Image Annotation through Adaptive Dependency Fusion

Fangxin Wang[1,2], Jie Liu[1,*], Shuwu Zhang[1,3], Guixuan Zhang[1], Yang Zheng[1], Xiaoqian Li[1,2]
1.Institute of Automation, Chinese Academy of Sciences, Beijing, China
2.University of Chinese Academy of Sciences, Beijing, China
3.AICFVE, Beijing Film Academy, Beijing, China
[*]corresponding author
email address: jie.liu@ia.ac.cn

***Abstract*—In order to improve the performance of image annotation, recently proposed methods build their model combining multiple dependencies from relations between image and label (image/label), between images (image/image) and between labels (label/label). However, most of these methods cannot make multiple dependencies work jointly, and their performances is largely depending on the results predicted by image/label dependency. To address this problem, we propose an end-to-end image annotation model to associate these dependencies with the prediction path, which is composed of a series of labels in the order they are detected. Specially, our model can adaptively adjust the prediction path: from those easy-to-detect relevant labels to these hard-to-detect relevant ones. 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.**

***Index Terms*—image annotation, multiple dependencies, end-to-end, prediction path**

I. 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 [1]–[5] built image annotation model based on three basic dependencies: relations between image and label (image/label), between images (image/image) and between labels (label/label). In these dependencies, image/label dependency possesses advantage over others at extracting discriminative features, which is very important especially for multiple targets; label/label and image/image can make the model scale to real-world large dataset. Therefore, recently researches tend to combining multiple dependencies in their models. The common practice is first obtaining the initial annotations using image/label dependency, and then applying label/label or image/image dependency to refine the annotations. Even though this methods can, to some extent, keep discriminative and optimize generalization ability, its performance is largely affected by the initial annotation, since multiple dependencies do not actually work jointly. Wang et al. [6] proposed the prediction path, a series of labels in the predefined order to be recognized, to integrate image/label with label/label dependencies. However, different prediction paths may produce very different results, and it is still a great challenge to this method since it can be very costly to find the best path.

To address these problems, we put forward an end-to-end image annotation model (showed in **Fig. 1**) to associate these dependencies with the prediction path. In particular, 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 the 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 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 relevant labels can obtain better performance than previous methods.

The main contributions of the paper are as follows:

- Instead of applying multiple dependencies in stages in the traditional paradigms, we propose an end-to-end model to perform image annotation.
- Our model can adaptively adjust the prediction path for each image that from the easy-to-detect to hard-to-detect relevant labels.
- We associate multiple dependencies with proposed adaptive prediction path, and the experimental results on the COCO 2014 [7], IAPR TC-12 [8] and NUSWIDE [9] benchmarks show that our methods outperform the state-of-the-art methods.

II. 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 [10], pLSA [11]

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.

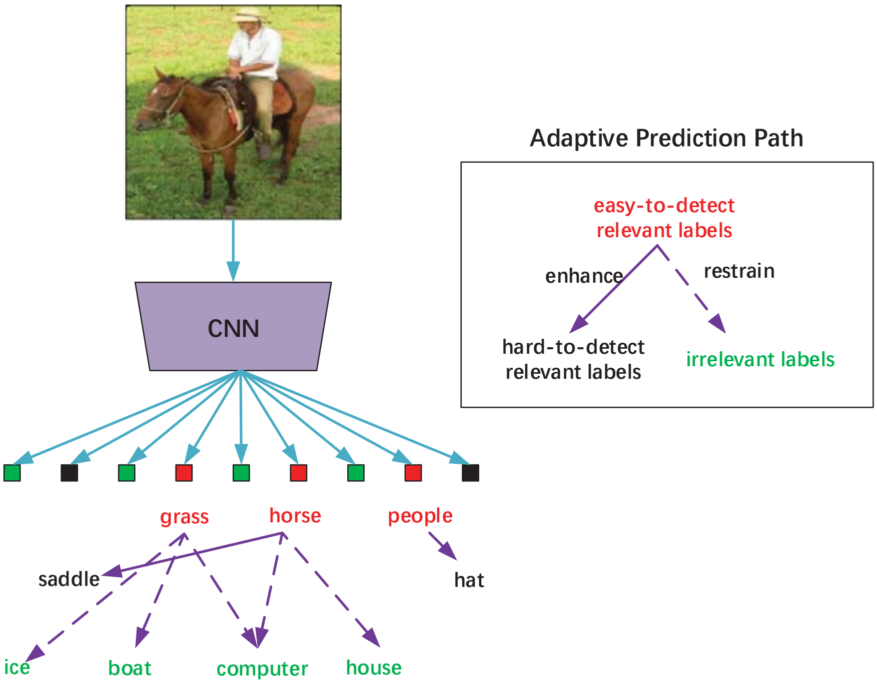

Fig. 1. The structure of the proposed image annotation models. The red and black labels are the easy-to-detect and the hard-to-detect relevant labels of the query image while the blue ones are irrelevant labels. Those easy-to-detect relevant labels, such as *people*, *horse* and *grass*, are first detected by BCE (image/label dependency). Since *horse* and *saddle* most probably appear in the same scene, even though the latter is visually hard to detect, we apply TM to force *horse* to enhance *saddle*. In the same way, we can make *horse* restrain the irrelevant labels such as *computer*.

and LDA [12] to model the joint distribution over images and labels. Zheng et al. [13] 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. [14] trained a model to build a shared feature space of both media by max-margin embedding method. Gong et al. [15] combined DNN with different top-k ranking loss functions to improve performance. Usunier et al. [16] 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. 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. [17] used Conditional Random Field to model the label/label dependency. Wu et al. [18] 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 desire. 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. [19] employed a cascading structure with Convolutional Neural Network (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. [6] put forward the CNN-RNN framework for multi-label classfication 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. Murthy et al. [20] and Uricchio et al. [21] 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.

III. Motivation

As some recent researches have proven that the fusion of multiple dependencies can further improve the annotation performance, subsequent annotation methods began taking

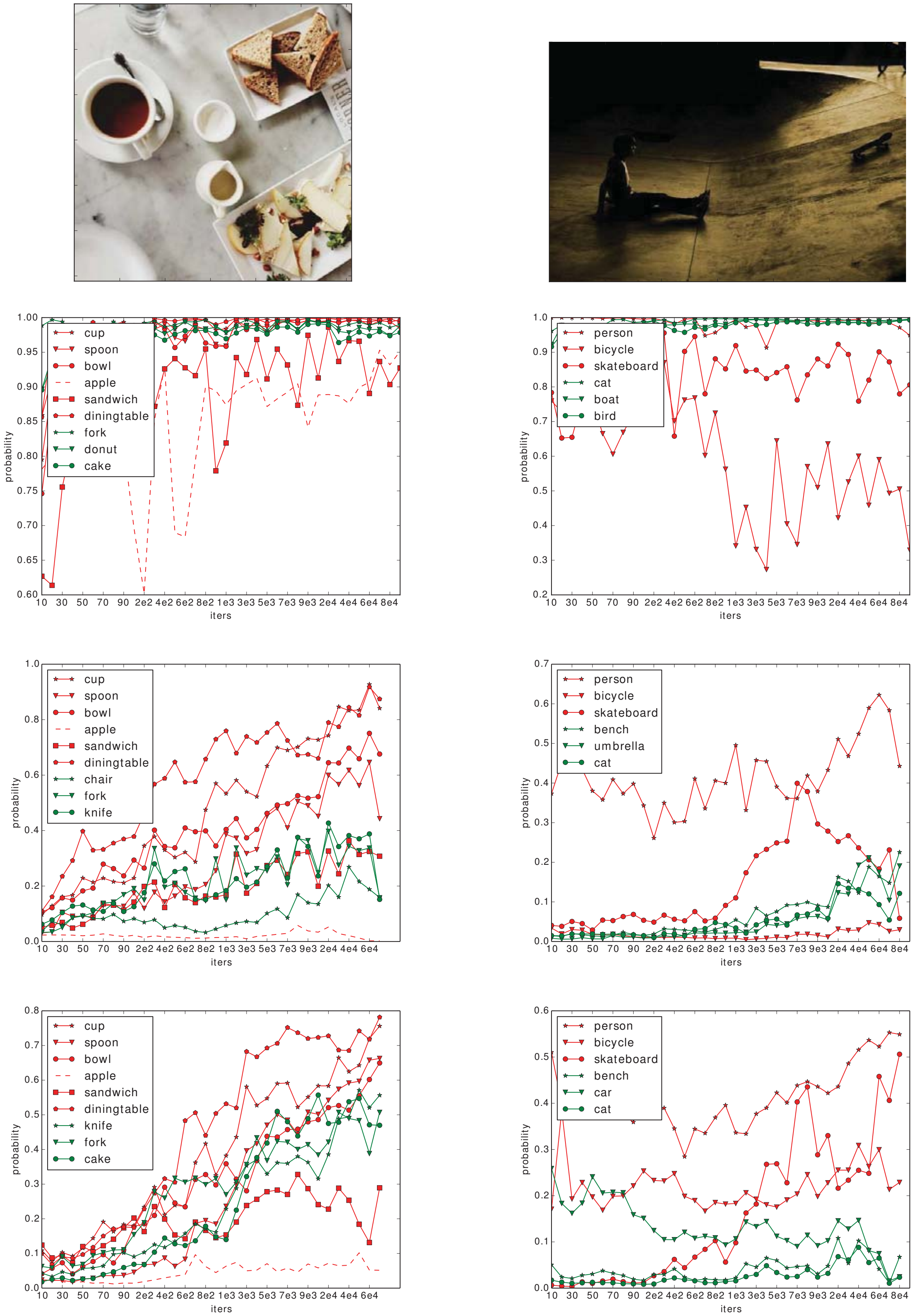

Fig. 2. Probabilities of labels for different models on COCO 2014: the 1st row shows the target images, the 2nd, 3rd and 4th rows show the results of WARP, BCE and the proposed model respectively. The red texts represent the relevant labels, and the green ones represent the irrelevant labels.

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. [6] 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 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 an 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, such as *diningtable* and *bowl*, while the other ones in the later stage, such as *bicycle* and *sandwich*, and they have different prediction paths 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 detected in the early stage the easy-to-detect relevant targets, and in the later stage the hard-to-detect relevant ones. In our paradigm, we assume that by the fusion of multiple dependencies, those easy-to-detect relevant targets can help to find the hard-to-detect relevant and filter out the irrelevant ones.

IV. The Proposed 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, where W are the parameters of the model.

A. Binary 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 score $\hat{z}_i$ 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:

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

The Binary Cross-Entropy loss optimizes the relationship between image I and label z_i, s 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 z_i is not ssociate with other predicted label $z_j (j \neq i)$, which means we can hardly detect those visually hard-to-detect targets.

B. 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 $\Delta\hat{z}_{i,k} = |\hat{z}_i - \hat{z}_k|^2$ is the score discrepancy between the relevant label y_i and irrelevant label y_k. he gradient with respect to 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 $I(\cdot)$ 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 the irrelevant pairs.

C. 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:

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

TABLE I
Configurations of evaluation datasets

Configurations	**Datasets**		
	COCO 2014	***IAPR TC-12***	***NUSWIDE***
No. of Images	122585	19627	209347
No. of Labels	80	201	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]

TABLE II
Image Annotation Results on COCO 2014 (k=3)

Methods	**Metrics (%)**		
	Precision@k	***Recall@k***	***mAP***
Softmax	59.00	57.00	50.65
BCE	59.30	**58.60**	57.90
WARP	59.30	52.50	54.80
CNN-RNN	66.00	55.60	-
Ours	**67.35**	53.56	**64.33**

where α s designed to make a trade-off between the two losses. With the integration of two dependencies, the performance of these later predicted objects can be boosted by those previous predicted targets.

V. Experimants

A. 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 showed in **Table I**. 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 distributions of labels, we also show the minimum and maximum number of labels per images [minimum, maximum].

B. 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 I** 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. As a consequence, 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.

TABLE III
Image Annotation Results on IAPR TC-12 (k=5)

Methods	**Metrics (%)**		
	Precision@k	***Recall@k***	***mAP***
SVM	31.00	29.00	34.00
BCE	40.18	32.95	37.36
WARP	36.72	27.78	34.99
kNN	39.00	29.00	36.00
KCCA + kNN	44.00	34.00	40.00
Ours	**49.65**	**37.24**	**42.55**

TABLE IV
Image Annotation Results on NUSWIDE (k=3)

Methods	**Metrics (%)**		
	Precision@k	***Recall@k***	***mAP***
Logistic	40.90	43.12	45.78
SVM	34.60	**60.60**	50.20
BCE	41.14	42.87	51.03
WARP	31.65	35.60	39.21
kNN	39.60	44.00	49.30
KCCA + kNN	40.20	50.50	51.70
CNN-RNN	40.50	30.40	-
Ours	**47.10**	44.56	**53.66**

C. Results

In our implementation, we adopt VGG16 [22] network pretrained on ImageNet [23] as the image extractor, which is also used in related methods. And the triplet sampler will randomly chooses a top-ranking relevant tag y_i, a lower-ranking relevant tag y_j and a top-ranking irrelevant tag y_k, respectively each time, so as to find out those visually hard-to-detect relevant targets. For the training processes, we use ADAM to optimize our model, and set the learning rate in feature layer 1×10^{-2}, and classifier layer 1×10^{-3}, weight decay rate 1×10^{-5}, and dropout rate 0.5. All the parameters involved are obtained through cross-validation. Besides, at the beginning, we train the network only with BCE loss for better prediction of those easy-to-detect targets, and after 30 epochs, we adopt the loss in (6) to make the two losses work jointly to find those hard-to-find targets. For the missing results on some datasets, we re-implement BCE and WARP methods.

We show the image annotation results on COCO 2014, IAPR TC-12 and NUSWIDE datasets in **Table II, III** and **IV**, respectively. In order to make an overall analysis on different dependencies, we carefully choose some models to compare with the proposed method: Softmax, Logistic, KCCA and BCE consider the image/label dependency; WARP gives consideration to the label/label dependency; kNN addresses the image/image dependency. In order to confirm the advantages of dependeny fusion, we also compare the performance of fusion of multiple dependencies: KCCA + kNN combine image/label and image/image dependencies. CNN-RNN takes both image/label and label/label dependencies into account.

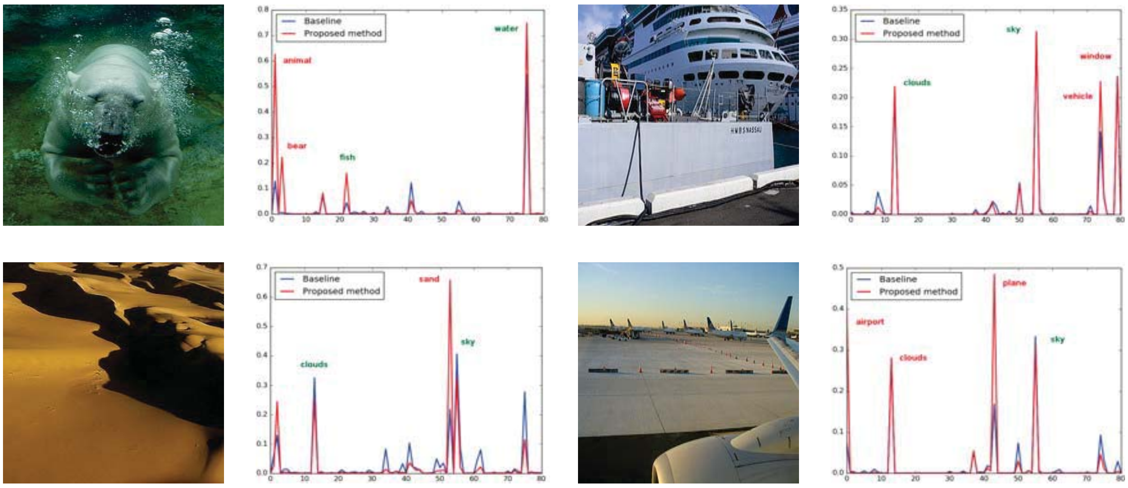

Fig. 3. Comparison between the baseline and the proposed model on image annotation, where red texts represent the relevant labels and green ones the irrelevant labels.

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 mage/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 kNN, can obtain 2% promotion, which proves that multiple dependencies can help to boost the detection performances.

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. Our method increases the mAP by 7%, 5% and 2% in three datasets respectively, compared to the baseline BCE.

In order to illustrate the effect of two dependencies to the prediction path, we show how the predictions change over time under three 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 models, respectively. Specially, after obtaining 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 *sandwich*, *cake* and *donut* (in the 1st query image), which leads to incorrect results. In contrast, BCE (the 3rd row) can distinguish semantic similar objects, but its inferences often cannot detect those visually hard-to-detect relevant objects such as *bicycle* (in the 2nd 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. 3** 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.

VI. Conclusion

In this paper, we propose an image annotation though adaptive dependency fusion. To detect those visually hard-to-detect relevant 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 relevant ones based on the label/label dependencies between them. This procedure forms an adaptive prediction path. 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.

References

[1] M. Wang, X. Xia, J. Le, and X. Zhou, "effective automatic image annotation via integrated discriminative and generative models," *Information Sciences*, vol. 3, no. 262, pp. 159-171, 2014.

[2] Yonghao He, Jian Wang, Cuicui Kang, Shiming Xiag, and Chunhong Pan, "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.

[3] 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.

[4] Changhu Wang, Shuicheng Yan, Lei Zhang, and Hongjiang Zhang, “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.

[5] S. Hamid Amiri and Mansour Jamzad, “Automatic image annotation using semi-supervised generative modeling”, Pattern Recognition, vol. 48, no. 1, pp. 174-188, 2015.

[6] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu, “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.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays et al. “Microsoft coco: Common objects in context,” in *Proc. of European Conference on Computer Vision*, pp. 740-755, September 6-12, 2014.

[8] Michael Grubinger, Paul Clough, Henning Muller, and Thomas Deselaers, “The IAPR TC-12 benchmark: A new evaluation resource for visual information systems,” in *International Workshop OntoImage*, pp. 13-23, May 22-23, 2006.

[9] Tat-Seng Chua, Jinhui Tang, Richang Hong, and Haojie Li, “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.

[10] Scott Deerwester, “Improving information retrieval with latent semantic indexing,” *Information Sciences*, vol. 100, no. 1-4, pp. 105-137, 1988.

[11] Thomas Hofmann, “Unsupervised learning by probabilistic latent semantic analysis,” *Machine Learning*, vol. 42, no. 1-2, pp. 177-196, 2001.

[12] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2012.

[13] 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 27-30, 2016.

[14] Sunho Park and Seungjin Choi, “Max-margin embedding for multi-label learning,” *Pattern Recognition Letter*, vol. 34, no. 3, pp.292-298, 2013.

[15] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe, “Deep convolutional ranking for multilabel image annotation,” *arXiv*: 1312.4894 [cs], 2014.

[16] Nicolas Usunier, David Buffoni, and Patrick Gallinari, “Ranking with ordered weighted pairwise classification,” in *Proc. of the 26th International Conference on Machine Learning*, pp. 1057-1064, June 14-18, 2009.

[17] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid, “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.

[18] Jiajun Wu, Yinan Yu, and 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.

[19] 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.

[20] 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.

[21] Tiberio Uricchio, Lamberto Ballan, Lorenzo Seidenari, and Alberto Del Bimbo, “Automatic Image Annotation via Label Transfer in the Semantic Space,” *Pattern Recognition*, vol. 71, pp. 144-157, 2017.

[22] Karen Simonyan, Andrew Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv*: 1409.1556 [cs], 2015.

[23] J. Deng, W. Dong, R. Socher, 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.