computer Vision*, pp. 5257-5266, October 22-29, 2017. Article (CrossRef Link).

[43] Feng Liu, Tao Xiang, Timothy M. Hospedales, Wankou Yang and Changyin Sun, "Semantic Regularisation for Recurrent Image Annotation," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4160-4168, July 21-26, 2017. Article (CrossRef Link).

[44] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu and Xiaogang Wang, "Learning Spatial Regularization with Image-Level Supervisions for Multi-label Image Classification," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2027-2036, July 21-26, 2017. Article (CrossRef Link).

Fangxin Wang received the master degree in Comunicaiton and Information System from Capital Normal University in 2014. He is now the PhD candidate of Institute of Automation, Chinese Academy of Sciences (CASIA). His current research interests include pattern recognition, image annotation, image processing, and video analysis.

Jie Liu received the PhD degree in pattern recognition and intelligent systems from Institute of Automation, Chinese Academy of Sciences(CASIA) in 2011. He is now an associate research fellow at CASIA. His research interests include pattern recognition, deep learning, image processing and especially the applications to scene text detection and recognition.

Shuwu Zhang received the PhD degree from Chinese Academy of Sciences in 1997. Currently, he is a professor of Institute of Automation, Chinese Academy of Sciences. His research interests are focused on digital content analysis, digital right management, and web-based cultural content service technologies.

Guixuan Zhang received the B.S. degree in Measurement and Control Technology in 2012 from University of Science and Technology Beijing, China, and the Ph.D. degree in Pattern Recognition from University of Chinese Academy of Sciences, in 2017. He is currently an Assistant Researcher in the Institute of Automation, Chinese Academy of Sciences. His research interests include image retrieval, machine learning and pattern recognition.

Yang Zheng received the PhD degree in Control Science and Engineering from University of Science and Technology Beijing in 2018. Hi is working at Institute of Automation, Chinese Academy of Sciences(CASIA).His current research interests include pattern recognition, object detection, image processing, and video analysis.

Xiaoqian Li received the BE degree in Automation from NanKai University, China, in 2015. Currently she is working toward the PhD degree in pattern recognition and intelligent systems at the Institute of Automation of Chinese Academy of Sciences, Beijing, China. Her research interests include pattern recognition and computer vision.

Wei Liang received Ph.D. degree in Pattern Recognition and Intelligent System from Institute of Automation,Chinese Academy of Sciences(CASIA) Beijing, China, in 2006. He was a Postdoctoral Research Fellow at the North Carolina State University, Raleigh, NC, USA in 2015. Currently, he is with CASIA as an associate research fellow. His research interests include image and video siganl processing, machine learning.

Yuejun Li is now the PhD candidate of Institute of Automation, Chinese Academy of Sciences (CASIA). His research insterests include text mining, fake review detection, community detection.