Deeply Supervised Salient Object Detection

Hongzhi Liu

Jun 6, 2018

Abstract

Recent progress on saliency detection is substantial, benefiting mostly from the explosive development of Convolutional Neural Networks (CNNs). Semantic segmentation and saliency detection algorithms developed lately have been mostly based on Fully Convolutional Neural Networks (FCNs). Today, I read a thesis written by Mingming Cheng, who is from Nankai University. His team propose a new saliency method by introducing short connections to the skip-layer structures within the HED architecture. Their method produces state-of-the-art results on 5 widely tested salient object detection benchmarks, with advantages in terms of efficiency, effectiveness and simplicity over the existing algorithms. Furthermore, based on their method, HUAWEI successfully developed intelligent self photo applications which is used in flagship mobile phones, providing powerful support for HUAWEI's intelligent photo taking.

1. Overview of the Framework

The goal in salient object detection is to identify the most visually distinctive objects or regions in an image. Salient object detection methods commonly serve as the first step for a variety of computer vision applications. Inspired by cognitive studies of visual attention, computational saliency detection has received great research attention in the past two decades. The Holistically-Nested Edge Detector (HED) [5] model, which explicitly deals with the scale space problem, has lead to large improvements over generic FCN models in the context of edge detection. However, the skip-layer structure with deep supervision in the HED model does not lead to obvious performance gain for saliency detection.

As demonstrated in Fig. 1, Dr. Cheng observes that deeper side outputs encodes high level knowledge and can better locate salient objects and shallower side outputs capture rich spatial information. This motivated him to develop a new method for salient object detection by introducing short connections to the skip-layer structure within the HED architecture. By having a series of short connections from deeper side outputs to shallower ones, their new framework offers two advantages mentioned in the paper [2]. One is that high-level features can be transformed to shallower side-output layers. And the other is that shallower side-output layers can learn rich low-level features.

2. Deep Supervision with Short Connections

As pointed out in most previous works, a good salient object detection network should be deep enough such that multi-level features can be learned. Besides, it should have multiple stages with different strides so as to learn more inherent features from different scales.

2.1. HED-based saliency detection

The team propose a top-down method to reasonably combine both low-level and high-level features for accurate saliency detection. They shall start out with the standard HED architecture as well as its extended version, a special case of this work, for salient object detection and gradually move on to the proposed architecture. The side objective function of HED can be given by as Equation 1:

$$L_{side}(W, w) = \sum_{m=1}^{M} \alpha_m l_{side}^{(m)}(W, w^{(m)}), \quad (1)$$

where α_m is the weight of the mth side loss and $l_{side}^{(m)}$ side denotes the image-level class-balanced cross-entropy loss function for the mth side output. Besides, a weighted fusion layer is added to better capture the advantage of each side output. The fusion loss at the fusion layer can be expressed as Equation 2:

$$L_{fuse}(W, w, f) = \sigma(Z, h(\sum_{m=1}^{M} f_m A_{side}^{(m)})), \quad (2)$$

Therefore, the final loss function can be given by as Equation 3:

$$L_{final}(W, w, f) = L_{fuse}(W, w, f) + L_{side}(W, w), \quad (3)$$

HED connects each side output to the last convolutional layer in each stage of the VGGNet [4]. Each side output

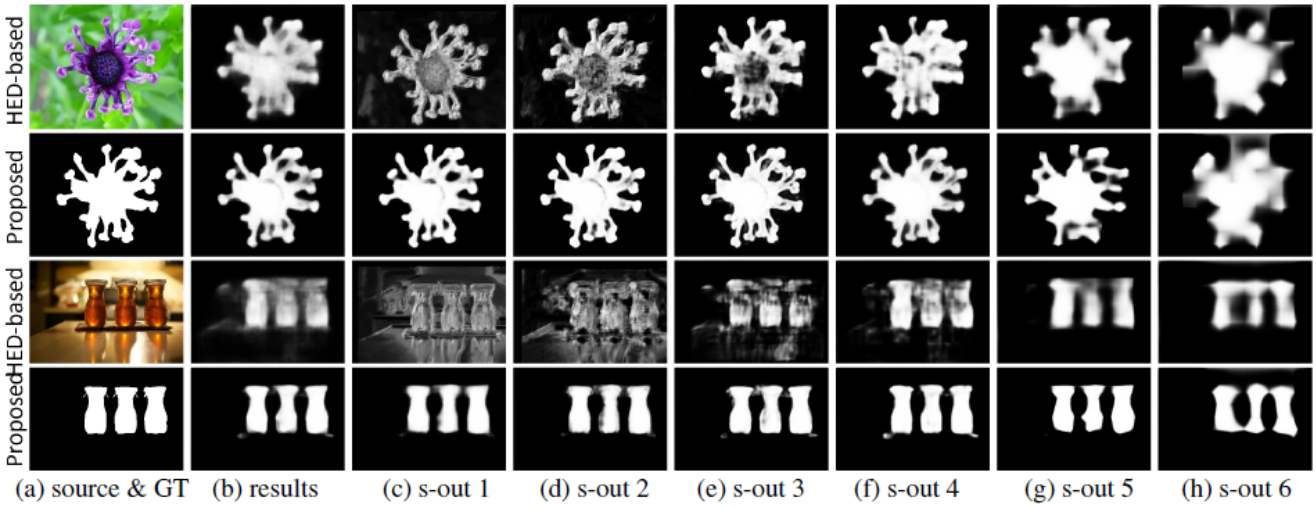

Figure 1. Visual comparison of saliency maps produced by the HED-based method and theirs. Though saliency maps produced by deeper side output look similar, because of the introduced short connections, each shallower side output can generate satisfactory saliency maps and hence a better output result.

Table 1. The performance of different architectures on PASCALS dataset [3]. "*" represents the pattern used in this paper.

No.	Architecture	F_β
1	Hypercolumns [1]	0.818
2	Original HED [5]	0.791
3	Enhanced HED	0.816
4	Pattern 1	0.816
5	Pattern 2	0.824
6	Pattern 3*	0.830

is composed of a one-channel convolutional layer with kernel size 1×1 followed by an up-sampling layer for learning edge information.

2.2. Short Connections

Their approach is based on the observation that deeper side outputs are capable of finding the location of salient regions but at the expense of the loss of details, while shallower ones focus on low-level features but are short of global information. These phenomenons inspire them to utilize the following way to appropriately combine different side outputs such that the most visually distinctive objects can be extracted.

3. Experimental Results of the Method

The team's experiment with different design options and different short connection patterns to illustrate the effectiveness of each component of our method. The performance is listed in Table 1. As can be seen from Table 1, with the increase of short connections, their approach gradually achieves better performance.

References

[1] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015. 2

[2] Q. Hou, M. M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr. Deeply supervised salient object detection with short connections. In *CVPR*, 2017. 1

[3] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014. 2

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[5] S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV*, 2015. 1, 2