Social image annotation based on image captioning

HAIYU YANG, HAIYU SONG*, WEI LI, KEXIN QIN, HAOYU SHI, QI JIAO
School of Computer Science and Engineering Dalian Minzu University, Dalian, 116600 CHINA

Abstract—With the popularity of new social media, automatic image annotation (AIA) has been an active research topic due to its great importance in image retrieval, understanding, and management. Despite their relative success, most of annotation models suffer from the low-level visual representation and semantic gap. To address the above shortcomings, we propose a novel annotation method utilizing textual feature generated by image captioning, in contrast to all previous methods that use visual feature as image feature. In our method, each image is regarded as a label-vector of k user-provided textual tags rather than a visual vector. We summarize our method as follows. First, the image visual features are extracted by combining the deep residual network and the object detection model, which are encoded and decoded by the mesh-connected Transformer network model. Then, the textual modal feature vector of the image is constructed by removing stop-words and retaining high-frequency tags. Finally, the textual feature vector of the image is applied to the propagation annotation model to generate a high-quality image annotation labels. Experimental results conducted on standard MS-COCO datasets demonstrate that the proposed method significantly outperforms existing classical models, mainly benefiting from the proposed textual feature generated by image captioning technology.

Keywords—Automatic image annotation, social image, image captioning, textual modal feature.

Received: March 14, 2021. Revised: March 25, 2022. Accepted: April 26, 2022. Published: May 19, 2022.

1. Introduction

With the popularity of digital imaging devices and social media (such as WeChat and Facebook), image data is generated on a massive scale and spread on the Internet. Images have become the most important information-sharing content on social media. Automatic image annotation (AIA) techniques can learn the model from the training dataset, and predict the semantic labels for the test (unseen) image. In recent years, it has been an active research topic due to its great potential applications in image retrieval, image understanding, and image management [1]. In the past 20 years, some representative AIA approaches have been proposed and great achievements have been made, such as MBRM [2], JEC [3], and and 2PKNN [4]. Recently, deep learning models are extensively being used for various computer vision tasks and shown a breakthrough performance, which mainly contributes to end-to-end feature extraction through convolution neural networks. The deep learning-based AIA is a quite new but promising direction for AIA [5]. Although many successes have been achieved in annotation models and deep learning, results of existing image annotation methods is far from satisfaction. Different from the traditional visual feature-based image annotation method, we propose a novel image annotation method that first generates a textual statement describing the image content through the image description generation technology; then, a text feature vector describing the described image is constructed based on the text statement; finally, calculates the similarity from the text feature vectors between images to complete the image annotation by propagating the annotation model.

Experimental results show that our method outperforms most existing image annotation models, which mainly benefits from the proposed textual feature generated by image captioning technology. The textual tag feature inherently reduces the semantic gap between the high-level labels and low-level feature.

2. Related Work

2.1 Nearest neighbor model based approaches

Recently, many nearest-neighbor models based AIA methods have been proved to be quite successful [5]. The representative models based on nearest neighbors include JEC [3], TagProp [6], 2PKNN [4]. The Joint Equal Contribution (JEC) model is one of the most classical nearest-neighbor models [3]. The JEC model utilizes various low-level image features and a simple combination of basic distance measures to find nearest neighbors of a given image. It creates a family of very simple and intuitive baseline method for image annotation.

The two-pass KNN (2PKNN) model represents a classical solution to solve problems related to label-imbalance and weak-labeling [4]. It identifies all related semantic neighbors for each label by selecting K similar images in the vocabulary to boost rare labels. Due to its successfully solving class-imbalance problem, the 2PKNN makes great achievement in terms of popular evaluation metrics including per-label precision and recall and still is one of the most famous and influential image annotation approaches[5].

2.2 Deep learning based approaches

Recently, Convolutional Neural Networks (CNNs) have shown great performance in many computer vision tasks by extracting effective feature vectors from images [7][8][9][10][11]. Jia, the creator of the Caffe, proposes CNN+WARP model [10]. Different from the aforementioned annotation methods based on CNN model, the D2IA approach is based on generative adversarial network (GAN) model. The D2IA aims to create semantically relevant, yet distinct and diverse labels [12]. Deep learning models are extensively being used for various computer vision tasks and shown a breakthrough performance, which mainly contributes to end-to-end feature extraction through convolution neural networks, but the application of deep learning for AIA is still in its early stage [13]. The deep learning-based AIA is a quite new but promising direction for AIA [5].

3. The Proposed Method

3.1 Architecture

In order to further improve the semantic level, an image annotation method completely based entirely on text feature information is proposed, and the sentences generated by image description are used to improve the text semantics and automatically annotate the image as the text features of the image. The proposed method is mainly divided into three modules, namely, the image feature extraction module, the image description generation module, and the image annotation module. The main workflow is shown in Figure 1.

Unlike traditional convolutional neural network based models to extract it global visual features, this paper uses Faster R-CNN network [14] combined with ResNet101 network [15] to extract deep visual features of images, extracting one feature vector for each target detection box. In the image description generation module, a sentence of descriptive text sentence is generated using a mesh-connected transformer model for each image. Different from the traditional Transformer model, which loses some information in the input image features during information transmission, the mesh connection-based Transformer model adopted in this paper can fully retain the visual information of the image, and then generate a more complete image description statement. In the image annotation module, the text features are constructed according to the image description statements generated by the previous module, the text vector is used instead of the mainstream image visual features, and the propagation annotation method is used for image annotation, and finally propagate appropriate labels to the test image.

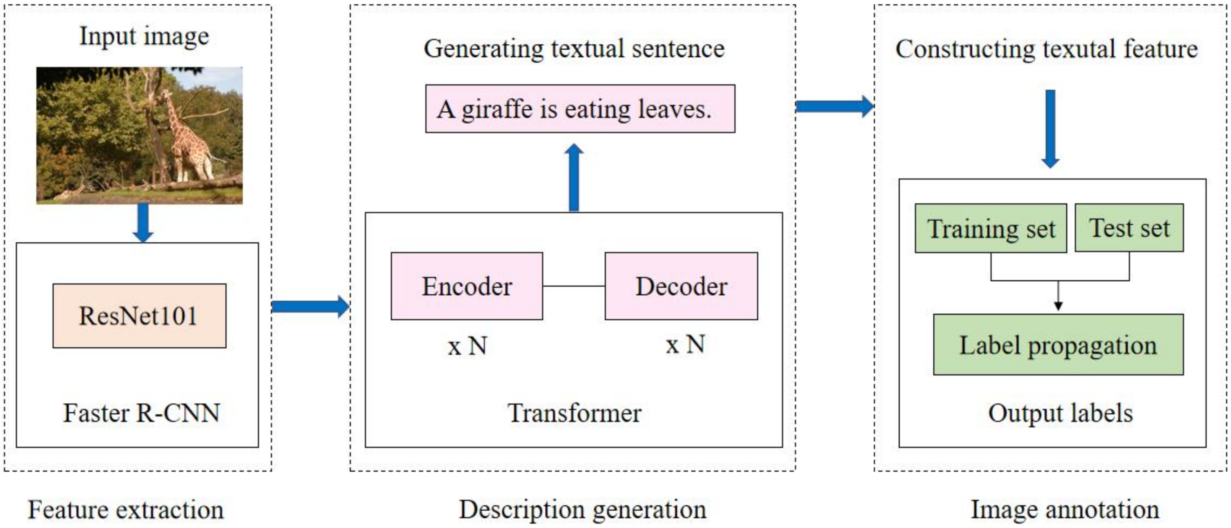

Fig. 1 Workflow Chart

3.2 Image description generation model network structure

Most of the existing models lack the full application of image data, and cannot parallelize the input features. In this experiment, the improved Transformer is used as a description generation model, taking advantage of the characteristics that it can process all input vectors in parallel to generate a smoother and more consistent statement description for human expression.In the image description generation method used in this chapter, the main process is to input the extracted image features into the mesh Transformer network model, use the unique self-attention mechanism and cross-attention method to directly encode and decoding the feature information of the image, and finally output the description text consistent with the image content.Image description Generation model specific network structure is shown in Figure 3.4.

After extracting the input image using Faster R-CNN, the feature vector is first position coded (Position Embedding, PE). This method corrects the problem that Transformer cannot obtain sequence sequence information, position coding adds position information to the input vector, and Transformer distinguishes the image characteristics of different locations through this information.The equation for position coding are shown in Equation (1) and (2).The p represents the bit of the current input vector.

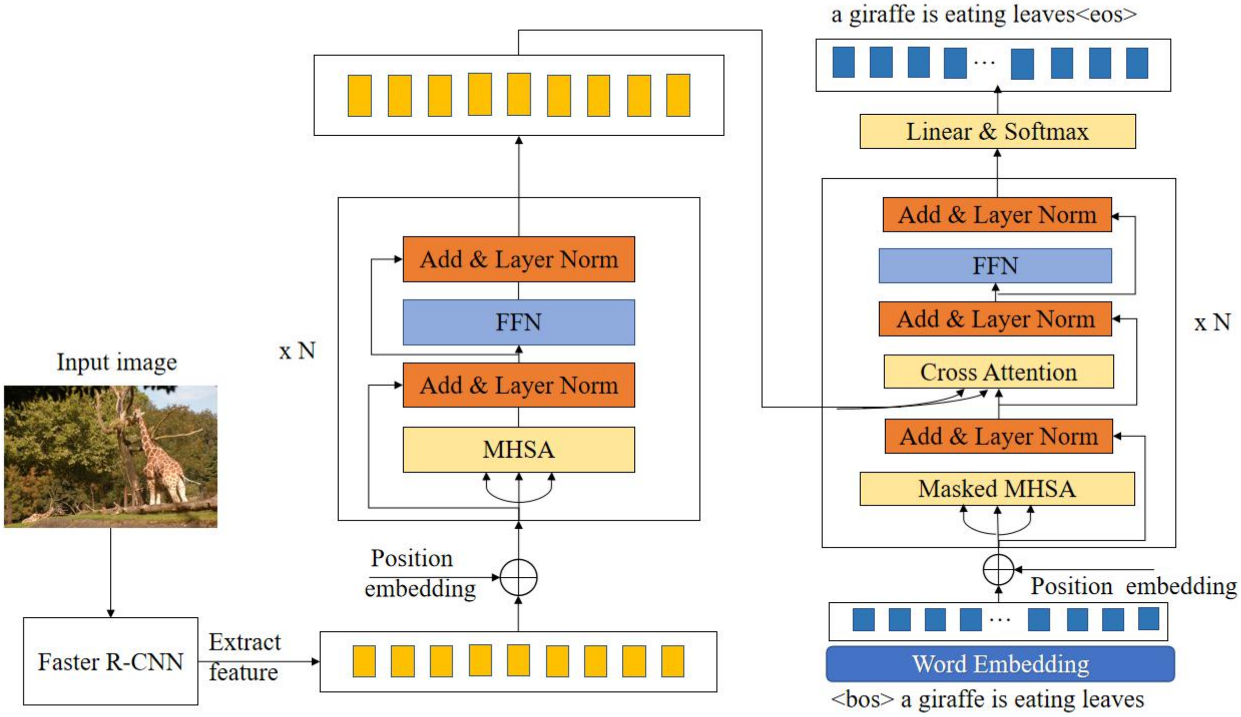

Fig. 2 Image caption generating model network structure

$$PE(p,2i)=sin\left(\frac{p}{1000^{\frac{2i}{d_m}}}\right) \quad (1)$$

$$PE(p,2i+1)=cos\left(\frac{p}{1000^{\frac{2i}{d_m}}}\right) \quad (2)$$

The core part of the network is the self-attention mechanism, which can directly calculate the correlation between the current vector and any other vectors. For each feature of the input vector, each feature multiple the input vector X by the matrix as, and obtains three different vectors, query (Q), key (K) and value (V). The equation for the concentration calculation is shown in Equation (3).

$$Attention(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

First, the transposition of matrix Q and K is then divided to get the attention score, representing the length of the vector in K. Because the vector is multiplied by too large, it is scaled to a certain extent and processed by softmax after all calculations are completed, which is equivalent to calculating the weight of each feature v.

Multi-Head Self-Attention (MHSA) uses the number of heads h to further get the q, k, v are divided into h parts, its calculation steps are to set h different self-attention, the input X into each self-attention, get the h weighted matrix, and then the h weighted matrix by columns, a linear transformation to get the final output result. The calculation equation of this procedure is shown in Equation (4).

$$\alpha_i=\sigma\left(W_i[Y_{t-1},C\left(X_i,Y_{t-1}\right)]+b_i\right) \quad (4)$$

$$h_i=Attention\left(QW_i^{q},KW_i^{k},VW_i^{v}\right) \quad (5)$$

In Equation (5), the corresponding Q, K, and V of each head are obtained by mapping through pairs.Specifically, for the learnable parameters W_i^q, W_i^k andW_i^v.

In the Add & Layer Norm layer, the input vector X is the same vector dimension calculated by attention, so it can be added directly, Add means a residual connection is introduced, Layer Norm sets the input of each network structure to the same mean and variance, both make the network training accelerate convergence.The calculation formulas such as (6) and (7).

$$LayerNorm(X + MHAtt(X)) \tag{6}$$

$$LayerNorm(X + FeedForward(X)) \tag{7}$$

The feature matrix output from the previous layer is then input into the feedforward neural network FFN part of the encoder, which consists of two activation function layers, first the ReLU activation function layer and second the linear activation function layer, using and representing the parameters of the first layer, using and representing the second layer parameters.The equation is shown in (8).

$$FFN(X) = max(0, XW_1 + b_1)W_2 + b_2 \tag{8}$$

In the decoder section, unlike the encoder, it has two multiple attention layers, the first is the mask operation in the self-attention part, the decoder can parallelize during training, the last decoder output words as input to predict the next decoder, pass the reference text and the corresponding output to the decoder, when predicting the i-output, should mask the word after i+1, which is used before the softmax of Equation (3). The second difference is that the decoder adds a layer of multiple attention, which captures the information of the encoder, where the Q matrix used by the decoder is calculated by the output of the previous decoder, while the K and V matrices come from the encoding matrix of the encoder.Finally, the decoder generates the description statement word-by-word via softmax.

3.3 Transformer Network Connection Mode

A six-layer Transformer model is adopted to connect each layer of encoder to each layer to realize the real cross-attention between the encoder and the decoder, so that the feature information of the image is more fully utilized.In the cross-attention section, the decoder retains the previous layer of q-information after self-attention calculation, receives the k and v information from the encoder after self-attention calculation, and combines the new q, k, and v vectors for cross-attention calculation.This mesh connection structure makes each decoder layer use the output of all the encoder layer, more making full use of the characteristics of the image, making the generated statement more fit to the image content.

In the model network connection structure, the cross-attention method of encoding decoding is different from the original network, which transmits the output X of each layer encoder and the word Y generated at the previous encoder to the decoder, and transmits the K and V to the second multi-head attention layer of the decoder, and the cross-attention F (X, Y) of the network connection is shown in Equation (9).

$$F(X, Y_{t-1}) = \sum_{i=1}^{N} \alpha_i . C(X_i, Y_{t-1}) \tag{9}$$

where X_i represents the ender layer encoded at the t time, Y is the word output at the previous time, representing the cross attention of the encoder to the decoder, the calculation Equation is shown in (10).

$$C(X, Y_{t-1}) = Attention(W^q Y_{t-1}, W^k X_i, W^v X_i) \tag{10}$$

where α_i represent weight matrix the same size as the cross-attention results was used, calculated to measure the degree of correlation between the cross-attention results for each coding layer and the query q entered by itself. The calculation Equation is shown in (11).

$$\alpha_i = \sigma(W_i[Y_{t-1}, C(X_i, Y_{t-1})] + b_i) \tag{11}$$

Where the weight matrix W_i, which represents the sigmoid activation function, is a learnable bias vector.

D. Propagation annotation model

Given the test image (I_i), we can obtain its k-nearest-neighb or images from the training dataset according to their similariti es. We can obtain its K neighbor training images in the textual label feature space for each current image I_i and rank the label s for image I_i according to their probability scores of:

$$P(I_i \mid w) = \sum_{j=1}^{K} sim_{text}(I_i, I_j) * P(w \mid I_j) \tag{12}$$

where, $sim_{test}(I_i, I_j)$ is the textual similarity between I_i and I_j. Based on probability theory, the probability of assigning a label w_k to I_i can be defined the posterior probability as follow:

$$P(w_k \mid I_i) = \frac{P(I_i \mid w_k)P(w_k)}{P(I_i)} \tag{13}$$

where, $P(w_k)$ is the probability of the label w_k. The best label for the test image I_i will be given by the following:

$$y^* = \arg\max_k P(w_k \mid I_i) \tag{14}$$

4. Experiment and Analysis

4.1 Dataset

We conduct experiments on the MS-COCO benchmark datasets.Both labels and social tags of images are textual English words. The Microsoft COCO (MS-COCO) dataset is used for image recognition, segmentation, and captioning. It contains 123 thousand images of 170,339 user -provided noisy tags and 80 expert-provided ground truth labels. Following previous works[16][17], we only keep 1,000 frequent tags and remove the images without any expert label, which leaves us with 123,286 images including 82,782 training images and 40,504 test images; each image being annotated with 2.9 labels on average. The refined MS-COCO dataset is the same as some research works [16][17].

4.2 Evaluation metrics

The precision, recall, and F1-measure are the most popular metrics in the information retrieval and AIA communities.Recently, per-label metrics (per-label precision, recall, and F1-measure) are most widely used in the AIA community.For each label , per-label precision (P_L) is defined as the number of images correctly predicted over the total number of images predicted with this label, and per-label recall (R_L) is defined as the number of images correctly predicted over the total number of images having this label in its ground-truth or manual annotations.The values are averaged over all the labels in the vocabulary as,

$$P_L = \frac{1}{C} \sum_{i=1}^{C} \frac{N_i^c}{N_i^p} \tag{15}$$

$$R_L = \frac{1}{C}\sum_{i=1}^{C}\frac{N_i^c}{N_i^g} \tag{16}$$

where C is the number of labels (keyword or classes) , N_i^c is the number of images correctly annotated for label i, N_i^p is the number of predictions for label i, and N_i^g is the number of images having ground-truth label i.

Many researchers have pointed out that the per-label metric s are biased toward infrequent labels, because making them co rrect could have a very significant impact on final accuracy [1 8].Therefore they propose per-image metrics [18][19][20][21]. The per-image precision (P_I) and recall (R_I) are defined as,

$$P_I = \frac{\sum_{i=1}^{C} N_i^c}{\sum_{i=1}^{C} N_i^p} \tag{17}$$

$$R_I = \frac{\sum_{i=1}^{C} N_i^c}{\sum_{i=1}^{C} N_i^g} \tag{18}$$

The F1-measure is used for comprehensive performance evaluation by combing precision and recall.The same equation can be used for both per-label metric and per-image metric as,

$$\text{F1-measure} = \frac{2\times \text{P}\times \text{R}}{\text{P}+\text{R}} \tag{19}$$

4.3 Experimental Settings and Details

This section continues with experiments using the MS-COCO dataset, summarized in sentence after generating text description statements for all images at the previous stage.In npy file make dictionary by description statement and saved in dictionary.In the npy file, using the dictionary to build the image, first remove the stop words in sentences, such as "a", "the", "and" that "," of "," on " words and the conjunctions used to generate the sentence, take the first 1,000 words with the highest frequency as the tag label text used in the image annotation, extract the tag tag included in each image as the tag feature of the image, and save it as feat_train.The m and the feat_test.For the m-file, enter the subsequent image annotation model.

In the image annotation stage, the annotation model parameters are set for the MS-COCO dataset, and the first N labels with the highest probability are taken as the annotation words of the image. Since the currently used image dataset contains 2.9 real annotation words per image, the number of output annotation words N is set to 3. Using the multi-level label propagation annotation method, the model is set to 30 classes, then determine the cluster category of the test image to 0.4, finally select 50 nearest neighbor images in the second and third steps, set the European distance when calculating the distance between the labeled image and the training set, then converting the distance into the image similarity, and finally passing three dimension words for the test image.

4.4 Experimental results

For a fair comparison, to compare with deep learning based image annotation models on large-scale datasets, we carry out our experiments on the same benchmark datasets (MS-COCO dataset) and predict a fixed length of annotations (three labels) for each test image.We compare our method and some representative methods using per-label metrics (precision, recall, F1-measure), per-image metrics. In addition, we use the hybrid F1-measure (called H-F1) combining per-label F1-measure and per-image F1-measure with the harmonic mean [1].

We compare our method with state-of-the -art models, including two nearest-neighbor models JEC and TagProp, and the state-of-the-art nearest neighbour based model 2PKNN , deep learning model CNN+WARP .The experiment results on MS-COCO is summarized in Table I. As can be seen from Table I, our method significantly outperforms the other methods (non-deep as well as deep learning based methods) on large-scale datasets in terms of most evaluation metrics, which mainly benefits from high -level semantic features and accurate neighbors.That might largely benefit from the textual feature , as it can capture high-level semantic concept .As far as the most important metric per-label F1 and comprehensive metric H-F1 are concerned, our method generally outperforms all models .

Table. I Performance evaluation on MS-COCO dataset

method	P_L	R_L	$F1_L$	P_I	R_I	$F1_I$	H-F1
JEC	49.12	41.35	44.90	49.39	61.10	54.63	49.29
TagProp	56.43	56.15	56.29	56.78	69.59	62.54	59.25
2PKNN	62.21	45.81	52.76	51.27	61.88	56.08	54.37
LC-KSVD	44.05	41.81	42.90	50.03	59.95	54.54	48.02
CNN+WARP	54.09	55.31	56.19	57.54	70.03	63.18	59.48
Ours	**68.59**	**58.49**	**63.14**	**60.44**	**73.00**	**66.13**	**64.60**

5. Conclusion and Future Outlook

We present a novel image annotation based on textual feature generated by image captioning technology. Our proposed method has several advantages. 1)To our knowledge, this is the first published work that use textual feature as only feature vectors instead of supplement feature. 2)Our proposed method can reduce the semantic gap and alleviate the issue of weak-labeling. 3) Our proposed method which can provide high level semantic in compared with traditional visual feature. Our method can find visually and semantically similar neighbor images, which can reduce the semantic gap and improve the performance. 4) In contrast to the traditional NN models paying more attention to frequent labels and classical, 2PKNN paying more attention to rare labels, our method can improve performance in both per-label and per-image metrics.

Extensive experiments demonstrate that our method can significantly outperforms competitive methods in terms of almost all evaluation metrics. Even though nearest neighbors based annotation models are concept-clear, structure-intuitive, and effective, there are several shortcomings. First, these methods will be time-consuming and space consuming if the number of the training image dataset is huge. Second, the performance of nearest neighbor model based AIA methods may be influenced by the size of training datasets.

In the future, we will explore Graph Neural Network into representation learning of multi-modal information. In addition, we are interested in exploring multi-modal feature space to further improve annotation performance.

References

[1] Y. Niu, Z. Lu, J. Wen, T. Xiang, and S. Chang, ‘‘Multi-modal multi-scale deep learning for large-scale image annotation,’’ IEEE Trans. Image Process., vol. 28, no. 4, pp. 1720–1731, Apr. 2019.

[2] S. L. Feng, R. Manmatha, and V. Lavrenko, ‘‘Multiple Bernoulli relevance models for image and video annotation,’’ in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2004, pp. 1002–1009.

[3] A. Makadia, V. Pavlovic, and S. Kumar, ‘‘Baselines for image annotation,’’ Int. J. Comput. Vis., vol. 90, no. 1, pp. 88–105, 2010.

[4] Y. Verma and C. V. Jawahar, “Image annotation by propagating labels from semantic neighbourhoods,’’ Int. J. Comput. Vis., vol. 121, no. 1, pp. 126–148, Jan. 2017.

[5] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, “A survey and analysis on automatic image annotation,’’ Pattern Recognit., vol. 79, pp. 242–259, Jul. 2018.

[6] Y. Zheng, T. Takiguchi, and Y. Ariki, “Image annotation with concept level feature using PLSA+CCA,’’ in Proc. 1 7th Int. Conf. Multimedia Modeling (Lecture Notes in Co mputer Science), vol. 6524, K. Lee, W. Tsai, H. M. Liao, T. Chen, J. Hsieh, and C. Tseng, Eds. Taipei, Taiwan: Spr inger, 2011, pp. 454–464, doi: 10.1007/978-3-642-17829-0_43.

[7] M. Koskela and J. Laaksonen, “Convolutional network features for scene recognition,’’ in Proc. 22nd ACM Int. Conf. Multimedia, K. A. Hua, Y. Rui, R. Steinmetz, A. Hanjalic, A. Natsev, and W. Zhu, Eds. Orlando, FL, USA: ACM, 2014, pp. 1169–1172.

[8] K. Simonyan and A. Zisserman, ‘‘Very deep convolutional networks for large-scale image recognition,’’ in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), Y. Bengio and Y. LeCun, Eds.San Diego, CA, USA: arXiv, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, ‘‘DeCAF: A deep convolutional activation feature for generic visual recognition,’’ in Proc. Int. Conf. Mach. Learn. (ICML), vol. 32, Jun. 2014, pp. 647–655.

[10] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, ‘‘Deep convolutional ranking for multilabel image annotation,’’ in 2nd Int. Conf. Learn. Represent. (ICLR), Y. Bengio and Y. LeCun, Eds. Banff, AB, Canada: arXiv, 2014. [Online]. Available: http://arxiv.org/abs/1312.4894

[11] V. N. Murthy, S. Maji, and R. Manmatha, ‘‘Automatic image annotation using deep learning representations,’’ in Proc. 5th ACM Int. Conf. Multimedia Retr., A. G. Hauptmann, C. Ngo, X. Xue, Y. Jiang, C. Snoek, and N. Vasconcelos, Eds. Shanghai, China: ACM, 2015, pp. 603–606.

[12] B. Wu, W. Chen, P. Sun, W. Liu, B. Ghanem, and S. Lyu, ‘‘Tagging like humans: Diverse and distinct image annotation,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018,pp.7967–7975.http://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Tagging_Like_Humans_CVPR_2018_paper.html

[13] P. K. Bhagat and P. Choudhary, ‘‘Image annotation: Then and now,’’ Image Vis. Comput., vol. 80, pp. 1–23, Dec. 2018.

[14] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6):1137–1149.

[15] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:770–778.

[16] A. Dutta, Y. Verma, and C. V. Jawahar, ‘‘Automatic image annotation: The quirks and what works,’’ Multimedia Tools Appl., vol. 77, no. 24, pp. 31991–32011, Dec. 2018.

[17] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, ‘‘Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 652–663, Apr. 2017.

[18] Q. Zhang and B. Li, ‘‘Discriminative K-SVD for dictionary learning in face recognition,’’ in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 2691–2698.

[19] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, ‘‘Semantic regularisation for recurrent image annotatio n,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4160–4168.

[20] J. Zhang, Q. Wu, J. Zhang, C. Shen, and J. Lu, ‘‘Mind your neighbours: Image annotation with metadata neighbourhood graph co-attention networks,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2956–2964.

[21] D. Tian and Z. Shi, ‘‘A two-stage hybrid probabilistic topic model for refining image annotation,’’ Int. J. Mach. Learn. Cybern., vol. 11, no. 2, pp. 417–431, Feb. 2020.

HAIYU YANG is currently pursuing the B.S. degree with the School of Computer Science and Engineering, Dalian Nationalities University, China. His current research interests include computer vision, image processing, and deep learning.

HAIYU SONG Received the B.S.,M.S.,and Ph.D. degrees in computer software and theory from Jilin University, in 1996, 2003, and 2012, respectively. He is currently a professor with the School of Computer Science and Engineering, Dalian Nationalities University, China. His current research interests include image understanding, computer vision, and machine learning.

WEI LI Received the B.S. and M.S. degrees in computer software and theory from Jilin University, in 2002, 2005, and the Ph.D. degree in traffic information and control from Jilin University, in 2020. She is currently a lecturer with the School of Computer Science and Engineering, Dalian Nationalities University, China. Her current research interests include image understanding, computer vision, and machine learning.

KEXIN QIN is currently pursuing the M.S. degree with the School of Computer Science and Engineering, Dalian Minzu University, China. Her current research interests include computer vision, image processing, and deep learning.

HAOYU SHI is currently pursuing the M.S. degree with the School of Computer Science and Engineering, Dalian Minzu University, China. His current research interests include computer vision, image processing, and deep learning.

QI JIAO is currently pursuing the M.S. degree with the School of Computer Science and Engineering, Dalian Minzu University, China. Her current research interests include computer vision, image processing, and deep learning.

Sources of funding for research presented in a scientific article or scientific article itself

This work was supported in part by the Liaoning Natural Science Foundation Projects under Grant 2019-ZD-0182, in part by the Scientific research project of Liaoning Provincial Education Department under Grant LJKZ0024, in part by the National Social Science Foundation Project under Grant 16BMZ010, in part by the Creative Research Talents from Universities of Liaoning Province under Grant LR2016071, and in part by the National Statistical Science Research Project under Grant 2019LY12.

Creative Commons Attribution License 4.0 (Attribution 4.0 International , CC BY 4.0)