P1-5

17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

FBNet: FeedBack-Recursive CNN for Saliency Detection

Guanqun Ding[1,2], Nevrez İmamoğlu[2], Ali Caglayan[2], Masahiro Murakawa[1,2], Ryosuke Nakamura[2]
[1]Graduate School of Science and Technology, University of Tsukuba, Tsukuba, Japan
[2]Artificial Intelligence Research Center,
National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
s2030174@s.tsukuba.ac.jp,
{nevrez.imamoglu, ali.caglayan, m.murakawa, r.nakamura}@aist.go.jp

Abstract

Saliency detection research has achieved great progress with the emergence of convolutional neural network (CNN) in recent years. Most deep learning based saliency models mainly adopt the feed-forward CNN architecture with heavy burden of parameters to learn features via bottom-up manner. However, this forward only process may ignore the intrinsic relationship and potential benefits of top-down connections or information flow. To the best of our knowledge, there is not any work to explore the feedback connection especially in a recursive manner for saliency detection. Therefore, we propose and explore a simple, intuitive yet powerful feedback recursive convolutional model (FBNet) for image saliency detection. Specifically, we first select and define a lightweight baseline feed-forward CNN structure (∼4.7MB), then the high-level multi-scale saliency features are fed back to the low-level convolutional blocks in a recursive process. Experimental results show that the feedback recursive process is a promising way to improve the performance of the baseline forward CNN model. Besides, despite having relatively few CNN parameters, the proposed FBNet model achieves competitive results on the public saliency detection benchmarks.

1 Introduction

Attention mechanism plays an important role for the perception in human visual system (HVS) [1, 2]. By imitating the similar mechanism of HVS, attention modeling research (e.g. prediction of saliency maps) aims to find the most attractive locations or regions from a visual stimuli (e.g. an image) [3]. Saliency detection research has been explored extensively in the past decade since it can be integrated to various vision tasks for improved results such as object tracking[4], person re-identification[5], data augmentation[6], and video streaming [7].

Generally, saliency detection research can be categorized into two tasks: human eye fixation prediction (also referred as saliency map prediction) [3, 8] and salient object detection [9]. The former represented in Gaussian-like saliency map and it focuses on predicting the fixations or locations of images where the human eyes are most attracted to [3, 8]. On the other hand, the latter aims to identify the regions of images or objects where focus of attention belongs to the salient object/s [9]. In this paper, we conduct the exploration studies focusing on the prediction of saliency maps based on supervised CNNs, which has been extensively studied by researchers in the past several decades [8, 10, 11]. Unlike the early traditional models of evaluating the contrast on low-level features [3, 12, 13], supervised saliency models [8, 10] have achieved a significant progress with the great breakthrough of deep learning techniques. However, most existing CNN saliency models mainly utilize the forward pathway to learn the visual representations [8, 10, 11]. In addition, CNN based saliency detection models mostly consist of large number of parameters and high computational cost, such as Shallow-Net in [10] (i.e. 2.5 GB model size). Therefore, it is highly desired to present a lightweight yet efficient saliency model to alleviate the gap between bottom-up and top-down contextual features from the observed visual stimuli.

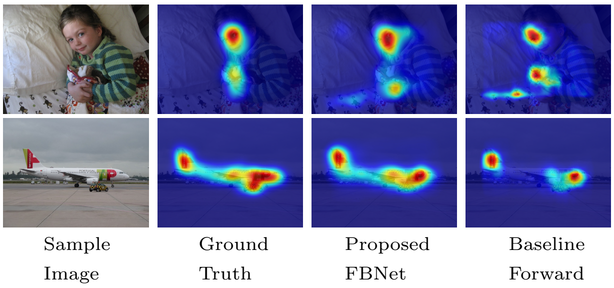

Figure 1. Samples for the visual comparison of the proposed FBNet with the baseline forward models on the validation set of SALICON [26].

Biologically, the feedback mechanism is generally used to amplify or inhibit a certain pathway and keep the balance of a system [14, 15]. Human brain and visual system also utilize the feedback mechanism to process complex cognition tasks [16, 17]. Inspired by this concept, some of the recent CNN models with recurrent

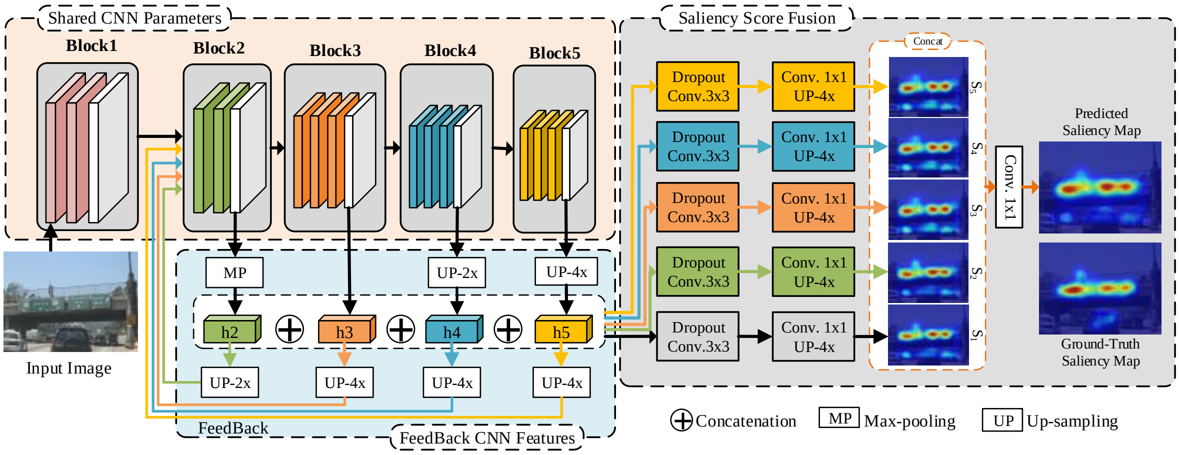

Figure 2. Overview of the proposed feedback saliency model.

and feedback structures have been explored to learn the semantic representations in various computer vision applications [18, 19, 20, 21, 22, 23, 25]. For instance, Zamir *et al.* establish feedback networks to demonstrate that feedback architectures are able to learn better representations than the feed-forward networks [18]. Liu *et al.* propose a weakly supervised geo-semantic segmentation model based on feedback neural networks [20]. In the study [23], the authors explore the effectiveness of the feedback features for satellite image classification by outperforming baseline feed-forward only model. Moreover, they present a feedback recursive model without the need of additional parameters [23]. Their model [23] is different from the traditional feedback networks, which require additional network parameters for each feedback connection.

Inspired by the feedback-recursive CNN approach in [23] and the simple yet efficient feed-forward saliency model in [8], in this work, we introduce a new model with feedback connections between CNN layers for image saliency prediction. **Our contributions in this study can be summarized as follows:**

i) We propose a novel and lightweight feedback convolutional model for image saliency detection, which includes bottom-up forward and top-down feedback contextual feature pathways with shared CNN parameters. Figure 1 shows some visual saliency samples of the proposed model with forward and feedback recursive features and forward only baseline model.

ii) The proposed FBNet model combines several salient cues from both one forward and four feedback processes of the same CNN encoder to obtain final saliency prediction. Therefore, the final saliency map model is influenced by these intermediate saliency scores from different semantic connections. These forward and feedback connections guide the network to learn more discriminating contextual representations.

iii) Finally, we also explore different numbers of feedback architectures to investigate the best way for the combination of the saliency features. Experimental results demonstrate that feedback connections outperform the forward structure, and the proposed model with few parameters achieves comparable performance on the public image saliency detection benchmarks.

2 Proposed Feedback-Recursive Networks

The proposed Feedback Network (FBNet) is demonstrated in Figure 2. As we can see from this architecture, the proposed model mainly includes three components: a feed-forward feature extractor with shared CNN parameters, a recursive module with feedback connections by using multi-scale features, and a saliency score fusion module with deeply supervisions.

2.1 Feed-Forward Feature Extractor

As shown in Figure 2, we first build a feed-forward feature extractor based on the ML-Net saliency model in [8], which uses the VGG [24] core network with 5 CNN blocks to learn the multi-scale features, then combines the features from the last three blocks $\{h3,h4,h5\}$ to obtain the saliency prediction [8]. However, unlike the ML-Net [8], in this work, we only neglect the input block *h1*, and fuse features from all the other four blocks $\{h2,h3,h4,h5\}$. In addition, to utilize feedback connections recursively, we set each layer to a fixed $c{=}64$ channels output, which also provides a more compact model compared to the ML-Net [8].

In order to obtain richer visual features for saliency detection, we up-sample and concatenate the multi-scale features learned from the last four blocks, and then feed them into the saliency score fusion module. The forward saliency score component (see the black-arrows in saliency score fusion module of Figure 2) contains a dropout layer, a convolutional layer with 3 × 3 kernel, a convolutional layer with 1 × 1 kernel and an up-sampling layer with a scale rate of 4. After obtaining the saliency score map (*score1*) for the forward pass, we calculate the *loss1* between *score1* and the eye fixation ground-truth map.

2.2 Feedback Recursive Module

In this work, we propose to recursively feed the learned multi-scale features back to the previous blocks with shared CNN parameters. To this end, intuitively, the network attempts to recursively aggregate contextual information through feedback connections to a holistic description. Since the first block includes the image input layer, we feedback the features to the second block.

To make full use of the features with forward bottom-up and feedback top-down manners, the feature *h2* from *block2* is fed back to the feed-forward feature extractor with the same CNN parameters to obtain the feedback features (see the green arrows in Figure 2). Similar to the forward saliency score component, the *score2* is predicted in a similar manner by using the feedback features. In order to gain compatible *h2* features with the *block2*, an up-sampling layer is used to re-scale the features.

Similar to the feature *h2*, the feature *h3* (orange arrows) from *block3*, the feature *h4* (blue arrows) from *block4*, and the feature *h5* (yellow arrows) from *block5* are also recursively fed back to the feed-forward feature extractor using the same CNN parameters, thereby we obtain the saliency score maps *score3*, *score4*, and *score5*. Then, we compute the losses between these score maps and the eye fixation map. To combine the forward and feedback saliency features, we concatenate the 5 predicted scores after obtaining the 5 score maps from a forward pass and 4 feedback processes, and then generate a final saliency score from a 1×1 convolutional layer.

2.3 Loss Functions for Training FBNet

We calculate several losses from the forward and feedback outputs to supervise and optimize the parameters of the proposed feedback-recursive network. MSE (Mean Square Error) loss is employed to measure the distance between predictions and labels as in the use of the ML-Net [8].

First, the total loss from the $m = \{1, 2, 3, 4, 5\}$ saliency scores (i.e. saliency maps of the single forward and the four feedback passes) can be represented as follows:

$$Loss_{score} = \frac{1}{m}\frac{1}{w}\frac{1}{h}\sum_{k=1}^{m}\sum_{i=1}^{w}\sum_{j=1}^{h}\|S_{i,j}^{m} - G_{i,j}\|^2 \qquad (1)$$

where w and h denote the width and height of an input image; the $S_{i,j}^{m}$ and the $G_{i,j}$ represent saliency value of a location (i, j) in the m^{th} saliency score map and ground truth map. Finally, we measure the cost between the final fused saliency prediction and ground truth map by the following function:

$$Loss_{fuse} = \frac{1}{w}\frac{1}{h}\sum_{i=1}^{w}\sum_{j=1}^{h}\|S_{i,j}^{fuse} - G_{i,j}\|^2 \qquad (2)$$

where the $S_{i,j}^{fuse}$ represents the saliency value of a location (i, j) in the final fused saliency map. The overall loss for optimizing the proposed FBNet saliency model can be calculated as:

$$Loss = Loss_{score} + Loss_{fuse} \qquad (3)$$

3 Experimental Evaluation

We first investigate different feedback connections based on the feed-forward CNN model and evaluate their performances in a large saliency detection benchmark to demonstrate the effectiveness of the proposed feedback-recursive saliency network. Then, we compare the proposed approach with several existing saliency detection methods and present detailed analyses based on the experimental results.

Datasets: We conduct the comparison experiments using the SALICON [26] benchmark, which is a well-known public benchmark of eye fixation predictions. The large scale SALICON [26] dataset officially contains 10,000 images in training set, 5,000 images in validation set, and 5,000 images in testing set.

Evaluation Metrics: Similar to the studies in [8, 10, 11], we report the performance results by using the popular metrics including Pearson's linear correlation coefficient (CC), area under ROC curve (AUC), shuffled AUC (sAUC), normalized scanpath saliency (NSS), similarity (SIM), and Kullback-Leibler Divergence (KLDiv). The AUC includes the AUC$_{Judd}$ and the AUC$_{Borji}$. Note that the larger the values of CC, AUC, sAUC, NSS, and the smaller the value of KLDiv, the better the performance of the saliency method. We refer the reader to [8, 10, 11] for more details about the metrics.

Implementation Details: Our source code is implemented on an Ubuntu operating system using the popular Pytorch library. SGD optimizer is used for training with a batch size, momentum and weight decay values as 10, 0.9 and 1e-4, respectively, and learning rate is set to 1e-3 with a decay rate of 0.1 every five epochs until learning stops when it reaches epsilon.

3.1 Quantitative and Qualitative Evaluation

In order to demonstrate the superior performance of the proposed feedback model, we first show the visual comparison samples from the validation set of SALICON [26]. Figure 1 shows that feedback model can

Table 1. Performance evaluation for different number of feedbacks on the validation set of SALICON [26].

Method	AUC$_{Judd}$ ↑	AUC$_{Borji}$ ↑	sAUC ↑	CC ↑	NSS ↑	KLdiv ↓	SIM ↑
Baseline_Forward	0.8129	0.7509	0.6384	0.5564	1.1334	2.9065	0.5505
FBNet_1_FeedBack	0.8254	0.7774	0.6622	0.6581	1.3628	2.5772	0.6042
FBNet_2_FeedBack	0.8745	0.8087	0.6732	0.6887	1.4137	1.3316	0.6362
FBNet_3_FeedBack	0.8850	0.8243	0.6854	0.7239	1.4520	1.1096	0.6398
Proposed FBNet	**0.9054**	**0.8376**	**0.7151**	**0.7841**	**1.6091**	**0.8787**	**0.6900**

Table 2. Performance comparison of the proposed FBNet with the baseline forward and deep models in the literature on the testing set of SALICON [26]. Note that "*" denotes the unreported results in the studies.

Metric Method	AUC-B ↑	sAUC ↑	CC ↑	SIM ↑	SIZE (MB)
Baseline_Forward	0.7700	0.6420	0.5620	0.5490	**4**
Proposed FBNet	0.8430	0.7060	**0.7850**	**0.6940**	**4.7**
MLNet_Forward	0.7880	0.6540	0.5950	0.5770	4
MLNet_FBNet	0.8390	0.6970	0.7660	0.6770	4.7
ML-Net [8]	0.8660	0.7680	0.7430	*	123.7
DeepNet [10]	0.8580	0.7240	0.6220	0.6090	103
ShallowNet [10]	0.8364	0.6698	0.5957	0.5198	2500
SalGAN [27]	**0.8840**	**0.7720**	0.7810	*	130
BMS [12]	0.7899	0.6935	0.4268	*	*
GBVS [13]	0.7899	0.6303	0.4212	0.4460	*
Itti [3]	0.6669	0.6101	0.2046	0.3870	*

capture more details from the visual stimuli than the forward model (e.g. the middle region of the airplane). In other words, the feedback connections effectively provide abundant visual features from high-level layers. Table 1 presents the quantitative results on the validation set of SALICON [26]. As seen from the results, it is clear that the feedback architecture performs better than the forward structure in all metrics. Especially, the model with 4 feedback connections demonstrates a dramatic performance boost over the forward structure. Furthermore, the proposed FBNet model with four feedbacks (from the blocks {*h5,h4,h3,h2*} in Figure 2) is better than the model with three feedbacks (from the blocks {*h5,h4,h3*}), two feedbacks (from the blocks {*h5,h4*}), and one feedback (from only {*h5*}), which are investigated as an ablation study. The ablative results in Table 1 show that the more the number of feedback pathway with top-down manner, the better the performance.

We also compare the proposed model with several existing saliency detection models on the testing set of SALICON [26]. Note that the prediction results can only be evaluated by submitting them to the official website of SALICON [26]. As seen from the Table 2, FBNet variants outperform the baseline forward model. Furthermore, the proposed FBNet with the feature fusion on four blocks {*h5,h4,h3,h2*}) performs better than both forward and our feedback version of MLNet [8] with multi-scale fusion on three blocks {*h5,h4,h3*} (given as MLNet_Forward and MLNet_FBNet in Table 2, note that their number of channel is $c = 64$). Additionally, the proposed feedback model with few parameters (∼4.7MB) achieves comparable performance among the existing saliency methods, which also confirms that employing feedback connections in this way is highly effective to improve the performance of saliency detection methods. The proposed FBNet performs better than existing methods based on CC and SIM by giving 0.7850 and 0.6940 values, respectively, and the performance on other metrics are also close to the existing methods.

In summary, the proposed model with feedback connections achieves a significant improvement on the baseline forward-only model for saliency prediction. Moreover, our approach yields competitive results on the public saliency detection benchmarks. However, the limitation of the proposed FBNet is that it will consume more memory and time since the recursive iteration compared with the feed-forward baseline model.

4 Conclusion

In this paper, a novel and lightweight feedback-recursive CNN is proposed to learn abundant contextual features for saliency detection. Our FBNet model recursively feed back the multi-scale features from high-level layers to the low-level layer. Thus, the proposed FBNet model remains compact while learning rich saliency features effectively. The outputs from forward and feedback pathways are jointly supervised by the saliency label, enforcing the model to learn more discriminating features. Our experimental results show that using feedback connections in a deep learning model as in our work is a promising way to provide abundant richer visual features for saliency detection. As a future work, conventional recurrent and feedback connections can be jointly explored on the saliency detection task.

Acknowledgments

This paper is in part based on the results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan.

References

[1] D. Norman, and T. Shallice: "Attention to Action," *Consciousness and Self-regulation*, pp.1–18, 1986.

[2] H. Pashler: "The Psychology of Attention," *MIT press*, 1999.

[3] L. Itti, C. Koch, and E. Niebur: "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.20, no.11, pp.1254–1259, 1998.

[4] P. Zhang, T. Zhuo, W. Huang, K. Chen, M. Kankanhalli: "Online object tracking based on CNN with spatial-temporal saliency guided sampling," *Neurocomputing*, vol.257, pp.115–127, 2017.

[5] R. Zhao, W. Ouyang, X. Wang: "Unsupervised salience learning for person re-identification," *IEEE Conference on Computer Vision and Pattern Recognition*, pp.3586–3593, 2013.

[6] A F M Shahab Uddin, Mst. Sirazam Monira, Wheemyung Shin, TaeChoong Chung, Sung-Ho Bae: "SaliencyMix: A Saliency Guided Data Augmentation Strategy for Better Regularization," *International Conference on Learning Representations*, 2021.

[7] C. Ozcinar, N. Imamoglu, W. Wang, and A. Smolic: "Delivery of omnidirectional video using saliency prediction and optimal bitrate allocation" *Signal, Image and Video Processing*, Springer, pp.1-8, 2020.

[8] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara: "A Deep Multi-level Network for Saliency Prediction," *International Conference on Pattern Recognition*, pp.3488–3493, 2016.

[9] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, P. H. Torr: "Deeply Supervised Salient Object Detection with Short Connections," *IEEE Conference on Computer Vision and Pattern Recognition*, pp.3203–3212, 2017.

[10] J. Pan, E. Sayrol, X. Giro-i-Nieto, K. McGuinness, and N. E. O'Connor: "Shallow and Deep Convolutional Networks for Saliency Prediction," *IEEE Conference on Computer Vision and Pattern Recognition*, pp.598–606, 2016.

[11] R. Droste, J. Jiao, and J. A. Noble: "Unified Image and Video Saliency Modeling," *European Conference on Computer Vision*, pp.419–435, 2020.

[12] J. Zhang and S. Sclaroff: "Saliency Detection: A Boolean Map Approach," *IEEE International Conference on Computer Vision*, pp.153–160, 2013.

[13] J. Harel, C. Koch, and P. Perona: "Graph-Based Visual Saliency," *International Conference on Neural Information Processing Systems*, pp.545–552, 2006.

[14] C. Cosentino, and D. Bates: "Feedback Control in Systems Biology," *Crc Press*, 2011.

[15] R. Thomas, and R. d'Ari: "Biological Feedback," *CRC press*, 1990.

[16] S. J. Ashford and L. L. Cummings: "On the Mechanisms of the Feedback Control of Human Brain-wave Activity," *Mind/Body Integration*, pp.325–340, 1979.

[17] V. S. Ramachandran, and E. L. Altschuler: "The Use of Visual Feedback, in Particular Mirror Visual Feedback, in Restoring Brain Function," *Brain*, vol.132, no.7, pp.1693–1710, 2009.

[18] V. S. Ramachandran, and E. L. Altschuler: "Feedback Networks," *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1308–1317, 2017.

[19] C. Cao, X. Liu, Y. Yang, et al.: "Look and Think Twice: Capturing Top-down Visual Attention with Feedback Convolutional Neural Networks," *IEEE International Conference on Computer Vision*, pp.2956–2964, 2015.

[20] X. Liu, A. Zhang, T. Tiecke, A. Gros, and T. S. Huang: "Feedback Neural Network for Weakly Supervised Geosemantic Segmentation," *arXiv:1612.02766*, 2016.

[21] M. F. Stollenga, J. Masci, F. Gomez, J. Schmidhuber: "Deep Networks with Internal Selective Attention Through Feedback Connections," *International Conference on Neural Information Processing Systems*, pp.3545–3553, 2014.

[22] Y. Tang, X. Wu, and W. Bu: "Deeply-Supervised Recurrent Convolutional Neural Network for Saliency Detection," in *Proceedings of the 24th ACM international conference on Multimedia (MM '16)*, pp. 397–401, 2016.

[23] N. Imamoglu, M. Kimura, H. Miyamoto, A. Fujita and R. Nakamura: "Solar Power Plant Detection on Multi-Spectral Satellite Imagery using Weakly Supervised CNN with Feedback Features and m-PCNN Fusion," in *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 183.1-183.12., September 2017.

[24] K. Simonyan, A. Zisserman: "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations*, 2015.

[25] V. Poliyapram, N. Imamoglu, R. Nakamura: "Recurrent Feedback CNN for Water Region Estimation from Multitemporal Satellite images," in *Proc. SPIE 11155, Image and Signal Processing for Remote Sensing XXV, 111550T*, 2019.

[26] M. Jiang, S. Huang, J. Duan, Q. Zhao: "SALICON: Saliency in Context," *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1072–1080, 2015.

[27] J. Pan, C. Ferrer, K. McGuinness, N. O'Connor, J. Torres, E. Sayrol, and X. Giro-i-Nieto: "SalGan: Visual Saliency Prediction with Generative Adversarial Networks," *arXiv:1701.01081*, 2017.