Title for citation

SCIENCE CHINA
Information Sciences

• Supplementary File •

Multi-attention based Cross-domain Beauty Product Image Retrieval

Zhihui Wang[1], Xing Liu[1], Jiawen Lin[1], Caifei Yang[1] & Haojie Li[1*]

[1]*International School of Information Science and Engineering, Dalian University of Technology, Dalian* 116086,*China*

Appendix A Experiments

Appendix A.1 Dataset

The experiments are conducted on Perfect-500K, a large-scale image dataset of beauty products, released by 2018 ACM Multimedia Grand Challenge, which has 520,727 images from major e-commerce sites, e.g., Amazon, Cult Beauty, Yahoo Shopping Mall, etc. After cleaning up the invalid images, there are still about 480K images left. Parts of images in the Perfect-500K dataset are shown in Figure A1. All of the images are from e-commerce sites, with no real-world examples provided in advance. Each image has a text description information as its label and a unique id. For validation images, it consists of mapping of validation image ID to the respective training image ID, which is used to evaluate our approach.

Since the Perfect-500K dataset is a non-classified dataset provided by the Perfect Half Million Beauty Product Image Recognition Challenge, annotation information of all images is their text description. It is difficult to build an effective basic network using text descriptions. In addition, because the ImageNet dataset only contains very rare beauty product categories, the pre-trained network on Imagenet can not accurately and robustly describe the product objects. In order to learn an effective and generalized feature representation for images of beauty products, it is necessary to establish a dataset with clear label information. However, it is impossible to classify those images in Perfect-500K one by one.

In this paper, we use TF-IDF algorithm to analyze word frequency statistics of all text descriptions in the annotation information of Perfect-500K dataset and manually count 44 rough categories. Sequentially, we extract approximately 35,000 images from Perfect-500K dataset associated with these 44 categories based on the category keywords and construct a "few-shot" dataset named Perfect-30K containing 44 categories. These 44 categories include lipstick, facial cleanser, sunscreen, eye shadow, razor, mask, etc. Each of them contains about 800 images. Some categories of the Perfect-30K dataset are shown in Figure A2.

Appendix A.2 Performance Metrics

To evaluate the effectiveness of the algorithm and ensure fairness with other algorithms, we evaluated our algorithm using Mean Average Precision@7 (MAP@7). MAP@7 averages the average accuracy score of the top 7 rankings for all query images. Note that larger MAP@7 value indicates better performance. The calculation formula is as follows.

$$AP(q) = \frac{\sum_{k=1}^{7}(p(k) \times rel(k))}{Nrel}, \tag{A1}$$

$$MAP@7 = \frac{\sum_{q=1}^{Q} AP(q)}{Q}, \tag{A2}$$

where q represents a query image, Q represents the number of query images, $p(k)$ represents the accuracy of the top k retrieval results, $rel(k)$ is an indicator function equals to 1 if the item of rank k is a relevant image with query image. $Nrel$ represents the total number of relevant images with query image.

* Corresponding author (email: hjli@dlut.edu.cn)

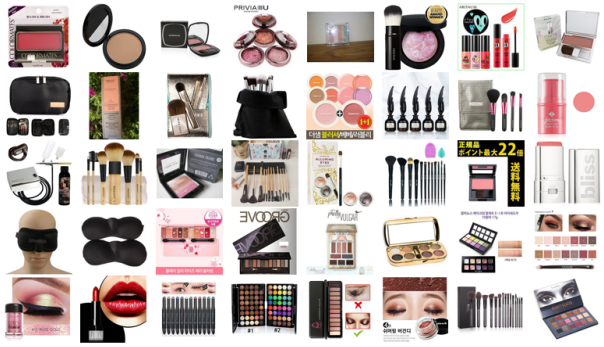

Figure A1 Examples of the Perfect-500K dataset.

Figure A2 Examples of some categories of the Perfect-30K dataset. The categories from top to bottom are dryer, eye mask, lipstick, moisturizer, and spray.

Appendix A.3 Multi-attention Classification Network Implementation Details

The proposed classification network is implemented using PYTorch framework. We divide the Perfect-30K dataset into a train dataset and a validation dataset according to the ratio of 8:2, and calculate the mean and variance of the three channels of the RGB image of the Perfect-30K dataset [0.8276, 0.8057, 0.7937] and [0.2334, 0.2479, 0.2563], respectively.

In the training phase, we first train the significant branch network separately. After the convergence of the saliency branch network, the backbone network and the text attention branch network are combined for end-to-end training. This training method can solve the problem of non-convergence of the network. In addition, we normalize all images in the dataset using the mean and the variance and resize them to 224*224, and use parameters of the ResNet50 pre-trained on ImageNet dataset to initialize the backbone network of the classified network. We set the momentum parameter to 0.90, the learning rate to 0.01, the weight decay to 0.0005 and the max iteration to 500. As the number of iterations increases, the learning rate will slow down to 0.00001.

Appendix A.4 Ablation Experiment of MANet

Two attention mechanisms, saliency attention and text attention, are proposed in our MANet. In this section, we use the Perfect-30K dataset to evaluate the impact of different attention mechanisms on classification accuracy. In addition, we also evaluate the impact of the end-to-end MANet and MANet using the DSS network as the saliency attention branch on classification accuracy. Table A1 gives results of the ablation experiment of MANet.

From Table A1, we observe that the saliency attention mechanism is superior to text attention mechanism. The main reason is that the product object region can provide more discriminative information for product classification than the text region. When using the saliency and text attention fusion mechanism, the classification accuracy of the beauty product image is always the highest regardless of the what network architectures selected. It shows that these two attention mechanisms are mutually reinforcing and promote the classification of beauty product images. Unsurprisingly, the classification network using only text attention mechanism has also achieved competitive accuracy. As we found when we summarized the characteristics of datasets, the text regions do provide discriminative information for product classification. In addition, we can also know that the performance of the MANet through end-to-end training is better than that of MANet using the DSS network as the saliency attention branch. Compared with the weighting method using saliency mask generated by DSS network, end-to-end training through joint optimization enables the network to self-adaptively learn a saliency attention weight coefficient map which is more suitable for the backbone network.

Appendix A.5 Comparison of Local Feature Aggregation Methods

In this paper, we propose a saliency-based regional maximum activation of convolutions(SR-MAC) to extract fine-grained features of beauty product images. In order to evaluate the effectiveness of our proposed local feature aggregation method and ensure the fairness of evaluation, we use VGG-16 pre-trained on ImageNet as feature extraction network, which is also adopted by RA-MAC method. Then the feature responses of VGG-16 network's last pooling layer are extracted, and we use different local feature aggregation methods to form their respective feature vectors. We evaluate different feature aggregation methods on Perfect-500K dataset, such as global max pooling (MAC), average pooling (SPoC), regional maximum activation of convolutions (R-MAC), regional maximum activations of convolutions with attention(RA-MAC), and our proposed saliency-based regional maximum activation of convolutions (SR-MAC). Table 2 gives the retrieval accuracy of different local feature aggregation methods in term of MAP@7.

The limitation of global max-pooling features (MAC) or average-pooling features (SPoC) is that such approaches are not compatible with the geometric-aware models involved in the final re-ranking stages. Moreover, global image feature can describe the whole image only and lack the ability to describe image details unable to generate discriminative features. As

Table A1 The effect of different attention mechanisms on classification accuracy. MANet with DSS represents the MANet using the DSS network as the saliency attention branch.

Network	Attention mechanism	Classification accuracy
Resnet50	n/a	65.17%
MANet with DSS	saliency attention	84.84%
	text attention	83.61%
	saliency and text attention	85.62%
End-to-end MANet	saliency attention	85.78%
	text attention	83.61%
	saliency and text attention	**87.26%**

Table A2 The retrieval accuracy of different local feature aggregation methods. $*$ numbers are reported in Lin et al. [1].

Network	Local feature aggregation	MAP@7
VGG-16	MAC [2]	0.195*
	SPoC [3]	0.175*
	R-MAC [4]	0.212*
	RA-MAC [1]	0.226*
	SR-MAC	**0.247**

to the RA-MAC method, it is unstable to use an unsupervised manner to gain attention region on the pre-trained network. From Table A2, we note that our proposed SR-MAC achieves the best performance, which indicates the effectiveness of SR-MAC. SR-MAC uses the salient region provided by the saliency branch in MANet independent of the feature extraction network to obtain the objects' locations, which ensures the accuracy of the feature extraction region.

Appendix A.6 Comparison with the State-of-the-art Methods

This section compares our method with other state-of-the-art methods. Table A3 reports the retrieval accuracy of different algorithms on Perfect-500K. For the fairness of comparison, we do not do any post-processing such as query expansion and data augmentation. In order to prove the main performance of MANet shown in Table A3 comes from its combination with saliency attention mechanism and text attention mechanism, one may seek Table A1 as reference. Using the same dataset (Perfect-30K), MANet shows higher classification accuracy than ResNet50. As can be seen from Table A3 , our method achieves the best result on Perfect-500K, which verifies the effectiveness of our method.

Our advantages are attributed to three points: 1) we use saliency and text attention mechanism to make the classification network pay more attention to the product objects and the text regions in the images; 2) we propose a robust local feature aggregation method, which eliminates the interference of background information and avoids to lose the key local areas in the product object region by using saliency mechanism; 3) we construct a well-labeled beauty product image dataset. Using this dataset, we can train the network to learn more accurate feature description for beauty products. We give three examples of our retrieval results as shown in Figure A3. From this figure, we can see the effectiveness of our approach intuitively. In particular, the false positive samples are quite similar to the query images, which reflects the difficulty of the beauty product image retrieval task.

Table A3 Comparison with the state-of-the-art methods. Ψ means that results are provided by authors. $ResNet-50^{\Phi}$ represents the result of using Perfect-30K dataset to train Resnet50 network and extract features of the last pooling layer for retrieval. MANet with DSS represents extracting the last pooling layer features of MANet using DSS nerwork as the saliency attention branch for retrieval. MANet represents extracting the last pooling layer features of the end-to-end MANet for retrieval.

Method	MAP@7
RA-MAC [1]	0.348^{Ψ}
MFF [5]	0.360^{Ψ}
Pre-trained ResNet50 [6]	0.207^{Ψ}
$ResNet-50^{\Phi}$	0.331
MANet with DSS	0.365
MANet	0.374
MANet with DSS+SR-MAC	0.389
MANet+SR-MAC	**0.395**

Figure A3 Examples of results achieved by our method. The blue box indicates the query images, the red boxes indicate the true positive samples.

References

1 Lin, Z., Yang, Z., Huang, F., and Chen, J. (2018) Regional maximum activations of convolutions with attention for cross-domain beauty and personal care product retrieval. *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 2073–2077. ACM.

2 Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014) CNN features off-the-shelf: An astounding baseline for recognition. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 512–519.

3 Babenko, A. and Lempitsky, V. S. (2015) Aggregating deep convolutional features for image retrieval. *CoRR*, **abs/1510.07493**.

4 Tolias, G., Sicre, R., and Jégou, H. (2016) Particular object retrieval with integral max-pooling of CNN activations. *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

5 Wang, Q., Lai, J., Xu, K., Liu, W., and Lei, L. (2018) Beauty product image retrieval based on multi-feature fusion and feature aggregation. *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 2063–2067. ACM.

6 Lim, J. H., Japar, N., Ng, C. C., and Chan, C. S. (2018) Unprecedented usage of pre-trained cnns on beauty product. *2018 ACM Multimedia Conference on Multimedia Conference*, pp. 2068–2072. ACM.