Image Tagging by Joint Deep Visual-Semantic Propagation

Yuexin Ma[1], Xinge Zhu[2], Yujing Sun[1], and Bingzheng Yan[3]

[1] University of Hong Kong, Hong Kong, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] National University of Defense Technology, Changsha, China
{yxma,yjsun}@cs.hku.hk
zhuxinge15@mails.ucas.ac.cn,bingzhengyan93@gmail.com

Abstract. Image tagging has attracted much research interest due to its wide applications. Many existing methods have gained impressive results, however, they have two main limitations: 1) only focus on tagging images, but ignore the tags' influences on visual feature modeling. 2) model the tag correlation without considering visual contents of image. In this paper, we propose a joint visual-semantic propagation model (JVSP) to address these two issues. First, we leverage a joint visual-semantic modeling to harvest integrated features which can accurately reflect the relationship between tags and image regions. Second, we introduce a visual-guided LSTM to capture the co-occurrence relation of the tags. Third, we also design a diversity loss to enforce that our model learns to focus on different regions. Experimental results on three challenging datasets demonstrate that our proposed method leads to significant performance gains over existing methods.

Keywords: Image Tagging · CNN-LSTM · Visual-Semantic

1 Introduction

Automatic image tagging aims to associate images with appropriate tags which reflect visual contents in the image. The problem is difficult because every real-world image usually contains multiple labels and has intricately spatial layout. Since the scale of visual data is growing fast and manually tagging images is expensive and time-consuming, there is a huge demand for image tagging.

Every tag is closely related to specific region of image. Modeling the connection between the tag and corresponding image region plays an important role in image tagging. Existing methods that model the relationship between images and tags can be roughly divided into parametric and non-parametric methods. One of the most popular parametric methods is the generative model [1], which is usually dedicated to maximize generative likelihood of image features and tags. On the other hand, many non-parametric nearest neighbor models [13] have been proposed. They transfer tags from nearest neighbor images to the query image based on the assumption that visually similar images are more likely to share

common labels. However, most existing approaches leave the tag information in the input feature space untouched and ignore the tags' influences on visual feature modeling. They are not rich enough to mine the complicated relationships between local image regions and tags. Moreover, different regions of the image have different weights with respect to different tags. Without considering tags' effect, these methods giving whole image the same weights are hard to recognize some objects, especially some small objects.

There is an observation that the tags associated with natural images do not appear in isolation, they appear correlatively and interact with each other. We refer it as tag correlation. For instance, the probability of an image being labeled with "sea" would be high if the image has been annotated with "boat" and "island". Instead of simply learning independent binary classifiers for each tag, many methods try to exploit tag correlation for improving tagging performance. These methods consider the tag correlation as additional information, which can be learned via sparse learning [1], graphical model [11], or dictionary learning [14]. However, two main drawbacks of these models mining the tag correlation need to be concerned. First, they explore the tag correlation in semantic level without considering visual contents of image. This manner causes the tag correlation less flexible when recognizing particular image, and has a negative effect if there exists remarkable tag correlation discrepancy between the training and test sets. Second, due to huge computational complexity, most methods only model the pairwise tag correlation rather than long range correlation among tags. It can not be able to well capture the intricate tag correlation.

Inspired by the great success from convolutional neural network (CNN) in image classification [5], many methods try to extend the CNN to multi-label image tagging [7]. In addition, recurrent neural networks have proven to be able to model the long range temporal dependencies [17]. Therefore, recent works [8] try to employ the RNN to model tag correlation for multi-label problem. However, these methods also suffer from aforementioned problems, including leaving the relationship between tags and visual features untouched in the input feature space and ignoring the effects of visual contents in the image during tag correlation exploitation.

In this paper, we propose a joint visual-semantic propagation (JVSP) model to address the foregoing problems. First, to sufficiently exploit the tag information in the input feature space, we introduce a joint visual-semantic modeling to explore the tag representation and visual features, and make them benefit from each other in a collaborative way. There are two steps in this process: 1) we incorporate the tag embedding into image features to obtain the integrated visual features, which characterize the relationship between tag and corresponding image region. 2) based on step one, we in turn refine the tag representation by fusing the integrated visual features into tag embedding. The refined tag representation will be less ambiguous and have a stronger connection with visual content of the image. Second, we introduce a visual-guided LSTM to model the tag correlation. Unlike previous methods that model tag correlation without considering image contents, we leverage the integrated visual features as an instruction, and feed

it into LSTM with refined tag representation together to capture the complex correlation among tags. Third, we design a diversity loss to enforce the proposed model concentrates on different image regions for extracting more discriminative features with respect to different labels. Finally, in this way, we cast the image tagging task into a visual-semantic propagation problem. Based on corresponding image regions and tag relation, we can propagate multiple tags by the visual-guided LSTM in an iterative way. We conduct extensive experiments on three large-scale datasets, showing that JVSP consistently outperforms the existing methods.

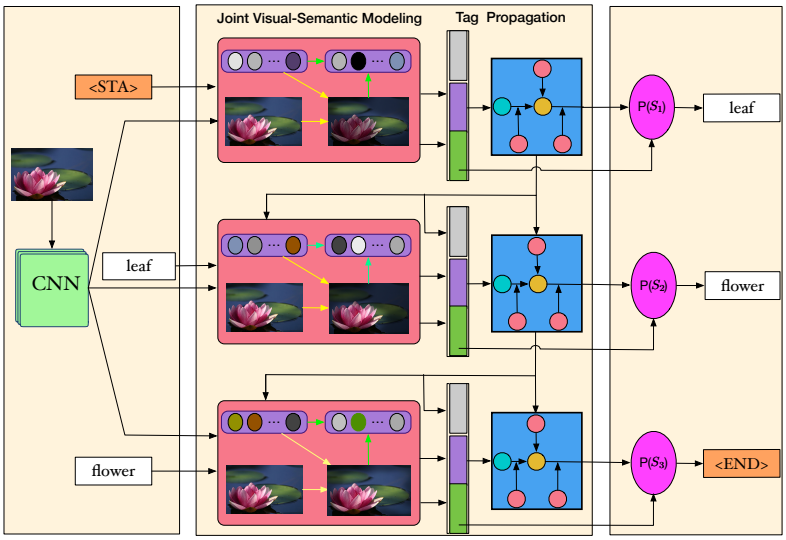

Fig. 1. Schematic of the proposed model. We unfold a complete tag propagation procedure of a two-label image to get three steps. As the pipeline shows, our model first extracts the image features from CNN and computes the tag embedding, then the joint visual-semantic modeling explores and integrates both of them. Finally, the visual-guided LSTM is applied to model the tag correlation and perform the tag propagation. In this way, we build an automatic framework to propagate tags in an iterative manner.

2 Proposed Method

We show an example pipeline of the proposed model in Figure 1. As shown in Figure 1, the overall architecture of the model consists of three components: feature extraction, joint visual-semantic modeling, and tag propagation.

2.1 Joint Visual-Semantic Modeling

To overcome the issue existing in most previous methods: ignoring the effect of tags on image feature representation, we introduce the joint visual-semantic modeling. Figure 2 describes the detailed structure of the joint visual-semantic

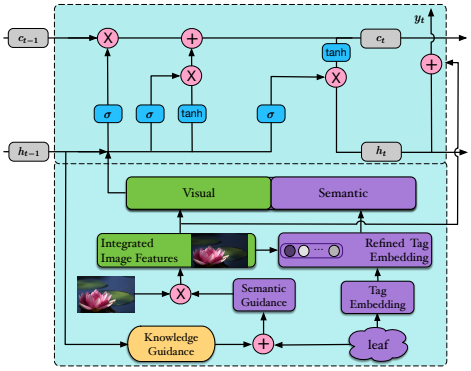

Fig. 2. The details of the joint visual-semantic modeling (bottom panel) and tag propagation (top panel).

modeling and tag propagation module. At time step t, a tag k is represented as a one-hot vector e_k. There are two different tag embedding matrices, U_l and U_f. Tag embeddings can be obtained by multiplying these two tag embedding matrices.

$$w_k = U_l e_k, w_f = U_f e_k \tag{1}$$

where w_k is used as the semantic guidance and w_f is applied to form the refined tag representation.

Since one tag propagation relies on its predecessors and the effect of previous tags is getting weaker and weaker with time steps, we introduce the **knowledge guidance**, which consists of preceding hidden states h_i (from step 1 to step $t-1$) to enhance the role of previous tags. It is beneficial to generate the semantic guidance for current tag. The knowledge guidance is defined as follows:

$$K_t = \frac{1}{t-1}(W_h \sum_{i=1}^{t-1} \lambda_i h_i) \tag{2}$$

Where λ_i is the maximum tag probability at time step i. Learnable matrix W_h is used to embed the knowledge guidance into the same vector space as w_k. We then obtain the **semantic guidance vector** g_t:

$$g_t = w_k + K_t \tag{3}$$

We regard the semantic guidance as a weighted mask, and then earn the integrated image features by following function:

$$V_t = \Phi(C(I) \odot g_t) \tag{4}$$

where V_t is the integrated image features at time step t, and $C(I)$ is the original image features extracted from CNN. $\odot$ represents the element-wise multiplication, and $\Phi(\cdot)$ denotes a non-linear transfer function, *i.e.*, Softmax, which is chosen based on cross-validation. As shown in Figure 2, the integrated image

features have high activations on specific regions of image under the semantic guidance.

After generating integrated image features, we in turn utilize it to refine the tag embedding. The operation can be expressed in the following steps:

$$\hat{g}_t = W_{gt}^T \tanh(w_f + U_{ig} V_t) \tag{5}$$

$$T_t = \Phi(w_f \odot \hat{g}_t) \tag{6}$$

where U_{ig} and W_{gt} are learnable weight matrices. V_t is the integrated image features obtained from Eq. 4. Furthermore, V_t is a visual guidance to refine the tag embedding. For a label of multiple senses, *e.g.*, "mouse", the direct tag embedding will always project it into same vector regardless of its senses in the image. In contrast, V_t can guide the tag embedding with its real senses in the image because of the differences in visual features. Thus, we get the weighted vector $\hat{g}_t$ by incorporating the integrated image features. We also apply dropout on $U_{ig}V_t$ in the training phase to overcome overfitting problem. Then we compute the refined tag representation T_t according to Eq. 6.

Finally, we combine the integrated image features and refined tag representation into a feature concatenation.

$$X_t = [V_t, T_t] \tag{7}$$

where $[\cdot]$ is the concatenation operation on two vectors.

2.2 Tag Propagation by Visual-Guided LSTM

Unlike the normal LSTM which only takes the text representation as the input, we introduce a visual-guided LSTM, which accepts the integrated image features as the guidance in both the input and output space to instruct the tag propagation. [17] points out that as the time step continues, the generation result of LSTM becomes "blind", for the reason that the role of image representation, which is only fed once at the beginning, becomes weaker and weaker. To figure out this problem, they extend the LSTM with semantic information which is a global guidance and does not change with time step. In our model, we employ the integrated image features as the visual guidance and feed it into LSTM at each time step. More importantly, the visual guidance is different from the global guidance mentioned above because it is relevant to current tag which is varying with time step.

The visual-guided LSTM neuron in our implementation is illustrated in the top panel of Figure 2. The interaction between states and gates is defined as follows:

$$\begin{aligned}
i_t &= \sigma(W_{iv}V_t + W_{it}T_t + W_{ih}h_{t-1}) \\
f_t &= \sigma(W_{fv}V_t + W_{ft}T_t + W_{fh}h_{t-1}) \\
o_t &= \sigma(W_{ov}V_t + W_{ot}T_t + W_{oh}h_{t-1}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \delta(W_{cv}V_t \\
&\quad + W_{ct}T_t + W_{ch}h_{t-1}) \\
h_t &= o_t \odot \delta(c_t)
\end{aligned} \tag{8}$$

where i_t stands for the input gate, f_t for the forget gate, o_t for the output gate. c_t denotes the state of the memory cell and h_t encodes the hidden state. All these variables are the key points to learn long-term correlation. $\sigma(\cdot)$ represents the sigmoid function and $\delta(\cdot)$ denotes the hyperbolic tangent function. $W_{[\cdot]}$ stand for the learnable weight matrices. The visual guidance V_t is the integrated image features, and T_t is the refined tag representation. Both of them are obtained from E.q 7.

The visual-guided LSTM generates label scores based on the hidden state h_t and visual guidance V_t. As the visual guidance is capable of describing the high-level visual concepts from specific regions, it is conducive to the tag prediction.

$$y_t = W_s^T(h_t + V_t) \tag{9}$$

where y_t represents the score vector of labels at time step t and W_s is a projection matrix. The predicted tag probability can be computed using the Softmax on the score y_t.

2.3 Diversity Loss

To make the focused region vary with time steps, we propose the diversity loss to compute the correlation between neighboring integrated image features:

$$\begin{aligned} \mathcal{L}_V =& \frac{1}{T-1}\sum_{t=2}^{T}\sum_{i=1}^{M} V_{t-1,i} \cdot V_{t,i} \\ s.t. \quad & V_{[\cdot,\cdot]} \geq \beta \end{aligned} \tag{10}$$

where $V_{t,i}$ is the i-th value of the integrated image features at time step t. T denotes the total number of time steps and M represents the length of integrated image features. β is a given threshold. Values larger than β can be regarded as high activation. In general, $\mathcal{L}_V$ will get a large value if the focused regions between two temporally adjacent image features are similar. We then apply the negative log probability as the classification loss. The final loss function is obtained by combining both of them:

$$L = \frac{1}{T}\sum_{t=1}^{T} -\log \frac{\exp(\hat{y}_t)}{\sum_{j=1}^{\phi}\exp(y_t^j)} + \alpha\mathcal{L}_V \tag{11}$$

where y_t follows E.q (9) and the formulation of $\mathcal{L}_V$ follows E.q (10). $\hat{y}_t$ is the score of the predicted tag at time step t. ϕ is the size of label vocabulary. α is a constant weighting factor.

3 Experiments

We evaluate our proposed model on three widely used datasets for image tagging: ESP Game dataset [16], NUS-WIDE dataset [6] and MS-COCO dataset [15].

For fairness of comparison, we follow previous work [13] to evaluate methods. The precision (P) and recall (R) of all labels are applied as evaluation

metrics. Based on these two measures, we also compute the F1-score $F1 = 2 * P * R/(P + R)$, which takes care of the trade-off between precision and recall. Additionally, the number of tags with non-zero recall value ($N+$) and the mean average precision (MAP) measure are reported for evaluation.

3.1 Implementation Details

The visual features are extracted from VGG network[3] pretrained on the ImageNet dataset [4]. The dimensionality of image features and tag embedding are both set to 512. We apply fixed input state dimension of 1024 and hidden state dimension of 512 for all visual-guided LSTM neurons. Learning rate is scheduled as 10^{-3} and a staircase weight decay is applied after a few epochs. β_1, β_2 in adam are set to 0.8 and 0.999. We set the threshold $\beta = 0.01$ and the weighting factor $\alpha = 0.5$ based on cross-validation. Dropout with rate 0.5 is applied on the top of output state in the LSTM neuron. For testing images, we annotate each of them with top k labels.

3.2 Performance on NUS-WIDE

Method	C-P	C-R	C-F1	O-P	O-R	O-F1	MAP
Multi-edge graph[2]	-	-	-	35.0	37.0	36.0	-
KNN[6]	32.6	19.3	24.3	42.9	53.4	47.6	-
Softmax	31.7	31.2	31.4	47.8	59.5	53.0	-
WARP[7]	31.7	35.6	33.5	48.6	60.5	53.9	-
CNN-RNN[8]	**40.5**	30.4	34.7	49.9	61.7	55.2	56.1
JVSP	33.1	**46.2**	**38.5**	**52.9**	**65.8**	**58.6**	**68.5**

Table 1. Performance evaluation on NUS-WIDE dataset for $k = 3$.

NUS-WIDE dataset is widely used as the benchmark for image tagging and multimedia retrieval due to its large amount of images and high quality annotations. As each label has enough training images, we put the more frequency tags before the less frequency ones in the training time. Not only the precision and recall scores (C-P and C-R) over all classes, but we also report the overall precision (O-P), overall recall (O-R) and overall F1 (O-F1) scores, where the average is taken over all testing images.

We compare the performance of our method against several state-of-the-art approaches, *i.e.* KNN [6], WARP [7], CNN-RNN [8]. Table 1 shows the comparison results. Comparing with these methods using the learning to rank framework *i.e.* WARP [7], the proposed method obtains superior results in all evaluation terms. As compared with [8] that proposes a CNN-RNN based model, our approach can gain a favorable performance improvement, which indicates the effectiveness of jointly considering visual features and tag representation. In particular, we can see our method achieves 16% higher recall and 12% higher

MAP than the CNN-RNN based method, because our model can focus on local image regions to precisely recognize some objects, especially small objects, which is the key to significantly improve the performance. Furthermore, we also quantitatively verify the capability of recognizing small objects in Section 3.3.

3.3 Performance on MS-COCO

In the MS-COCO dataset, we view the object annotations as the labels. There are 80 object types and each label has enough training images, so we use the same frequent-first strategy as in the NUS-WIDE dataset. As the number of objects per image varies considerably, we do not set the limitation of minimum k time steps during tag propagation in the testing stage. We measure the performance in the same setting as NUS-WIDE dataset.

Table 2 compares the performance of the proposed method to existing approaches on MS-COCO dataset. Several competitors participate in the comparison, including WARP [7], multi-label binary cross entropy, and CNN-RNN model [8]. From Table 2, it is obvious that the proposed method achieves better performances compared with other methods.

Method	C-P	C-R	C-F1	O-P	O-R	O-F1	MAP
Softmax	59.0	57.0	58.0	60.2	62.1	61.1	47.4
WARP[7]	59.3	52.5	55.7	59.8	61.4	60.7	49.2
Binary cross-entropy	59.3	**58.6**	58.9	61.7	65.0	63.3	-
No RNN [8]	65.3	54.5	59.3	68.5	61.3	65.7	57.2
CNN-RNN[8]	66.0	55.6	60.4	69.2	**66.4**	67.8	61.2
JVSP	**67.6**	58.1	**62.5**	**72.3**	65.7	**68.8**	**62.7**

Table 2. Performance evaluation on MS-COCO dataset.

Most previous methods give the whole image the same weights. This setting harms the performances, especially when the image contains some small objects. Rather than using the whole image with uniform weights, our model incorporates tag information into the visual features to highlight different regions conditioned on different tags. To verify the efficiency of recognizing small objects in our model, we sample some small objects in the MS-COCO dataset, including "bottle", "cell phone", "bowl", "spoon", "fork", "knife" and "mouse", then compare their per-class precision and recall scores with the CNN-RNN [8]. The comparison results are shown in Figure 3. It can be observed that JVSP gains a significant improvement than CNN-RNN [8] among these small objects.

3.4 Performance on ESP Game

ESP Game dataset contains 268 different labels, and some labels include only less than 100 training images. Since some rare labels are easier to be ignored, we raise the priority of the rarer labels in the training stage and apply a rare-first order of labels. For a fair comparison, we annotate each image with 5 tags.

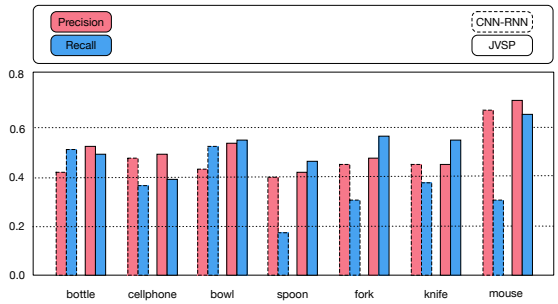

Fig. 3. Performances on small objects between CNN-RNN [8] and JVSP.

Method	Features	P	R	F1	N+
TagProp[13]	HC	39	27	32	239
2PKNN[12]	HC	**51**	23	31.7	245
SFR[1]	HC	37	24	29	238
SLED[14]	HC	49	30	37	253
CNN-R[10]	Caffe-Net	44.5	28.5	34.7	248
2PKNN	VGG-16	40	23	29	250
RIA [9]	VGG-16	32	32	31	249
CCA-KNN[10]	VGG-16	46	36	40.4	**260**
JVSP	VGG-16	50.3	**36.4**	**42.2**	258

Table 3. Performance evaluation on ESP Game dataset for $k = 5$.

Table 3 shows the results of the proposed model and existing approaches on ESP Game dataset. Comparing with these models [13,12] leaving the tag correlation untouched, JVSP has gained marked improvement on R, $F1$ and $N+$. Different from these methods [1,14] which explore tag correlations without studying the visual contents of image, our model utilizes the visual-guided LSTM to capture tag co-occurrence correlation. All above methods use the hand-crafted features, and we also compare with the methods [10,12] which use deep visual features. We can see our model outperforms all other deep models in most evaluation terms. Compared with these methods [9] which applied the RNN to model the tag relation, our proposed model also achieves much better performance since we consider the tags' influences on image feature modeling and utilizes the visual features to guide the tag propagation.

4 Conclusion

In this paper, we proposed a joint visual-semantic propagation model for image tagging. The joint visual-semantic modeling exploits tag information and image features in the input feature space to mine the relationship between tag and local image region, and a visual-guided LSTM is introduced to model the tag correlation and propagate multiple tags in an iterative way. We train our model with a novel diversity loss in an end-to-end manner. We conducted experiments

on three benchmark datasets. The experimental results demonstrated that our model achieves superior performance to the state-of-the-art methods, especially for some small objects.

References

1. Sun, F., Tang, J., Li, H., Qi, G. J., Huang, T. S.: Multi-label image categorization with sparse factor representation. IEEE TIP, 23(3), 1028-1037 (2014)
2. Liu, D., Yan, S., Rui, Y., Zhang, H. J.: Unified tag analysis with multi-edge graph. In Proceedings of the 18th ACM international conference on Multimedia (pp. 25-34) (2010)
3. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
4. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In CVPR 2009. (pp. 248-255) (2009)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In Proceedings of CVPR (pp. 770-778) (2015)
6. Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from National University of Singapore. In Proceedings of the ACM international conference on image and video retrieval (p. 48) (2009)
7. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep convolutional ranking for multilabel image annotation. arXiv preprint arXiv:1312.4894 (2013)
8. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE CVPR (pp. 2285-2294) (2016)
9. Jin, J., Nakayama, H.: Annotation order matters: Recurrent image annotator for arbitrary length image tagging. arXiv preprint arXiv:1604.05225 (2016)
10. Murthy, V. N., Maji, S., Manmatha, R.: Automatic image annotation using deep learning representations. In Proceedings of the 5th ACM on ICMR (pp. 603-606) (2015)
11. Wang, Hua, Heng Huang, and Chris Ding.: "Image annotation using multi-label correlated green's function." IEEE ICCV, (2009)
12. Verma, Y., Jawahar, C. V.: Image annotation using metric learning in semantic neighbourhoods. In ECCV (pp. 836-849) (2012)
13. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In ICCV (pp. 309-316) (2009)
14. Cao, X., Zhang, H., Guo, X., Liu, S., Meng, D.: SLED: Semantic label embedding dictionary representation for multilabel image annotation. IEEE TIP, 24(9), 2746-2759 (2015)
15. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L.: Microsoft coco: Common objects in context. In ECCV (pp. 740-755) (2014)
16. Von Ahn, L., Dabbish, L.: Labeling images with a computer game. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 319-326). ACM (2004)
17. Jia, X., Gavves, E., Fernando, B., Tuytelaars, T. (2015).: Guiding the long-short term memory model for image caption generation. In ICCV (pp. 2407-2415). (2015)