SCIENCE CHINA
Information Sciences

February 2020, Vol. 63 120112:1–120112:3
https://doi.org/10.1007/s11432-019-2721-0

• LETTER •
Special Focus on Deep Learning for Computer Vision

Multi-attention based cross-domain beauty product image retrieval

Zhihui WANG, Xing LIU, Jiawen LIN, Caifei YANG & Haojie LI*

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

Received 28 July 2019/Revised 9 October 2019/Accepted 12 November 2019/Published online 14 January 2020

Citation Wang Z H, Liu X, Lin J W, et al. Multi-attention based cross-domain beauty product image retrieval. Sci China Inf Sci, 2020, 63(2): 120112, https://doi.org/10.1007/s11432-019-2721-0

Dear editor,

In recent years, the Perfect Half Million Beauty Product Image Recognition Challenge has been held by ACM MultiMedia 2018 [1] for beauty product image retrieval task, and the Perfect-500K dataset has been released, acting as a large-scale beauty product dataset. Retrieval methods exploit classic CNN models to extract features and conduct either fusion or post-process to enhance accuracy of feature description (e.g., [2–4]). Others are inclined to design network architecture to achieve the idential effect (e.g., [5–7]). By observing the dataset, we argue that the beauty product objects are conspicuous and the text regions of images display noticeable discrimination.

To concentrate on the salient objective area, as well as the the prominent text content, we propose an end-to-end multi-attention classification network MANet, accounting for the basic feature extraction. Besides, a saliency-based regional maximum activation of convolutions (SR-MAC) module for feature representation is proposed to reduce the effect of background regions unrelated to the salient region on MANet's convolution activation and to increase the feature weight of regions related to the salient region; it is capable of objectively aggregating multiple local features and making the feature representation of beauty product images more discriminative.

Besides, word frequency statistics of the text description of each image in Perfect-500K is analyzed using the TF-IDF algorithm and 44 categories are roughly counted. Subsequently, some images are extracted from Perfect-500K dataset associated with these 44 categories and a well-labeled "few-shot" dataset is built, named Perfect-30K.

The proposed method for beauty product image retrieval is illustrated in Figure 1, consisting of the offline part and the online part. Given a query image online, we use MANet to extract the basic feature tensor, and SR-MAC is employed to aggregate local features from it. After post-process with L2 normalization, final features are obtained for the query.

Multi-attention classification network. The proposed MANet has three branches and employs a full convolution structure: it is composed of the saliency attention mechanism, the backbone network, as well as the text attention mechanism. Because various branches have different tasks, different network structures are designed.

Saliency attention mechanism. The saliency attention mechanism in MANet covers the "up-to-down" and "down-to-up" processes of feature learning. The "up-to-down" process learns the high-level semantic information of images, which is capable of finding the location of salient regions but at the expense of the loss of details. Besides, the "down-to-up" process merges a wide range of outputs, so the most visually distinctive objects can be extracted.

During the training phase, the pseudo-saliency mask is adopted to fine-tune the saliency branch

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

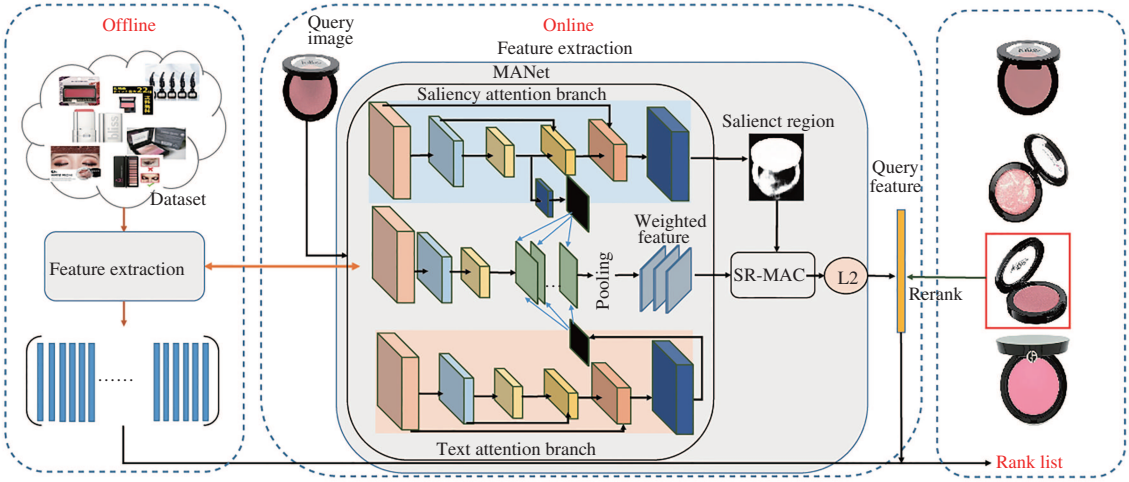

Figure 1 (Color online) Method overview. The whole framework includes two parts: the offline part and the online part.

network. The region with values greater than the mean value of saliency mask in the pseudo-saliency mask is treated as the product object region. The value below the mean in the mask is set to 0; otherwise, it is set to 1. The network is trained by minimizing the following loss function:

$$L_S = 1 - \sum_{i=1}^{N} \frac{2|f(X_i, \Theta) \bigcap Y_i|}{|f(X_i, \Theta)| + |Y_i|}, \tag{1}$$

where $X = \{X_i\}_{i=1}^N$ denotes the set of training images, and $\{Y_i\}_{i=1}^N$ denotes pseudo-saliency mask corresponding to the training images. All the parameters of saliency branch network are defined as Θ. f refers to the saliency attention model.

Text attention mechanism. To draw upon text information, text attention mechanism is adopted to make the network notice the features of the text regions in the learning process. EAST [8] refers to a powerful pipeline that yields fast and accurate text detection; thus the text mask output from it is directly adopted as the output of the branch. Besides, the text mask is adjusted by (2) to retain the weighting effect of saliency attention branch on backbone network feature tensor:

$$T = \text{sigmoid}(T_m) + 1, \tag{2}$$

where T_m refers to the text mask and T is the adjusted text mask. The feature X of backbone network's last convolution layer after being weighted by saliency attention corresponding to channel k with the spatial location of (i, j) is expressed as X_{kij}, and the adjusted text mask T with location of (i, j) is denoted by T_{ij}. The final weighted feature tensor $\tilde{X}$ is produced using the text attention mechanism as written in

$$\tilde{X_k} = T_{ij} \otimes X_{kij}, \tag{3}$$

where $\otimes$ represents element-wise multiplication.

The text mask generated by EAST is applied to weight our feature map because training text detection requires text region annotation, which is beyond the scope of our research. So in order to avoid extra supervised information of text region, text attention branch does not participate in our training, whose parameters are frozen, and we update the parameters of backbone network and saliency attention branch only.

SR-MAC feature representation. The global feature from MANet expresses the overall information of the image, with no discriminative details. To aggregate the local features and obtain discriminative features of the beauty product images, an SR-MAC module is proposed. Besides, the feature of backbone network's last convolutional layer is weighted by saliency attention and text attention is acted as the basic feature tensor to extract local features.

In accordance with R-MAC [9] method, we obtain convolution responses $X^R = \{X^1, \ldots, X^r, \ldots, X^m\}$ corresponding to m regions. We define the regional feature vector:

$$X^r = [f_1^r, \ldots, f_i^r, \ldots, f_K^r], \tag{4}$$

where $f_i^r = \max(C_i^r)$ denotes the maximum activation of the ith channel on the considered region r. Moreover, these local regions are defined on the space Ω of all valid positions for the considered feature map (and not on the input image plane).

Our proposed SR-MAC uses the saliency attention mechanism to assign different weights to respective regions. The local region weight is calculated as follows:

$$W_r = \frac{R \bigcap S}{R}, \tag{5}$$

where R denotes the local region, and S indicates the saliency region. The weighted feature of the local region r is calculated as follows:

$$\acute{X}^r = W_r \cdot [f_1^r, \ldots, f_i^r, \ldots, f_K^r]. \quad (6)$$

The final SR-MAC feature is represented as follows:

$$\Phi = [\acute{F}_1, \ldots, \acute{F}_i, \ldots, \acute{F}_K], \quad \acute{F}_i = \sum_{r=1}^{m} \mathrm{norm}(\acute{f}_i^r), \quad (7)$$

where $\acute{f}_i^r = W_r \cdot f_i^r$.

Experiments. In this study, TF-IDF algorithm is applied to analyze word frequency statistics of all text descriptions appearing in Perfect-500K dataset and 44 rough categories are counted. Sequentially, approximately 35000 images are extracted from Perfect-500K dataset associated with the 44 categories based on the category keywords, and a "few-shot" dataset is built, named Perfect-30K. These 44 categories include lipstick, sunscreen, razor, mask, each of which contains nearly 800 images. Lastly, the Perfect-30K dataset is split into a train dataset and a validation dataset according to the ratio of 8:2.

Compared with RA-MAC [2], MFF [3], and pre-trained ResNet50 [4], which are competitors of Half Million Beauty Product Image Recognition Challenge 2018, our proposed MANet achieves the optimal result on Perfect-500K with 0.395 MAP@7, which is higher than those of RA-MAC with 0.348 MAP@7, MFF with 0.360 MAP@7, and pre-trained ResNet50 with 0.207 MAP@7, respectively.

The advantages of this study are interpreted as follows: (1) saliency and text attention mechanism are used to make MANet pay more attention to the product objects and the text regions in the images; (2) a robust local feature aggregation method is proposed, eliminating the interference of background information and retaining the key local areas in the product object region by using saliency mechanism; (3) a well-labeled beauty product image dataset is built, and the network is trained on it to learn more accurate feature description for beauty products. For more detailed experimental results, please refer to the supplement materials.

Conclusion. An end-to-end multi-attention classification network MANet for beauty product images retrieval is proposed, focusing on the features of saliency regions and text regions in the images and suppressing the interference of irrelevant information. To take the details of beauty product images, an SR-MAC feature representation module is proposed. The feature obtained by SR-MAC eliminates the interference of object-independent region in MANet's convolution activation and enhances the feature weight of regions related to the salient region. Besides, a "few-shot" beauty product dataset, Perfect-30K, with 44 categories for training our proposed MANet is constructed. The retrieval performance of our method on the Perfect-500K dataset outperforms the state-of-the-art methods, which indicates the effectiveness of our method.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 61772108, 61932020, 61976038).

Supporting information Experiments. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Cheng W-H, Jia J, Huang J. Half Million Beauty Product Image Recognition. 2018. https://challenge2018.perfectcorp.com/
2 Lin Z, Yang Z, Huang F, et al. Regional maximum activations of convolutions with attention for cross-domain beauty and personal care product retrieval. In: Proceedings of ACM Conference on Multimedia, 2018. 2073–2077
3 Wang Q, Lai J X, Xu K, et al. Beauty product image retrieval based on multi-feature fusion and feature aggregation. In: Proceedings of ACM Conference on Multimedia, 2018. 2063–2067
4 Lim J H, Japar N, Ng C C, et al. Unprecedented usage of pre-trained CNNs on beauty product. In: Proceedings of ACM Conference on Multimedia, 2018. 2068–2072
5 Sun H Q, Pang Y W. GlanceNets–efficient convolutional neural networks with adaptive hard example mining. Sci China Inf Sci, 2018, 61: 109101
6 Zhong J, Sun Y X, Yu Y L, et al. Attribute-guided network for cross-modal zero-shot hashing. IEEE Trans Neural Netw Learn Syst, 2018. doi: 10.1109/TNNLS.2019.2904991
7 Li H J, Wang X H, Tang J H, et al. Combining global and local matching of multiple features for precise item image retrieval. Multimedia Syst, 2013, 19: 37–49
8 Zhou X, Yao C, Wen H, et al. East: an efficient and accurate scene text detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 5551–5560
9 Tolias G, Sicre R, Jegou H. Particular object retrieval with integral max-pooling of CNN activations. In: Proceedings of the 4th International Conference on Learning Representations, San Juan, 2016