DISTINCT CLASS SALIENCY MAPS FOR MULTIPLE OBJECT IMAGES

Wataru Shimoda & Keiji Yanai
Department of Informatics, The University of Electro-Communications, Tokyo
shimoda-k@mm.inf.uec.ac.jp, yanai@cs.uec.ac.jp

ABSTRACT

This paper proposes a method to obtain more distinct class saliency maps than Simonyan et al. (2014). We made three improvements over their method: (1) using CNN derivatives with respect to feature maps of the intermediate convolutional layers with up-sampling instead of an input image; (2) subtracting saliency maps of the other classes from saliency maps of the target class to differentiate target objects from other objects; (3) aggregating multi-scale class saliency maps to compensate lower resolution of the feature maps.

1 INTRODUCTION

Recently, a convolutional neural network (CNN) trained with only image-level annotation has been known to have the ability to localize trained objects in an image. Simonyan et al. (2014) proposed class saliency maps based on the gradient of the class score with respect to the input image, which showed weakly-supervised object localization could be done by back-propagation-based visualization. However, their class saliency maps are vague and not distinct (Fig.1(B)(C)). In addition, when different kinds of target objects are included in the image, the maps tend to respond to all the object regions. Although they adopted GrabCut for weakly-supervised segmentation based class saliency maps, their method is unable to distinguish multiple object regions (Fig.1(D)(E)).

To resolve the weaknesses of their method, we propose a new method to generate CNN-derivatives-based saliency maps. The proposed method can generate more distinct class saliency maps which discriminate the regions of a target class from the regions of the other classes (Fig.1(F)(G)). The generated maps are so distinct that they can be used as unary potentials of CRF directly (Fig.1(H)).

To make class saliency maps clearer, we propose three improvements: (1) using CNN derivatives with respect to feature maps of the intermediate convolutional layers with up-sampling instead of an input image; (2) subtracting saliency maps of the other classes from saliency maps of the target class to differentiate target objects from other objects; (3) aggregating multiple-scale class saliency maps to compensate lower resolution of the feature maps.

(A) sample image | Simonyan et al. (2014) (B) motorbike (C) person | GrabCut (D) motorbike (E) person | our saliency maps (F) motorbike (G) person | (H) CRF result

Figure 1: (From the left) A sample image, its class saliency maps with respect to "motorbike" and "person" by Simonyan et al. (2014), estimated regions of them by GrabCut, class saliency maps by the proposed method, and estimated regions by CRF.

2 PROPOSED METHOD

The proposed method consists of (1) extracting CNN derivatives with respect to feature maps of the intermediate convolutional layers, (2) subtracting class saliency maps between the target class and the other classes, and (3) aggregation of multi-scale saliency maps.

In Simonyan et al. (2014), they regarded the derivatives of the class score with respect to the input image as class saliency maps. However, the position of an input image is the furthermost from the class score output on the deep CNN, which sometime causes weakening or vanishing of gradients.

Instead of the derivatives of the class score with respect to the input image, we use the derivatives with respect to feature maps of the relatively upper intermediate layers which are expected to retain more high-level semantic information. We select the maximum absolute values of the derivatives with respect to the feature maps at each location of feature maps across all the kernels, and up-sample them with bilinear interpolation so that their size becomes the same as an input image (Fig.2 (C)-(G)). Finally we average them to obtain one saliency map (Fig.2 (B)). The idea on aggregating of information extracted from multiple feature layers was inspired by the work of Long et al. (2015).

The class score derivative v_i^c of a feature map is the derivative of class score Sc with respect to the layer L_i at the point (activation signal) L_i^0:

$$v_i^c = \left. \frac{\partial S_c}{\partial L_i} \right|_{L_i^0} \tag{1}$$

v_i^c can be computed by back-propagation. After obtained v_i^c, we up-sample it to w_i^c with bilinear interpolation so that the size of an 2-D map of v_i^c becomes the same as an input image. Next, the class saliency map $M_i^c \in \mathcal{R}^{m \times n}$ is computed as $M_{i,x,y}^c = \max_{k_i} |w_{i,h_i(x,y,k)}^c|$, where $h_i(x, y, k)$ is the index of the element of w_i^c. Note that each value of the saliency map is normalized by $\tanh(\alpha M_{i,x,y} / \max_{x,y} M_{i,x,y})$ for visualization in Fig.2 and all the other figures with $alpha = 3$.

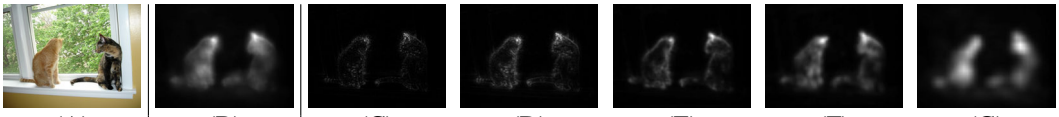

Figure 2: Class saliency maps obtained from the VGG16-net fine-tuned with the PASCAL VOC 2012 dataset. (A) an input image, (B) average of [(E)(F)(G)], (C) conv1_1, (D) conv2_1, (E) conv3_2, (F) conv4_2, (G) conv5_2

As shown in Fig.1(B)(C), the saliency maps of two or more different classes tend to be similar to each other especially in the image-level. The saliency maps by Simonyan et al. (2014) is likely to correspond to foreground regions rather than specific class regions. This problem is relaxed in the proposed methods, because we use saliency maps obtained from intermediate layers. However, the saliency regions of different classes are still overlapped with each other (Fig.3 (raw)).

To resolve that, we subtract saliency maps of the other candidate classes from saliency maps of the target class to differentiate target objects from other objects. Here, we assume that we use CNN trained with multi-label loss and select several candidate classes with a pre-defined threshold. The improved class saliency maps with respect to class c, $\tilde{M}_i^c$, is represented as:

$$\tilde{M}_{i,x,y}^c = \sum_{c' \in candidates} \max\left(M_{i,x,y}^c - M_{i,x,y}^{c'}, 0\right) [c \neq c'], \tag{2}$$

where $candidates$ is a set of the selected candidate classes. Fig.3 shows results without subtraction in the left (raw) and ones with subtraction in the right (diff).

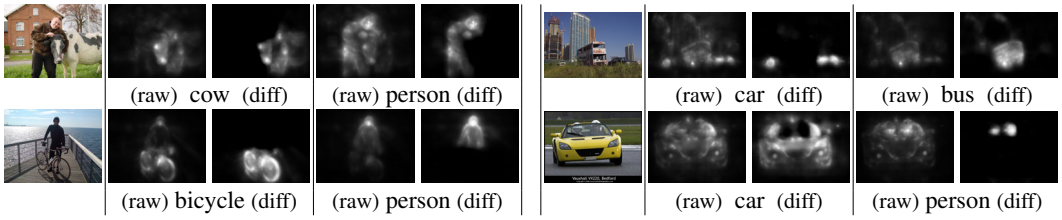

Figure 3: (Left) without subtraction of other class maps (Right) with subtraction of other class maps.

Recently, fully convolutional networks (FCN) which accept arbitrary-sized inputs are used commonly in works on CNN-based detection and segmentation such as Oquab et al. (2015) and Long et al. (2015), in which fully connected layers with n units were replaced with the equivalent convolutional layers having n 1×1 filters. We introduce FCN into multi-scale generation of class saliency maps. If the larger input image than one for the original CNN is given to the fully-convolutionalized CNN, the output becomes class score maps represented as $h \times w \times C$ where C is the number of classes, and h and w are larger than 1. To obtain CNN derivatives with respect

to enlarged feature maps, we simply back-propagate the target class score map which is define as $S_c(:,:,c) = 1$ (in the Matlab notation) with 0 for all the other elements, where c is the target class index.

Finally, the class specific saliency map M^c is obtained as follows:

$$M^c_{x,y} = \frac{1}{|S||L|} \sum_{j \in S} \sum_{l \in L} \tanh(\alpha M^c_{i,x,y}), \tag{3}$$

where L is a set of the layers for which saliency maps are extracted, S is a set of the scale ratios, and α is a constant which we set to 3 in the experiments. Note that we assume the size of M_i for all the layers are normalized to the same size as an input image before calculation of Eq.3.

In the experiments, we adopted guided back-propagation (GBP) (Springenberg et al. (2015)) as back-propagation method instead of normal back-propagation (BP) used in Simonyan et al. (2014). The difference between two methods is in the backward computation through ReLU. GBP can visualize saliency maps with less noise components than normal BP by back-propagating only positive values of CNN derivatives through ReLU as shown in Fig.4.

Figure 4: Obtained class saliency maps (Left) using BP (Right) using GBP.

3 Experiments and Discussion

We tested the proposed method with the PASCAL VOC 2012 dataset. We used 16-layered CNN, VGG-16 (Simonyan et al. (2015)) pre-trained with ImageNet 1000 categories, and fine-tuned it using PASCAL VOC augmented training dataset including image-level multi-label annotation of 20 classes provided by Hariharan et al. (2011) with Sigmoid entropy loss for multi-label training, random resizing of training images and global max pooling for multi-scale training following Oquab et al. (2015).

For obtaining final saliency maps with Eq.3, we extracted CNN derivatives from Conv3_1 to Conv5_3 of VGG-16 (totally 9 layers) with scales $\{1, 1.2, 1.4, 1.6\}$. Fig.5 shows both the results by Simonyan et al. (2014) and our method for three multiple object images and one single object images. Note that we did not use subtraction of saliency maps in case of only a single object image. From these results, it is shown that our method is much more effective for not only multiple object images but also single object images than the previous method. Fig.6 shows the results for images containing three or more objects. In even such cases, all the class saliency maps except for "chair" in the top-right sample were estimated successfully.

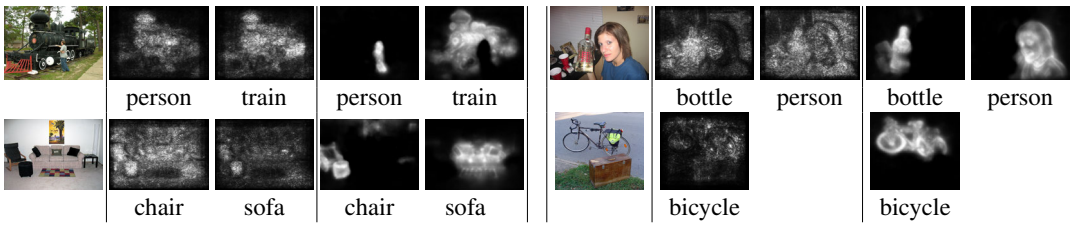

Figure 5: Obtained class saliency maps (Left) by Simonyan et al. (Right) by the proposed method.

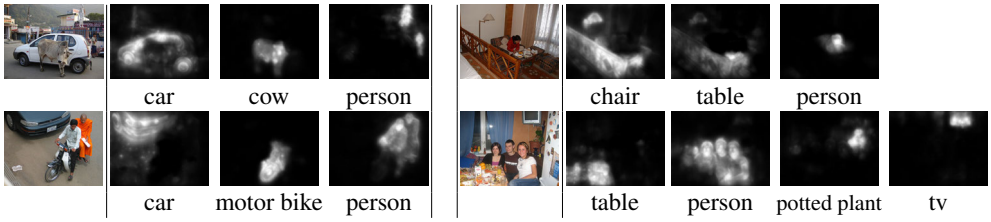

Figure 6: Class saliency maps by the proposed method for images containing three or more classes.

REFERENCES

B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and Malik. J. Semantic contours from inverse detectors. In *Proc. of IEEE International Conference on Computer Vision*, 2011.

J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? -weakly-supervised learning with convolutional neural networks. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proc. of International Conference on Learning Representations, Workshop Track*, 2014.

K. Simonyan, A. Vedaldi, and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of International Conference on Learning Representations*, 2015.

J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *Proc. of International Conference on Learning Representations*, 2015.