Users Personalized Sketch-Based Image Retrieval Using Deep Transfer Learning

Qiming Huo(✉), Jingyu Wang, Qi Qi, Haifeng Sun, Ce Ge, and Yu Zhao

State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing 100876, China
314762927@qq.com, {wangjingyu,qiqi,sunhaifeng1,gece,zhaoyu}@ebupt.com

Abstract. Traditionally, sketch-based image retrieval is mostly based on human-defined features for similarity calculation and matching. The retrieval results are generally similar in contour and lack complete semantic information of the image. Simultaneously, due to the inherent ambiguity of hand-drawn images, there is "one-to-many" category mapping relationship between hand-drawn and natural images. To accurately improve the fine-grained retrieval results, we first train a SBIR general model. Based on the two-branch full-shared parameters architecture, we innovatively propose a deep full convolutional neural network structure model, which obtains mean average precision (MAP) 0.64 on the Flickr15K dataset. On the basis of the general model, we combine the user history feedback image with the input hand-drawn image as input, and use the transfer learning idea to finetune the distribution of features in vector space so that the neural network can achieve fine-grained image feature learning. This is the first time that we propose to solve the problem of personalization in the field of sketch retrieval by the idea of transfer learning. After the model migration, we can achieve fine-grained image feature learning to meet the personalized needs of the user's sketches.

Keywords: Personalized sketch-based image retrieval
Deep full convolutional neural network · Transfer learning
Feature extraction

1 Introduction

The difference in the distribution of sketch and natural image statistics results in completely different image scopes. Using the computer to process hand-drawn image feature and to retrieve related natural pictures require some transformation [8] hand-drawn and natural images to make them in the same image domain. In addition, determining user intent from visual search queries is still a public challenge, especially in sketch-based image retrieval (SBIR). The hand-drawn images are ambiguous, and the same hand-drawing can express the semantics of different things. On the other hand, hand-drawings of the same object, drawn by different users, are also different, and eventually the search results after computer

W. Liu et al. (Eds.): KSEM 2018, LNAI 11061, pp. 160–168, 2018.
https://doi.org/10.1007/978-3-319-99365-2_14

operations are certainly different. Usually, there is a "one-to-many" relationship between the hand-drawing and categories of natural images. If it is desired that the computer can accurately recognize objects represented by the user-drawn image as humans do, it is necessary to add user relevance feedback information. The feedback allows the user to indicate to the system which of these instances are desired or related, and which are not. Based on feedback, the system modifies its search mechanism and tries to return a more optimal picture set to the user [9]. The feedback here serves as an effective tool for extracting image depth semantic information and doing fine grain image analysis.

The main contributions of this paper are as follows. (1) Based on the natural image cross-image scoping method, extracting the bottom pixel-level edge line information of the natural image, which is input to the improved deep full-convolution neural network simultaneously with the hand-drawn image information. After the training, the mean average precision (MAP) of the model evaluation is greatly improved compared with the traditional image algorithm [5–7,14] and the deep learning algorithm [1,2,12,13] in recent years; (2) As for the problem of "one-to-many" relationships between hand-drawing and the categories of natural images, we propose a data modeling method based on user feedback using the idea of transfer learning [10]. The way is using the user history feedback to adjust the spatial distribution of the subclass images and input hand-drawn image feature vectors on the basis of a general model. Determine the subspace [17] where the fine-grained natural image and the input hand-drawn image are located in the overall feature space of each category. The migrated model completes the fine-grained image retrieval task and satisfies the user's personalized requirements to the maximum extent possible.

2 Sketch-Based Image Retrieval General Model

As for the general model of sketch retrieval, our goal is to extract the complete image feature information as much as possible. The more complete the sketch feature is, the more the real content of the sketch can be expressed, and the more accurate the sketch matching is. At the same time, this step is also the basis for the training of personalization model training data. The quality of the common model directly affects the training of the feedback process.

2.1 Image Pre-processing and Feature Extraction

In the natural image contour extraction process, we use the global probability of boundary (gPb) edge detection algorithm and the dual threshold processing method to obtain the binary edge map, retaining the strongest edge information of 25% and removing 25% of the weakest edge information. Then, the canny edge extraction algorithm is used to perform the lag threshold processing. And the pixels connected to the strong edges are left and the isolated edge pixels are removed [2]. The resulting image after filling the remaining blank image to a size of 256×256 is then binary processed. The nature images are filled the blank and binary processed by the same way.

2.2 Establish a Joint Loss Function

The construction of label information is based on the relationship between the hand-drawn image X^S and the contour image X^C provided in the data set. First define the input tag Y, which value is 0 or 1. When the i-th hand-drawn image X_i^S and the contour image X_i^C are in the same category, it is a positive sample, and the triplet $<X_i^S, X_i^C, Y_i = 0>$ is constructed. Conversely, it is a negative sample, construct a triplet $<X_i^S, X_i^C, Y_i = 1>$. When the triplet is input into the neural network, X_i^S and X_i^C are used to calculate V_i^S and V_i^C, where $V_i^S = f_N(X_i^S)$ and $V_i^C = f_N(X_i^C)$, f_N is the neural network forward propagation calculation function. Here is the loss function [4]:

$$\sum_{i=0}^{batch\_size} (1 - Y_i)\frac{2}{Q}Ew_i^2 + Y_i \times 2Q \times e^{-\frac{2.77}{Q}Ew_i} . \tag{1}$$

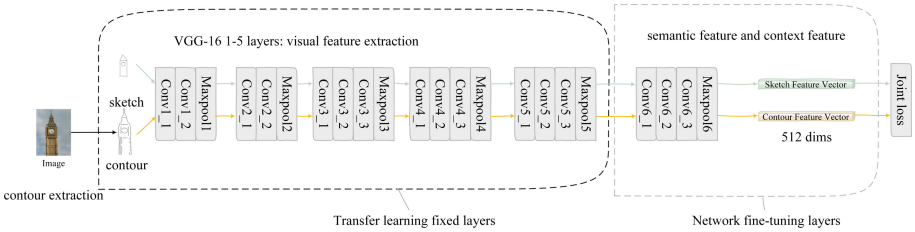


Fig. 1. Two-branch CNN structure and migration learning model outline diagram

Q is a constant, which is the maximum value of $Ew = ||V^S - V^C||_2$ when the final category is discriminated (Fig. 1).

2.3 Image Similarity Matching and Retrieval

In image retrieval, we calculate the Euclidean distance of the feature vectors of all the pictures in the hand-drawn image and the image library is calculated by traversing the entire list.

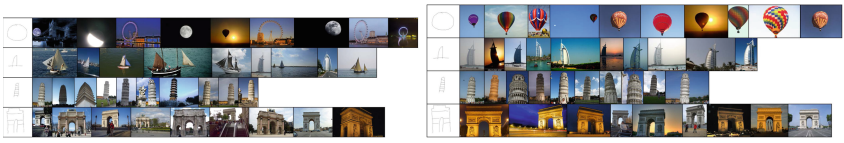
$$Sim_{common} = [euc_1, euc_2, euc_3 ..... euc_n]. \tag{2}$$

Take the index of the top K values as the most similar K pictures as candidate results to return to the user interface. For the R is the set of nature images.

$$index = \arg\min_{i \in R}^{K} Euc(i). \tag{3}$$

3 User Feedback Based on Transfer Learning

Through the discussion in the previous section, we can obtain a general model after the training process has converged. When the general model gives candidate results based on the low-level similarity matching, the user is provided with the choice of which pictures are of interest to the user. The system selects the records according to the user history, and judges the positive correlation sample, negative correlation sample, and general correlation sample [16]. Use this data as a training sample for a personalized model. By learning user feedback information, the model will change the distribution of input hand-drawn images and variously related samples in the feature space. Then, based on the rearranged distribution, the system determines the similarity measurement method and retrieves the image with the closest similarity to the user (Fig. 2).



(a) General model (b) Personalization model

Fig. 2. Different retrieval results of general model and personalization model

3.1 Data Construction

As for the construction of the training sample data set, here we define R is a correlation set of user set U, input hand-drawing set S, user feedback natural image outline image set FI, and construction training data set sampling natural image outline SI set, expressed as $R_{u,s,Fi,Si}$. The set FC represents the parent category of the Fi in the feedback data pair $<s, Fi>$, and the set SC represents the sub-category where Fi is located. Defining the input tag $Y \in \{0, 0.5, 1\}$ is used to train the input data of the neural network model. The actual meaning represented is the quantification value of the correlation degree. The rules defining the relationship between samples are defined as follows, in which samples $Si \in SI$. The tag Y rules are defined as follows:

$$Y = \begin{cases} 0, & \text{if } Si \in FC \quad and \quad Si \in SC \\ 0.5, & \text{if } Si \in FC \quad and \quad Si \notin SC \\ 1, & \text{if } Si \notin FC \quad and \quad Si \notin SC \end{cases}. \tag{4}$$

According to the user feedback result data, the training data is constructed with the positive correlation, negative correlation and general correlation sample 1:1:1 ratio when training samples are selected, and the input data format is a quadruple $< u, s, Si, Y >$.

3.2 Network Architecture

In this task, since a general correlation is added in this section, the two-branch independent joint loss function (1) based on strong and weak relations in the previous section needs to be rewritten as a three-branch independent function.

$$L_p(Ew, Y) = \delta_1 F_S(Ew) + \delta_2 F_M(Ew) + \delta_3 F_W(Ew). \tag{5}$$

where $Ew = ||V^S - V^C||_2$ is still the Euclidean Distance between the two output vectors, $F_S(Ew)$, $F_M(Ew)$, $F_W(Ew)$ are the loss functions selected for the positive correlation sample, the general correlation sample, and the negative correlation sample relationship respectively. The prefix term δ is defined here as an independent factor, which is determined according to the value of Y. In order to ensure the independence of branches, set: $\delta_1 = 2 \times |Y - 1| \times (0.5 - Y)$, $\delta_2 = 4 \times |Y - 1| \times Y$, $\delta_3 = 2 \times |Y - 0.5| \times Y$.

As for the loss function of each independent branch, to ensure that the overall feature space does not change, continue to use $F_W(Ew) = 2Q \times e^{-\frac{2.77}{Q}Ew}$ as the branch loss function. For positively correlated samples and generally related samples, since the two parts of loss function are to reduce the distance of output vectors, but magnitude of reduction should different from each other, we introduce a double-threshold method here to control the amplitude. Define: $F_S(Ew) = \frac{2}{Q}(Ew - Th1)^2$ and $F_M(Ew) = \frac{2}{Q}(Ew - Th2)^2$. $Th1$ and $Th2$ are set as thresholds, where $Th1 < Th2 < Q$.

3.3 Similarity Measure and Image Retrieval

The similarity metrics selected during the final test still use the Euclidean distance of the output eigenvector after using the personalized model, so that get a list of similarity results for personalized models:

$$Sim_{personal} = [euc_1, euc_2, euc_3 ..... euc_n]. \tag{6}$$

Let $w \in [0, 1]$ be the weighting factor, the final similarity list is:

$$Similarity = w \times Sim_{personal} + (1 - w) \times Sim_{common}. \tag{7}$$

Arranged from the smallest to the largest, the indexes of the top K values are the suitable pictures for the user's preference as the final fine-grained search results and returned to the user interface.

4 Experiments and Results

4.1 Dataset and Experiment Settings

The sketch-based image retrieval general model training experiment is based on the public data set Flickr15K. In order to effectively extend the training data and solve the model over-fitting problem during each batch of input data, we

also set up a random hand-drawn image/contour image cropping and flipping to perform data enhancement operations. We use the RMSProp algorithm to train the network for a total of 20 epochs on the Tensorflow platform. For the model to converge to the optimal result, we set the learning rate decay operation, the initial learning rate is set to 0.0001, and the learning rate decay is performed every 5 epochs, and the degree of decay is 0.5. Select 100 between the boundary Euclidean distance Q for the positive and negative sample pairs in the general model (Tables 1 and 2).

As for the personalized model, the feedback images randomly select the natural images in TOP-20 in the general model to simulate the single user selection operation. The sub-category according to the folder name can be determined to have a total of 60. Each experimental set of hand-drawn sketches randomly corresponds to 100 positive correlation samples (positive correlation samples can be repeated). The experiment has also the learning rate decay operation. Using the Adam optimization algorithm to train the network for a total of 20 epochs. Threshold value $Th1 = \frac{Q}{10}, Th2 = \frac{Q}{2}$ is set in the personalized model.

4.2 Model Evaluation

For the general model and personalized model, we can both evaluate the intuitive perception and quantitative values. Intuitively, we can observe the TOP-N results of the search to visually feel the matching of the hand-drawn images with the resulting images. Measure the index of general image retrieval model can use the mean average accuracy (MAP) to determine the precision by category. This indicator can not only show the effect of retrieval on each specific category, but also play a very important role in the generalization of the entire search model (Fig. 3).

Table 1. Neural network structure

Layers	Kernal	Strides	Padding	Filters	Output size
Conv1_1	3 × 3	1	1	64	256 × 256 × 64
Conv1_2	3 × 3	1	1	64	256 × 256 × 64
MaxPool1	2 × 2	2	0		128 × 128 × 64
Conv2_1	3 × 3	1	1	128	128 × 128 × 128
Conv2_2	3 × 3	1	1	128	128 × 128 × 128
MaxPool2	2 × 2	2	0		64 × 64 × 128
Conv3_1	3 × 3	1	1	256	64 × 64 × 256
Conv3_2	3 × 3	1	1	256	64 × 64 × 256
Conv3_3	3 × 3	1	1	256	64 × 64 × 256
MaxPool3	2 × 2	2	0		32 × 32 × 256
Conv4_1	3 × 3	1	1	512	32 × 32 × 512
Conv4_2	3 × 3	1	1	512	32 × 32 × 512
Conv4_3	3 × 3	1	1	512	32 × 32 × 512
MaxPool4	2 × 2	2	0		16 × 16 × 512
Conv5_1	3 × 3	1	1	512	16 × 16 × 512
Conv5_2	3 × 3	1	1	512	16 × 16 × 512
Conv5_3	3 × 3	1	1	512	16 × 16 × 512
MaxPool5	2 × 2	2	0		8 × 8 × 512
Conv6_1	1 × 1	1	0	128	8 × 8 × 128
Conv6_2	3 × 3	1	1	128	8 × 8 × 128
Conv6_3	1 × 1	1	1	512	8 × 8 × 512
MaxPool6	8 × 8	8	0		1 × 1 × 512

Table 2. MAP in general model

Methods	MAP
Ours	**0.6449**
AFM + QE [15]	0.579
Triplet(fine-tuned final model) [3]	0.3617
Sketchy triplet [13]	0.3591
Query-adaptive re-ranking CNN [1]	0.3230
Triplet loss CNN [2]	0.2445
Siamese CNN [12]	0.1954
PeceptualEdge [11]	0.1837
GF-HOG [6]	0.1222
HOG [5]	0.1093
SIFT [7]	0.0911
SSIM [14]	0.0957

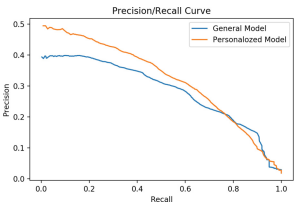


Fig. 3. P-R curves of general and personalized models tested on the entire data set.

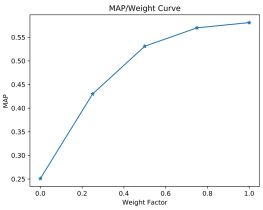


Fig. 4. Personalization Retrieval Model MAP and Weight Factor line chart.

For personalized training model evaluation, in order to make the personalized model contain both sketch content outline information and user feedback preference information, the effect of w is added to the evaluation during the calculation. Select 0, 0.25.0.5, 0.75 and 1 representative values from $[0, 1]$ as the value of w, and thus the MAP change curve is made according to the value of w as shown in the Fig. 4.

Table 3. Personalization models and general model through fine-grained search subcategories AP and MAP in Flickr15K comparison tables

Category	1	2	3	5	7	11	12	15	18	30	MAP
General model	0.039	0.012	0.104	0.032	0.44	0.218	0.189	0.784	0.602	0.087	0.2507
Personal model	0.645	0.57	0.417	0.335	0.677	0.393	0.337	0.994	0.654	0.993	0.6015

In the Table 3, it can be seen that while the general model has a high degree of accuracy in retrieving the parent category, it does not achieve good results in the retrieval of natural images in the fine-grained subcategories. The reason is that the sub-category sample features are randomly distributed in the parent category's feature space, so it is probably the correct sample from the same parent category appears, but the sub-category appears randomly. The improved personalized model successfully subdivides the spatial range of the sub-category features. When the input hand-drawn images are calculated to obtain features, according to the principle of distance similarity, the calculated features are located as close as possible to the sub-category samples required by the user, so that sub-category natural images can be efficiently retrieved.

5 Conclusions

In this article, we show how to use an improved depth full convolutional neural network to extract sketch features. We use popular universal datasets to verify the powerful feature extraction capabilities of our designed network. The high-level feature information learned by neural networks is used to improve the

accuracy of sketch retrieval. Based on the idea of transfer learning, the distribution of the migrated feature space is adjusted, and the similarity between the results of user input and historical feedback is further enhanced. The improved personalized model can not only retrieve pictures based on image content but also retrieve pictures based on user selection feedback. However, in the experiment, we used user feedback to enhance supervised sub-tag information to achieve fine-grained sketch retrieval. The accuracy of tagging and user selection operations are highly dependent on the tags. Error tagging or incomplete information of tag information and user's random selection of feedback activities the large probability affects the final retrieval rate. Therefore, in the future, realizing fine-grained sketch retrieval based on weak supervision information needs to design more powerful neural network networks and scientific loss function models.

Acknowledgment. This work was jointly supported by: (1) National Natural Science Foundation of China (No. 61771068, 61671079, 61471063, 61372120, 61421061); (2) Beijing Municipal Natural Science Foundation (No. 4182041, 4152039); (3) the National Basic Research Program of China (No. 2013CB329102).

References

1. Bhattacharjee, S.D., Yuan, J., Hong, W., Ruan, X.: Query adaptive instance search using object sketches. In: Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, pp. 1306–1315 (2016)
2. Bui, T., Ribeiro, L., Ponti, M., Collomosse, J.: Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network. Comput. Vis. Image Underst. **164**, 27–37 (2017)
3. Bui, T., Ribeiro, L.S.F., Ponti, M., Collomosse, J.P.: Generalisation and sharing in triplet convnets for sketch based visual search. CoRR abs/1611.05301 (2016)
4. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR 2005, pp. 539–546 (2005)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR 2005, pp. 886–893 (2005)
6. Hu, R., Collomosse, J.P.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval. Comput. Vis. Image Underst. **117**(7), 790–806 (2013)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
8. Ma, Z., Tan, Z., Guo, J.: Feature selection for neutral vector in EEG signal classification. Neurocomputing **174**(174), 937–945 (2016)
9. Macarthur, S.D., Brodley, C.E., Kak, A.C., Broderick, L.S.: Interactive content-based image retrieval using relevance feedback. Comput. Vis. Image Underst. **88**(2), 55–75 (2002)
10. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
11. Qi, Y., et al: Making better use of edges via perceptual grouping. In: CVPR 2015, pp. 1856–1865 (2015)
12. Qi, Y., Song, Y., Zhang, H., Liu, J.: Sketch-based image retrieval via siamese convolutional neural network. In: ICIP 2016, pp. 2460–2464 (2016)

13. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. ACM Trans. Graph. **35**(4), 119 (2016)
14. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: CVPR 2007 (2007)
15. Tolias, G., Chum, O.: Asymmetric feature maps with application to sketch based retrieval. In: CVPR 2017, pp. 6185–6193 (2017)
16. Xie, L., Wang, J., Zhang, B., Tian, Q.: Fine-grained image search. IEEE Trans. Multimed. **17**(5), 636–647 (2015)
17. Xu, P., et al.: Cross-modal subspace learning for fine-grained sketch-based image retrieval. Neurocomputing **278**, 75–86 (2018)