Local bi-directional funnel network for salient object detection

Zefeng Pan, Junxia Li and Ziyang Wang
B-DAT and CICAAEET, School of Automation, Nanjing University of Information Science and Technology, Nanjing, China
✉Email: junxia1999@163.com

Existing deep-learning-based saliency detection methods mainly design a sophisticated architecture to integrate multi-level convolutional features, for example, recurrent network or bi-directional message passing model, and achieve gratifying performance. However, the direct transmission of information with large span, for example, the information from the deepest layer and the shallowest layer, may lead to antagonistic problems. To address this problem, we propose a local bi-directional funnel network (LBDFN) to effectively integrate multi-level features for salient object detection. In the proposed local bi-directional funnel network, a local bi-directional feature integration (LBDFI) module is designed to fuse the features with similar properties from adjacent convolutional layers, which can solve the problem of mismatch between fusion features well. The extensive experiments on five saliency detection datasets clearly demonstrate that the proposed method outperforms the state-of-the-art approaches.

Proposed method: Our goal is to fuse the features with similar properties from adjacent convolution layers so as to avoid the information mismatch due to the large layer span, and thus to fully exploit the multi-layer and multi-scale salient features. As a matter of fact, different layers of the network are with different properties of the object. The features from the shallow layers are dominated by the salient detailed information, while the features from the deeper layers contain a lot of salient semantic information. Considering the above issues, we design a novel framework, named LBDFN, to effectively fuse multi-level features for salient object detection, which can optimize the features globally. Specifically, in LBDFN, we introduce a local bi-direction feature integration (LBDFI) module to fuse the features from adjacent convolution layers by a local bi-direction structure. Besides, the LBDFN can optimize the features globally.

The overall architecture is illustrated in Figure 1. In our network, we use the ResNet-50 network [5] as the backbone to extract local image features, which contains five convolutional blocks, each of which has several convolutional layers. Let Layer 1, Layer 2, Layer 3, Layer 4 and Layer 5 denote the features from the convolutional layers of five convolutional blocks. Next, the adjacent side-output feature maps are put into the LBDFI module to obtain the local bi-direction feature maps. The details of the LBDFI are shown in the blue box in Figure 1 and the visualization results of the feature maps are shown in Figure 2.

For the sake of narrative, we use Layer 1 and Layer 2, which belong to adjacent convolution layers, as examples. The layers’ features are divided into two groups: the original feature map and the new feature map. The new feature map is generated by the original feature map using a threshold function (see the second column in Figure 2). Note that the threshold function used here is the sigmoid activation, which aims to avoid the repeated information fusion and to generate information fusion rates in the range [0, 1]. It can be formulated as:

\[ F_{new} = \sigma(F_{original}) \]
where $F_{\text{new}}$ and $F_{\text{original}}$ denote the new feature map and the original feature map, respectively, and $\sigma$ represents the sigmoid threshold function.

Next, the new feature map of Layer 1 and the original feature map of Layer 2 are fused by the concatenation operation to obtain the feature map BD-1. Meanwhile, the new feature map of Layer 2 and the original feature map of Layer 1 are fused by the concatenation operation to obtain the feature map BD-2. Last, the feature maps BD-1 and BD-2 are fused by the sum-operation to obtain the local bi-directional feature map. The formula is as follows:

$$F_{\text{LBDFI-1}} = F_{\text{new-1}} \odot F_{\text{original-2}} \oplus F_{\text{new-2}} \odot F_{\text{original-1}},$$

where $F_{\text{LBDFI-1}}$ means the feature map of LBDFI $L - 1$. $F_{\text{new-1}}$ and $F_{\text{original-1}}$ denote the new feature map and the original feature map of Layer 1, respectively. $F_{\text{new-2}}$ and $F_{\text{original-2}}$ represent the new feature map and the original feature map of Layer-2, respectively. $\odot$ represents concatenation operation, and $\oplus$ denotes the sum-operation.

Generally, we fuse the adjacent convolution layers by the LBDFI to obtain complementary feature maps. Then, the adjacent complementary feature maps also are put into LBDFI to generate new complementary feature maps. Finally, we build a funnel shape network which can fuse multi-scale and cross-layer salient features. To note that, we only supervise the last feature map by the ground truth. And our loss function is the binary cross entropy loss which treats all pixels equally.

**Experimental results:** All experiments here are conducted using a single NVIDIA 1080Ti GPU. The proposed network is based on the current popular PyTorch framework. And the training dataset of our network is DUTS-TR. We evaluate our approach on five public saliency detection benchmarks, including SOD, ECSSD, PASCAL-S, DUTS-TE and HKU-IS. SOD, ECSSD, PASCAL-S, DUTS-TE and HKU-IS contain 300 images, 1000 images, 850 images, 5019 images and 4447 images, respectively. To evaluate the performance of the proposed method, we test our results based on two metrics including F-measure score and mean absolute error (MAE) score. Specifically, F-measure is an overall performance metric, which is computed by the weighted harmonic of precision and recall. MAE denotes the average absolute per-pixel difference between a predicted saliency map and its ground truth. Table 1 gives the quantitative results of the proposed approach against five state-of-the-art methods in terms of F-measure and MAE on the five datasets. From Table 1, we can find that our method can significantly outperform the state-of-the-art approaches on all datasets.

Figure 3 shows some saliency maps generated by the proposed method and the four state-of-the-art salience approaches for a subjective comparison. As can be seen, compared to other methods, our method generates more accurate saliency maps in various challenging cases. For images having a low contrast between objects and background (e.g. rows 1, 2, and 4), and having a complex circumstances (e.g. row 5 in Figure 3), our method still generates accurate saliency maps, while most of the existing saliency methods cannot effectively highlight the entire salient objects. Besides, for the images with inverted reflection in water (e.g. row 6) or detail structure (e.g. row 7), our method also can effectively detect the whole salient objects with well-defined boundaries.

**Conclusion:** We propose a novel LBDFN to fully exploit the multi-layer and multi-scale convolutional features for salient object detection. In LBDFN, an LBDFI module is designed to fuse the features with similar properties from adjacent convolution layers so as to avoid the information mismatch due to the large layer span. The experimental evaluation on five datasets demonstrates that the proposed approach provides more accurate saliency maps compared to the state-of-the-art saliency detection methods. We hope our work can provide valuable insights into enhancing the discriminant ability of feature representation for saliency detection.

**Acknowledgements:** This work was supported in part by the National Major Project of China for New Generation of AI under Grant 2018AAA0100400, in part by the National Science Fund of China under Grant 61702272 and Grant 61802199.

© 2021 The Authors. *Electronics Letters* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited. Received: 24 June 2020 Accepted: 12 October 2020 doi: 10.1049/el.2021.12021

**References**

1. Hou, Q.B., et al.: Deeply supervised salient object detection with short connections. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(4), 815–828 (2019)
2. He, J.Z., et al.: Bi-directional cascade network for perceptual edge detection. In: *IEEE International Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 3823–3832 (2019)
3. Zhang, L., et al.: A bi-directional message passing model for salient object detection. In: *IEEE International Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 1741–1750 (2018)

---

**Table 1. Quantitative evaluation in terms of F-measure and MAE scores**

| Dataset   | Ours | [6] | [8] | [7] | [4] | [3] |
|-----------|------|-----|-----|-----|-----|-----|
| ECSSD     | 0.04 | 0.042 | 0.041 | 0.046 | 0.056 | 0.045 |
| F-measure | 0.927 | 0.924 | 0.926 | 0.925 | 0.908 | 0.917 |
| PASCAL-S  | 0.111 | 0.123 | 0.145 | 0.121 | 0.16 | 0.121 |
| F-measure | 0.844 | 0.827 | 0.802 | 0.805 | 0.79 | 0.827 |
| SOD       | 0.094 | 0.111 | 0.125 | 0.104 | 0.124 | 0.108 |
| F-measure | 0.836 | 0.786 | 0.82 | 0.818 | 0.81 | 0.806 |
| DUTS-TE   | 0.047 | 0.052 | -   | 0.051 | 0.059 | 0.049 |
| MAE       | 0.841 | 0.815 | -   | 0.838 | 0.807 | 0.827 |
| HKU-IS    | 0.034 | 0.037 | 0.036 | 0.043 | 0.045 | 0.039 |
| F-measure | 0.918 | 0.907 | 0.904 | 0.906 | 0.901 | 0.91 |

The bold values show the best performance.
4 Chen, S.H., et al.: Reverse attention-based residual network for salient object detection. IEEE Trans Image Process 29, 3763–3776 (2020)
5 He, K.M., et al.: Deep residual learning for image recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 770–778 (2016)
6 Wang, W.G., et al.: Salient object detection with pyramid attention and salient edges. In: IEEE International Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp. 1448–1457 (2019)
7 Liu, N., Han, J.W., Yang, M.-H.: Picanet: Learning pixel-wise contextual attention for saliency detection. In: IEEE International Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 3089–3098 (2018)
8 Deng, Z., et al: R3Net: Recurrent residual refinement network for saliency detection. In: Proceedings of the Twenty-Seventh International Joint Conference Artificial Intelligence Stockholm, Sweden, (2018)