SCIENCE CHINA
Information Sciences

• LETTER •

August 2020, Vol. 63 189202:1–189202:3
https://doi.org/10.1007/s11432-018-9586-9

Salient object detection with side information

Qiuning LI, Yidong LI* & Congyan LANG

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, *China*

Received 1 May 2018/Revised 9 July 2018/Accepted 11 August 2018/Published online 12 March 2020

Citation Li Q N, Li Y D, Lang C Y. Salient object detection with side information. Sci China Inf Sci, 2020, 63(8): 189202, https://doi.org/10.1007/s11432-018-9586-9

Dear editor,

Saliency detection has recently attracted much attention owing to its applicability in several fields of computer vision and machine learning. Convolutional neural network (CNN) is especially successful in generating saliency map end-to-end in salient object detection. These methods can be grouped into two categories: (1) improving the structure of the networks, and (2) training the networks' parameters better than before. However, with the increasing in the number of the images, the side information of images becomes more and more numerous. In fact, the side information has been widely used in other applications with neural networks. The side information are used to improve the performance in graph matching algorithms and object tracking [1, 2]. Zhao et al. [3] deeply exploited the semantic information of the syntactic path based on RNN. Li et al. [4] took advantage of rich semantic information to enhance performance of exploring of indoor environments. The performance of saliency detection is related to many factors, not only visual features, but also semantic information of the images. In order to further improve the accuracy of salient object detection in images, we need to adopt the side information such as image-level tags in our convolutional neural network. Recent work by Zhou et al. [5] has shown that how CNN has remarkable localization ability. Wang et al. [6] proposed a saliency detection method with image-level weak supervision. Nevertheless, the image-level category labels predicted from network may not completely accurate. Therefore, we use image-level tags labeled from the object based on the TBS dataset [7].

We make the following contributions. (1) We use image-level tags labeled on the objects in images to improve the saliency detection results. (2) We extend the global average pooling (GAP) to predict the salient object in complex images and use it as a layer in the deep convolutional network. (3) We conduct extensive experiments and the results show that the proposed method achieves state-of-the-art performance on mean absolutely error (MAE), area under the ROC curve (AUC), and F-measure index on the TBS dataset.

Model and methodology. Our saliency detection model is composed by three main parts: (1) a CNN extracting low, medium and high level features for a given image; (2) a class activation mapping section; and (3) a fully connected conditional random field (CRF) to further improve the saliency map. The structure of the network is shown in Figure 1.

Classification-trained CNN. We obtain the parameters of the network by training the classification task on CNN. The results of the classification are the object tags of the TBS dataset. During the process of training classification, the CNN will extract the feature of the images from different categories in the dataset. This architecture includes 5 blocks from conv1 to conv5. We train the architecture on the popular VGG16, AlexNet [8] and GoogLeNet, which are well known for elegance and simplicity, and at the same time yield nearly state-of-the-art results in image classification and generalization properties. We delete the following part based on the method of adjusting the structure of network in [1]: the layers behind conv5-3 in

* Corresponding author (email: ydli@bjtu.edu.cn)

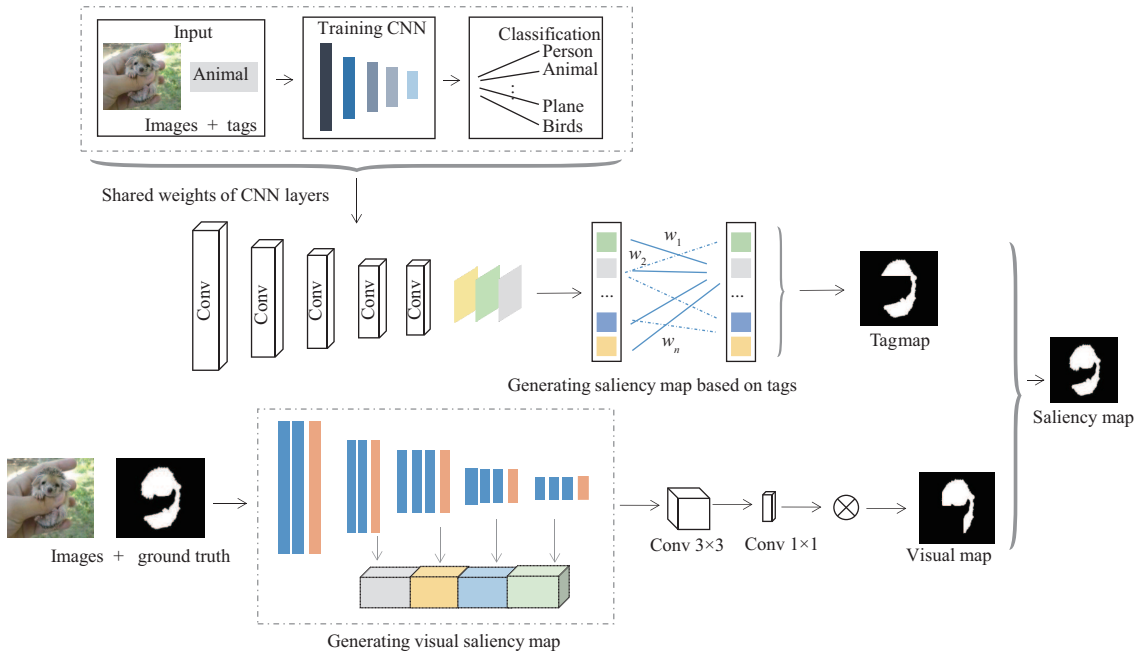

Figure 1 (Color online) The architecture of saliency detection with side information.

VGG16, the layers behind conv5 in AlexNet, and the layers behind pool4 in GoogLeNet.

Training. We use image-level tags of different categories in training stage. We randomly sample a minibatch containing N training saliency maps, and encourage the network to minimize a loss function inspired by three objects. The deviation between predicted values and ground-truth values y_i is weighted by a linear function $\alpha - y_i$. The overall loss function is

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}\left\| \frac{\frac{\phi(x_i)}{\max \phi(x_i)} - y_i}{\alpha - y_i} \right\|^2 + \lambda \|1 - U\|^2. \quad (1)$$

Class activation mapping. The model generates class activation maps (CAM) by using GAP. This process will generate the heatmap of different categories, and the first heatmap is the most discriminative localization of this category. The categories are object tags in the image dataset during training the CNN; therefore the class activation mapping will generate the most discriminative location in the images according to the object tags. A CAM for a particular category indicates the discriminative image regions used by the CNN to identify that category. We use the network architecture similar to Network in Network and GoogLeNet. GAP outputs the spatial average of the feature map of each unit at the last convolutional layer. A weighted sum of these values is used to generate the final output. For a given image, let $u_k(x, y)$ represent the activation of unit k in the last convolutional layer at spatial location (x, y). Therefore, for a given class a, ω_k^a is the weight corresponding to class a for unit k. P_a is the CAM for class a, where each spatial element is given by

$$P_a(x, y) = \sum \omega_k^a u_k(x, y). \quad (2)$$

Feature extraction network. We train another fully convolutional network for extracting the low-level and high-level visual features. The images and the corresponding groundtruth with marked salient regions are as the input of the CNN. This end-to-end network performs a non-linear combination of features extracted from the last convolutional layer to predict saliency maps based on visual feature. We regard the saliency map generated by this feature extraction network as visual map.

Smoothing method. We use a simple thresholding technique to segment the heatmap generated from the CNN. We merge the tag map and the visual map and generate the final saliency map. Some useful boundary information is lost in the obtained saliency maps. The fully connected CRF is incorporated to further improve the spatial coherence of the saliency maps from the deep network [9]. Therefore, we add a fully connected CRF to post-process the final output map. The energy function of CRF is given by

$$E(x) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j). \quad (3)$$

Experiments. In our experiments, we try to answer the following two key questions on saliency

detection. Q1: Are the associated tags beneficial for salient object detection in images with complex backgrounds? Q2: Compared with other state-of-the-art algorithms, how does the proposed method perform on the TBS dataset with image-level tags on main evaluation measures? We evaluate the performance of the proposed method on the TBS dataset. There are many semantic tags of objects in the images of TBS dataset. As mentioned above, we choose VGG16, AlexNet, and GoogLeNet as our pretrained model. We adopt the three popular metrics in this study. The AUC is one of the most widely used metrics for the evaluation of maps predicted from saliency models. We also use the second measure which directly computes the MAE between the estimated saliency map $\bar{S}$ and the binary ground truth $\bar{G}$, both normalized in the range [0,1]. The MAE score is defined as

$$\mathrm{MAE} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\bar{S}(x,y) - \bar{G}(x,y)|. \quad (4)$$

Moreover, we also report the F-measure

$$F_\beta = \frac{(1+\beta^2)\cdot \mathrm{mean(precision)} \cdot \mathrm{mean(recall)}}{\beta^2 \cdot \mathrm{mean(precision)} + \mathrm{mean(recall)}}. \quad (5)$$

Comparisons with other approaches. To answer the question Q1, we compare our method under two conditions: with side information and without side information. The method with side information performs better than that without side information on AUC and F-measure index with more clear edge in visual performance. To answer the question Q2, we compare our method with the other 6 approaches: hierarchical saliency detection (HS), minimum barrier distance (MBD), minimum barrier salient object detection (MB+), static and space-time visual saliency detection (SeR), saliency estimation using a non-parametric low-level vision model (SIM), and visual saliency detection by spatially weighted dissimilarity (SMD). Our method performs better than other methods on AUC, MAE, F-measure index. Because of space limitations, the detailed MAE, AUC, F-measure index and visual comparisons are shown in Appendix A.

Conclusion and future work. We proposed a salient object detection method with side information based on the image-level tags on TBS dataset. In the proposed model, the deep network is introduced to obtain the discriminative location of the salient region. The experimental results indicate that our method achieves better performance than the existing methods on the TBS dataset. In fact, the amount of images is an essential factor in training deep network, however, the dataset with semantic tags is rare. Therefore, how to train the deep network more efficiently with limited images will be an important challenge in future work. We will further improve the performance of saliency detection with semantic tags.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61672088, 61790575), and Fundamental Research Funds for the Central Universities (Grant No. 2018JBZ002).

Supporting information Appendix A. 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 Wang T, Ling H B. Gracker: a graph-based planar object tracker. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 1494–1501
2 Wang T, Ling H B, Lang C Y, et al. Graph matching with adaptive and branching path following. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 2853–2867
3 Zhao Y Y, Qin B, Liu T. Encoding syntactic representations with a neural network for sentiment collocation extraction. Sci China Inf Sci, 2017, 60: 110101
4 Li G S, Chou W S, Yin F. Multi-robot coordinated exploration of indoor environments using semantic information. Sci China Inf Sci, 2018, 61: 079201
5 Zhou B L, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2921–2929
6 Wang L J, Lu H C, Wang Y F, et al. Learning to detect salient objects with image-level supervision. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3796–3805
7 Liang Y, Lang C, Yu J, et al. Salient object detection of social images based on semantic tag context. Int J Sensor Netw, 2017, 23: 233
8 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012. 1097–1105
9 Li G B, Yu Y Z. Deep contrast learning for salient object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 478–487