Learning to Detect Salient Objects with Image-level Supervision

Chaonan Song

May 31, 2018

Abstract

Today I read the second half of the dissertation. In this part, Professor Wang introduced Weakly Supervised Saliency Detection and described the weak supervision training method. Professor Wang has done a series of experiments using existing DNN-based methods to train and evaluate publicly significant data sets. Professor Wang compared the effectiveness of various methods and finally confirmed the excellence of WSS.

1. Weakly Supervised Saliency Detection

The CNN used for image-level label prediction is usually composed of a series of convolutional layers followed by several fully connected layers. Although CNN trained on image-level labels, it has been shown that higher convolutional layers can capture distinct parts of the object and act as object detectors. Based on the above discussion, recent work on dense tag prediction tasks (eg, semantic segmentation) has begun to explore the full convolutional network (FCN) to maintain spatial location information.

1.1. Aggregation Through Global Smooth Pooling

The global pooling operation is performed independently in each channel of the score graph S. Professor Wang only considered the score graph of one channel. For GMP that only considers the maximum response value, perform aggregation by Equation 1.

$$s = \max_{\omega \in \Delta} \omega^T s \tag{1}$$

Although both GMP and GAP have been successfully used for fractional aggregation, they are not optimal for grounding image-level tags in the object area. GAP encourages the detector to have the same response in all spatial locations, which is unreasonable and leads to overestimation of the target area. An example is shown in Figure 1.

1.2. Foreground Inference Network

When the joint training with the GSP layer on the image level tag, the score map S generated by the FCN may capture the object region in the input image, each channel corresponding to the object category. For saliency detection, we do not pay special attention to object categories, but aim to discover all categories of salient object areas. To obtain such category-independent saliency maps, you can simply average the category score maps on all channels.

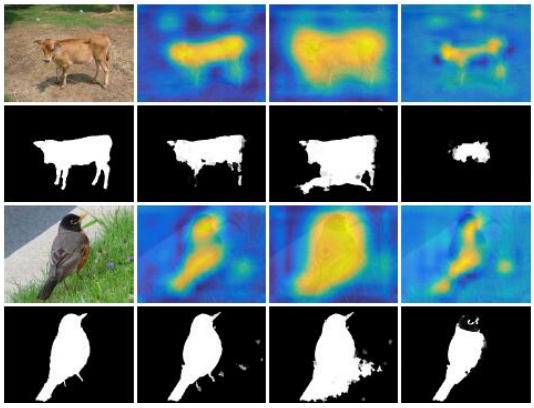

Figure 1. Comparison between different pooling methods. Firstand third rows: foreground maps produced by the FIN with different pooling methods. Second and fourth rows: refined saliency maps based on the foreground maps.

1.3. Pretrainingon Image-level Tags

The training image in the detection data set typically contains a plurality of objects from different categories, as opposed to an image classification data set having only one annotation category in each image. Therefore, the target detection data set is more suitable for solving significant co-occurrence problems Borji *et al.* [1].

2. Experiments

To facilitate fair comparison and effective model training, Professor Wang provided a large data set called DUTS, which contains 10,553 training images and 5,019 test images.

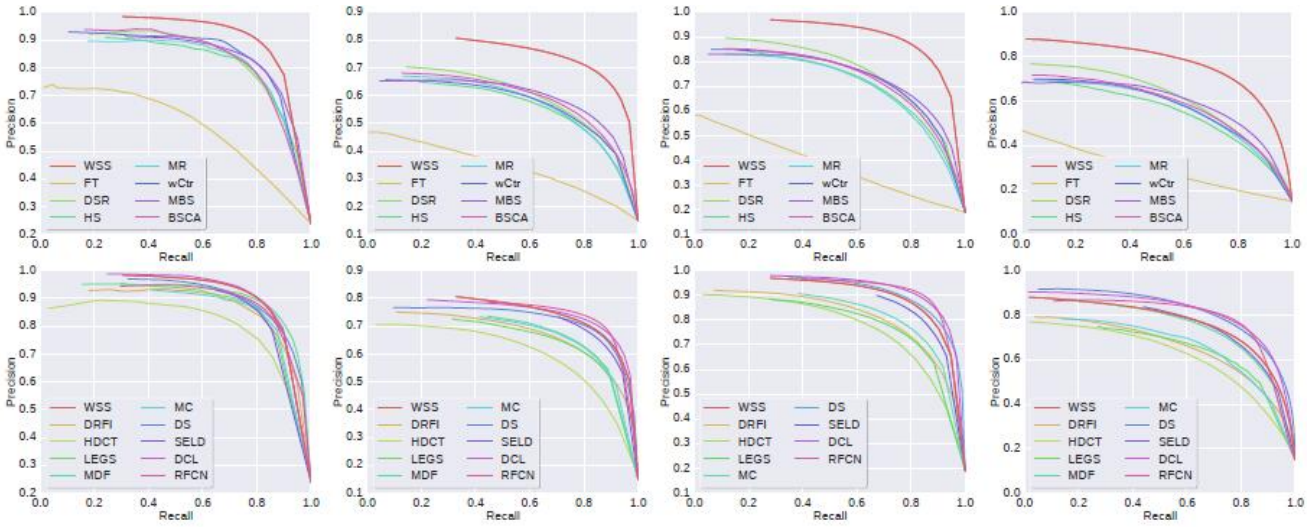

Figure 2. PR curves of unsupervised methods (first row) and fully supervised methods (second row). The proposed WSS significantly outperforms unsupervised methods and compares favorably against fully supervised methods

Table 1. The F_β measure of our method (WSS), the top 4 unsupervised methods, and top 7 fully supervised methods. All 7 supervised methods use DNNs supervised by pixel-level labels. The bold fonts denote the best methods in each setting. The speeds are in the last row.

	Unsupervised				Weakly	Fully						
	MR	wCtr	MBS	BSCA	WSS	LEGS	MDF	MC	DS	SELD	SELD	RFCN
ECSSD	0.690	0.676	0.673	0.705	0.823	0.785	0.807	0.796	0.826	0.810	0.829	0.834
SED	0.782	0.786	0.776	0.756	0.838	0.800	0.795	0.817	0.794	0.815	0.825	0.813
PASAL-S	0.583	0.597	0.604	0.597	0.720	¨C	0.705	0.687	0.655	0.714	0.710	0.747
THUS	0.542	0.528	0.547	0.536	0.663	0.607	0.636	0.610	0.626	0.634	0.657	0.694
HKU-IS	0.655	0.677	0.663	0.654	0.821	0.723	¨C	0.743	0.788	0.769	0.853	0.856
DUTS	0.510	0.506	0.511	0.500	0.657	0.585	0.673	0.594	0.632	0.628	0.714	0.712
FPS	6.71	6.76	76.9	0.67	62.5	0.52	0.04	0.44	8.33	1.80	2.18	0.60

2.1. Performance Comparison

Professor Wang compared the methods in each setting. Use WSS to compare with the two settings to provide a more comprehensive assessment. The PR curve in Figure 2 and the F_β in Table 1 all show that WSS is always superior to unsupervised methods and has a considerable margin, which is more advantageous than a fully supervised method. Meanwhile, WSS is also highly efficient with a real-time speed of 62.5 FPS, which is 8 times faster than supervised methods. Note that most of the saliency detection data sets contain huge amounts of objects not belonging to the 200 training categories. The superior performance of WSS confirms that WSS can well generalize to these unseen categories. Professor Wang also perform additional evaluations to verify the generalization ability of our method. We provide the quantitative and qualitative results on unseen categories, the MAE results, and the PR curves on PASCAL-S and ECSSD in the supplementary material due to limited space.

3. Conclusions

This paper proposes a two-stage training method based on image level weak supervision for saliency detection. In the first stage, two novel network designs, GSP and FIN, were proposed to estimate saliency maps by learning predictive image level category labels. In the second stage, the FIN is further fine-tuned using the estimated saliency map as a base fact. An iterative CRF was developed to improve the basic facts of the estimation and further improve performance. The extensive evaluation of the benchmark dataset validated the effectiveness of the method proposed by Professor Wang.

References

[1] A. Borji, M. Cheng, H. Jiang, and J. Li. Salient object detection: A survey. *arXiv preprint arXiv:1411.5878*, 2014.