Weakly-Supervised Segmentation by Combining CNN Feature Maps and Object Saliency Maps

Wataru Shimoda Keiji Yanai

The University of Electro-Communications, Tokyo, Japan

Objective

Weakly supervised segmentation

- Use only image-level annotation

Weakly supervised annotation

Person
horse
Car

Fully supervised annotation

Contributions

- Improved the method by Simonyan et al. [1]
- Combine BP-based map with CNN class maps for weakly supervised segmentation
- Achieved comparerable result in weakly-supervised segmentation with PASCAL VOC 2012

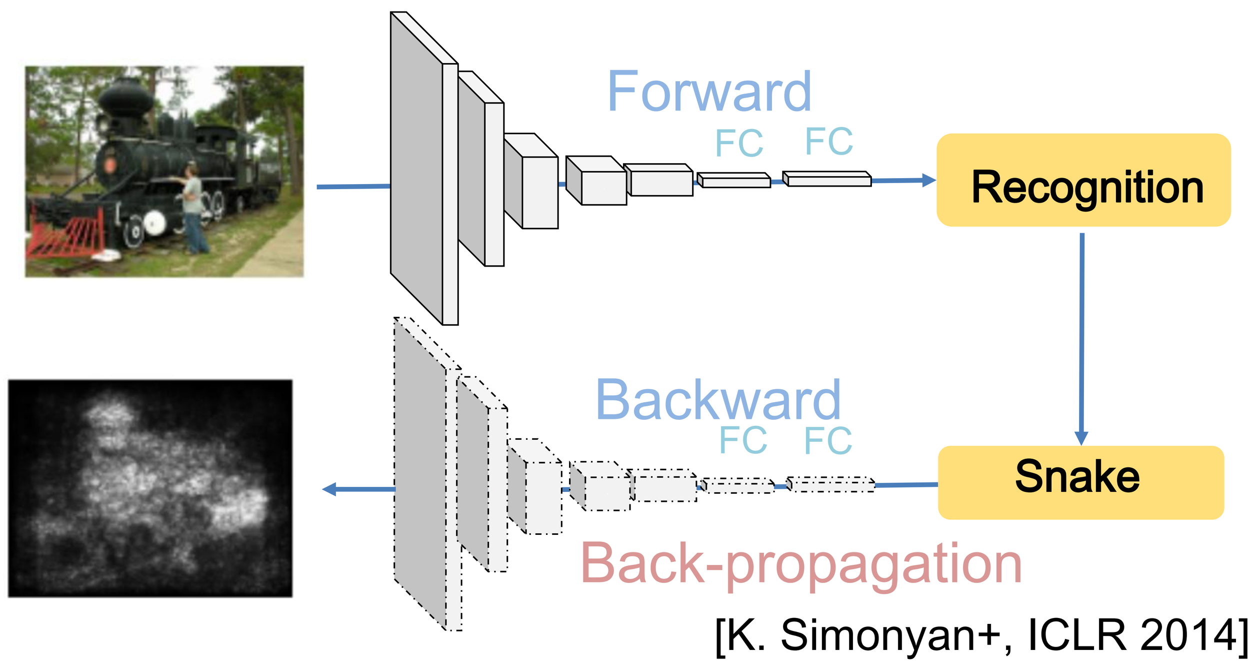
BP-based Visualization

Visualize class-specific saliency maps based on the derivatives of the class scores with respect to the input image

- proposed by K. Simonyan et al. at ICLR 2014 [1]
- Visualize contributed pixels on CNN classification
- Use derivatives obtained by back-propagation

The class score derivative vi of the i-th layer is the derivative of class score Sc with respect to the layer Li at the point (activation signal) Li

$$w_i^c = \frac{\partial S_c}{\partial w_i}\Big|_{I_0}$$



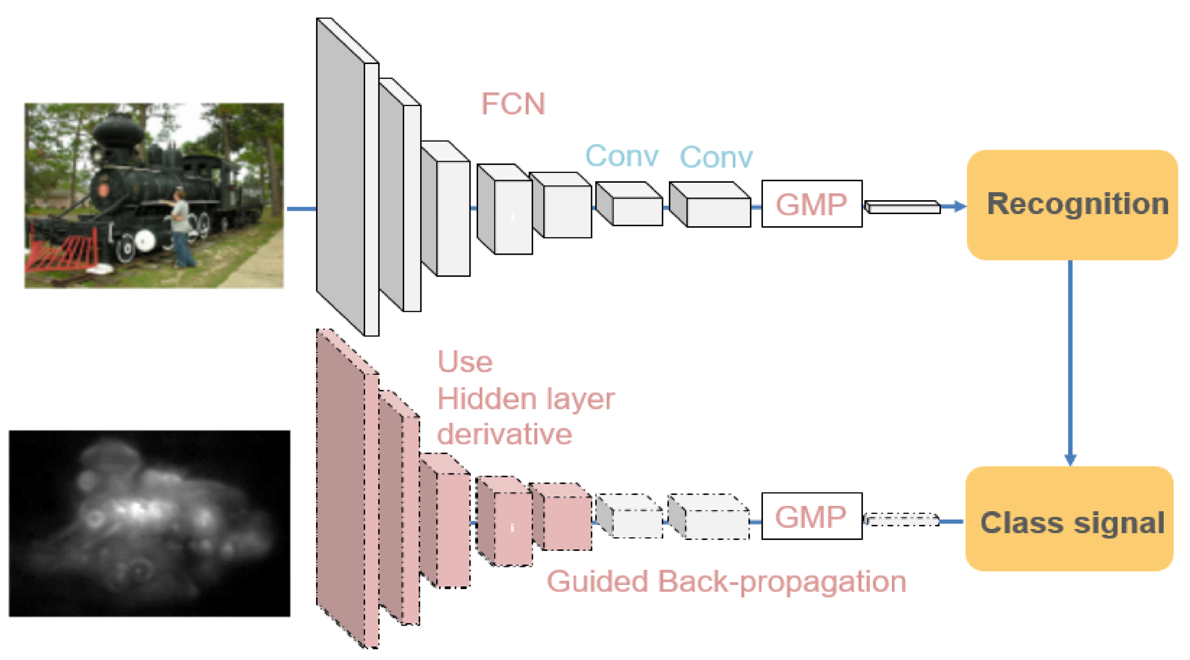
[K. Simonyan+, ICLR 2014]

White region means high derivative values which corresponds to the important pixels to enhance the given class score. (In the above fig. "Snake")

CNN Architecture

Improved points

- Fully Convolutional Net
- Guided back propagation [2]
- Use the derivatives of multiple intermediate layers



We back-propagate expected class scores generated by setting 1 for one of the top N-classes and 0 for the others. w_i^c represents up-sampled i-th layer derivative which is obtained by propagating class scores from the top layer.

.

Backward-base saliency maps

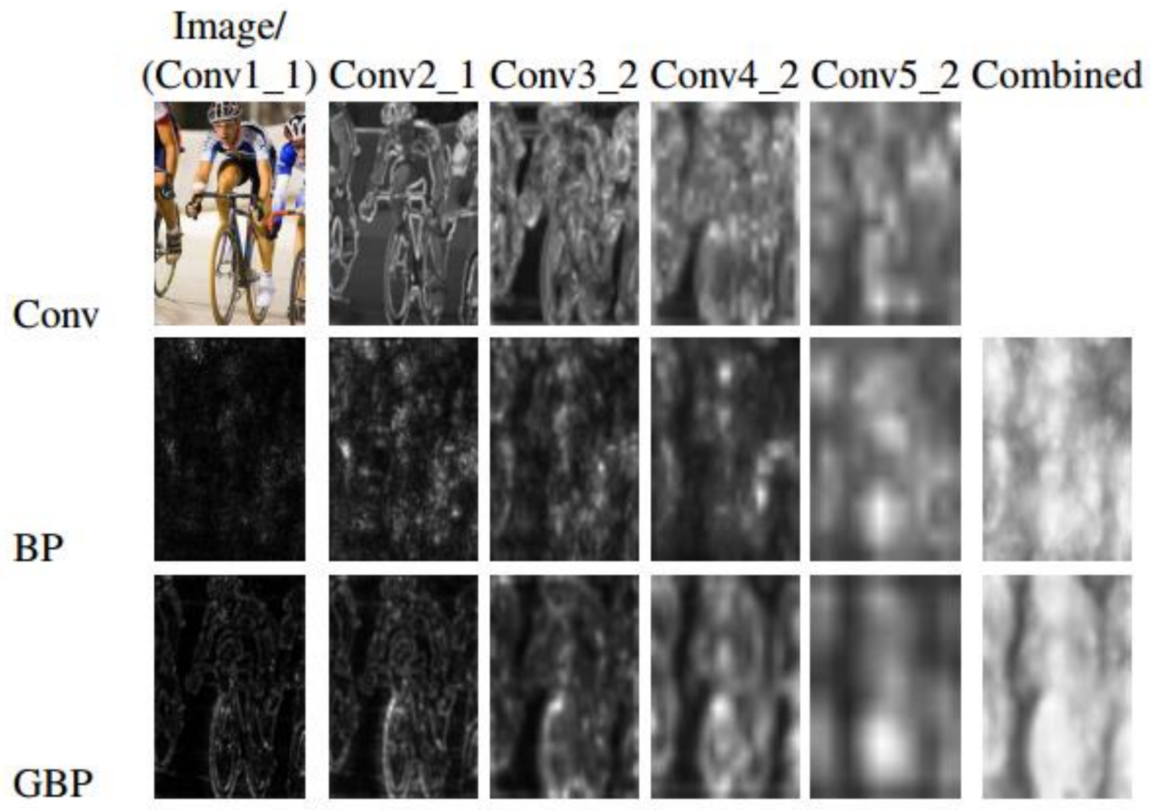
The class score derivative v_i^c of the i-th layer is the derivative of class score S_c with respect to the layer Li at the point (activation signal) $\mathrm{L_i}$

$$v_i^c = \frac{\partial S_c}{\partial L_i}\Big|_{\mathrm{L_i}}$$

Each class saliency maps $M_i^c \in \mathbb{R}^{m \times n}$ is calculated by:

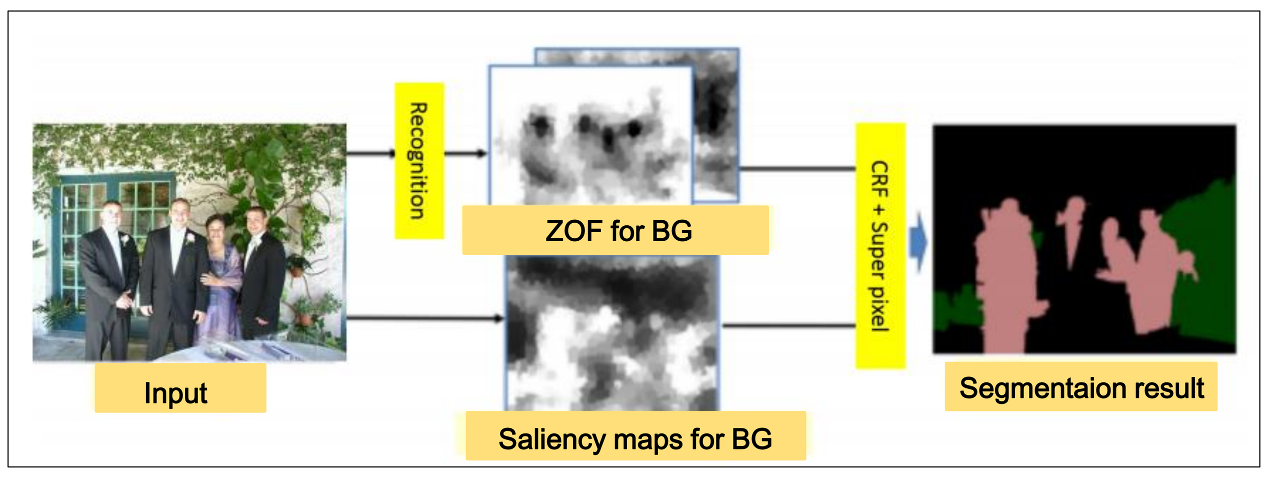
$$M_{i,x,y}^c = \max_{k_i} |w_{i,h_i(x,y,k)}^c| \quad ,$$

where $h_i(x, y, k)$ is the index of the element of w_i^c



Combine BP-based maps with featre maps

Zoom out feature



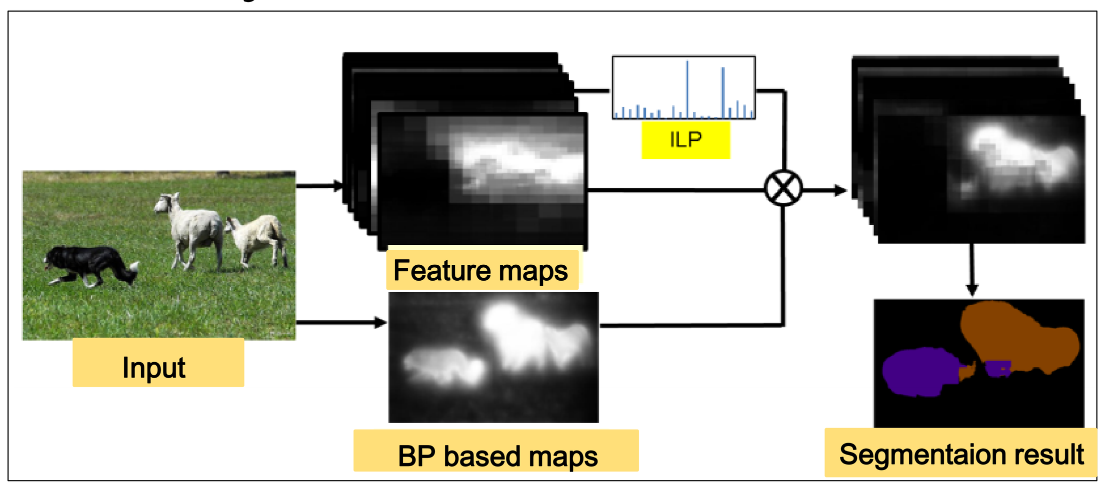
Steps

1. Recognize an image by forwarding
2. Back-propagation for each of the detected classes
3. Preparing Zoom-out feature by feature maps of eacy layer
4. Caluculate class probability maps by mi-SVM with feature maps
5. Unify the class maps and BP-based saliency maps by super-pixel-based CRF

Training

1. Training multi class recognition CNN
2. Training Mi-svm for CNN features

Fully convolutional network



Steps

1. Recognize an image by forwarding
2. Back-propagation for each of the detected classes
3. Obtaine class probability maps from output feature maps
4. Unify the class maps and BP-based saliency maps by ILP and thresholding

Training

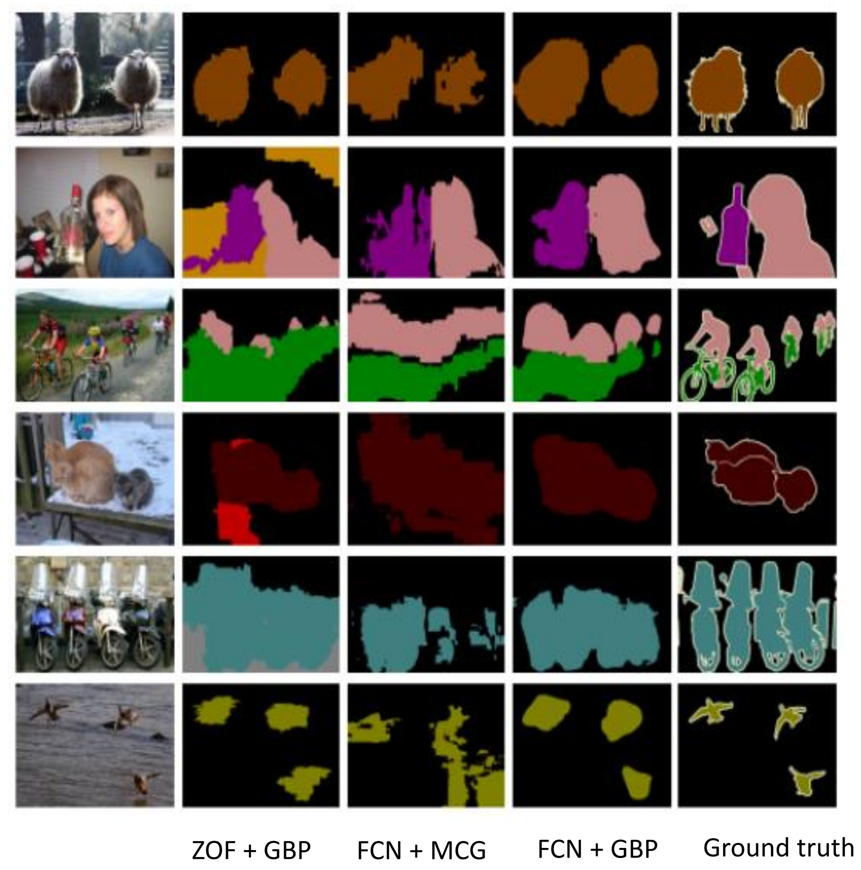
1. Training multi class recognition CNN

Experiments

Results on PASCAL VOC 2012

-A means using addtinal images for training

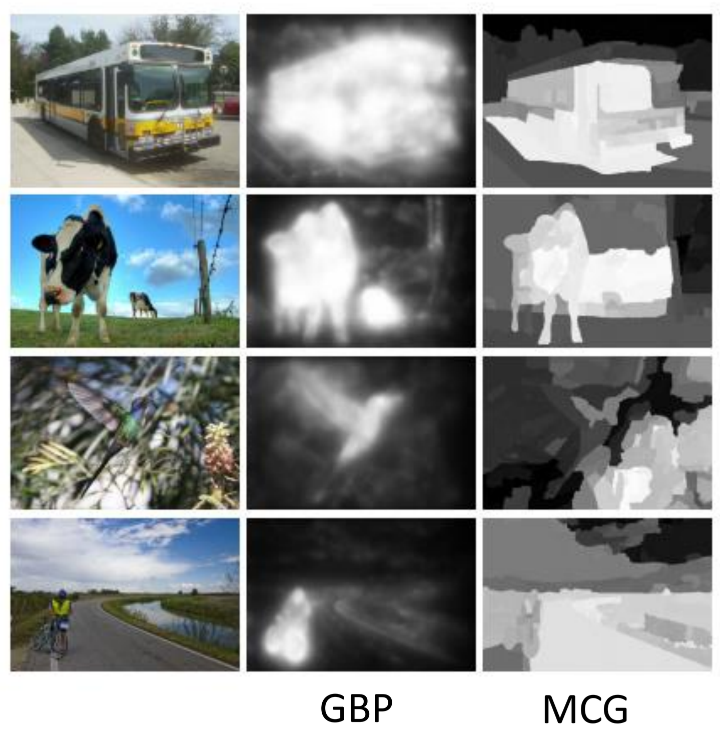
Method	A	Validation	test
MIL-FCN (ICLR 2015)	-	25.7	24.9
EM-Adapt(ICCV 2015)	-	38.2	39.6
CCNN (ICCV 2015)	-	34.5	35.5
MIL-seg (CVPR2015)	✓	42.0	40.6
STC (arXiv:1509.03150)	✓	49.8	51.2
Ours (ZOF+GBP)	-	38.1	37.7
Ours (FCN+GBP)	-	33.8	33.1
Ours (FCN+GBP)	-	41.4	40.7



Compare similar approach methods

- Training with global pooling
- Enhance FCN output with low-level objectness map

Method	Enhance	Mean IU
MIL-FCN	-	25.7
MIL-sppxl	Super-pixel refinement	36.6
MIL-bb	BB-proposal-based objectness map	37.8
MIL-seg	MCG-based objectness map	42.0
FCN + MCG	MCG-based objectness map	33.8
FCN + GBP	BP-based saliency map	41.4



References

[1] K. Simonyan et al. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. ICLR, 2014.

[2] J. Springenberg et al. Striving for Simplicity: The All Convolutional Net. ICLR, 2015.

[3] B. Hariharan et al. Semantic Contours from Inverse Detectors. ICCV, 2011.