Box-driven Class-wise Region Masking and Filling Rate Guided Loss for Weakly Supervised Semantic Segmentation

(2019)

Song, Chunfeng Huang, Yan Ouyang, Wanli Wang, Liang

Notes

1 Introduction

Bounding are capable of giving us a lot of information about the postition and placement of different objects in the image, but the main problem is that the majority of the pixels in a given bouncing box might belong to the background depending on the class, so the authors propose to: (1) learn class specific masks, that are then applied to each bounding box to get the segmentattion predictions, (2) and given that even after masking we'll still endup with a lot of false positive and negatives, they propose to calculate the segmentation loss only on the top most confident predictions, the rate depend on the class in question given that each calss have different filling rates (number of pixel that belongs to the foreground over the background pixels in a given bounding box) for different poses, this will help the model avoid having a leraning signal in the pixel that are wrongly annotated usign the bounding boxes.

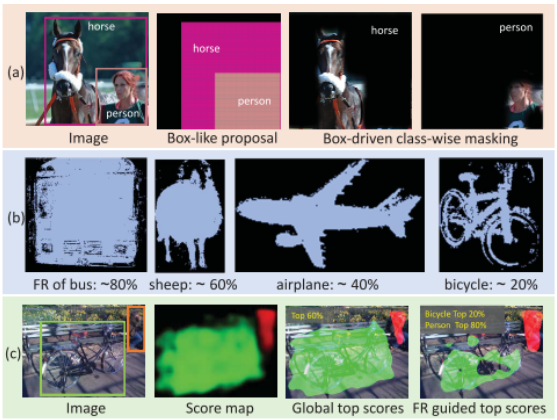

Figure 1. Weakly supervised segmentation with the box-level annotations. (a) The box-driven class-wise masking (BCM) model can learn specific masks for each class in region-level, and help remove the irrelevant regions of each class softly. (b) Based on the pixel-level segment proposals and the bounding boxes, we could calculate the mean pixel filling rates of each class, e.g., the sheep fills roughly 60% pixels of the box. (c) Via ranking the values of the score map, we can select the most confident locations for back propagation and ignore the weak ones. As shown in the picture, filling rate guided top scores selection is better than the global one.

2 Method

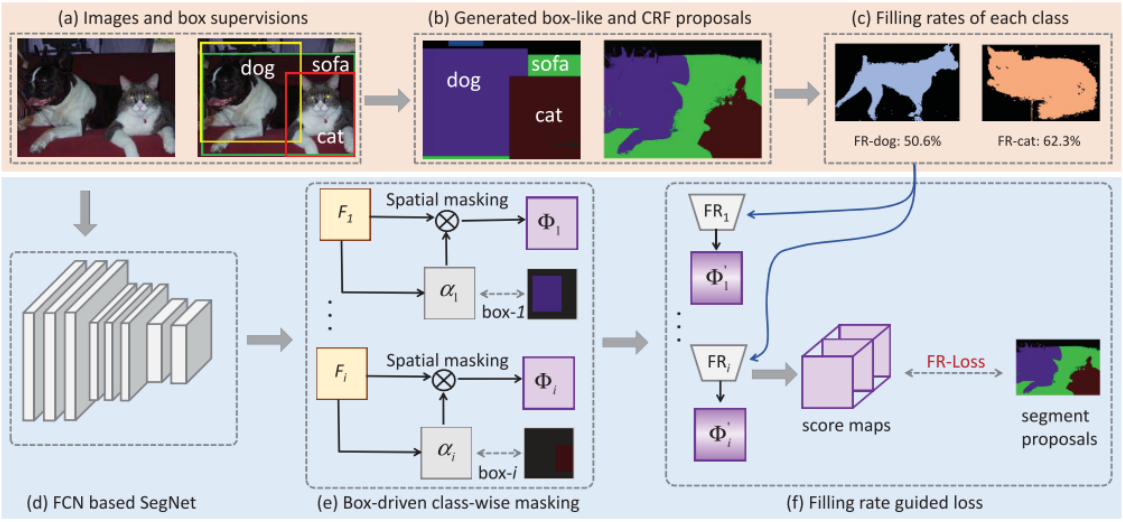

Figure 2. Pipeline of the proposed method. For a given image and its corresponding bounding boxes (a), we first generate the rectangle annotations (Box-like) and apply the unsupervised CRF [22] to generate segment proposals (b). We then calculate the mean filling rates of each class (c) with the CRF proposals and their corresponding boxes. With the images and segment proposals, we train the FCN based model (d), e.g., the DeepLab-LargeFOV network [5]. We add a box-driven class-wise masking (BCM) model (e) to generate class-aware masks via segmentation learning with box-like labels. The learned masks can implement spatial masking on the features of each class, separately. For each forward step, we rank the scores of each class in the prediction layer and adopt the filling rate guided loss (FR-loss) (f) to select the most confident locations for back propagation and ignore the weak ones. FR-loss could reduce the negative effects caused by the wrongly labeled pixels in the proposals.

Generating pseudo labels

First step is to create pseudo labels in the form of segmentation masks using the bounding box annotations, this is done by assigning all the pixel in a given box to the class of the bounding box, if we have some intersection we assign the pixel to the class of the smaller box, and them refine these masks using a fully connected CRF.

Box-driven Class-wise Masking

Now to train our segmentation network, we train attention maps to be binary class specific masks, these binary masks are trained using an MSE loss between the attention weights and the bounding boxes in the ground through (M_c), given that the attention weights inside the bounding box must be equal to one and outside it be equal to zero, the size of these attention maps (α) is the same as the size of the output predictions and are class spacific, the loss is as follows:

$$L_{bcm(c)} = \sum_{h=1}^{H} \sum_{w=1}^{W} \left\| M_{c(h,w)} - \alpha_{c(h,w)} \right\|_2^2$$

And these attention weights are applied to ourput feature afterwards, in an element wise manner:

$$\Phi_c = F_c \otimes \alpha_c$$

Filling Rate Guided Adaptive Loss

Given that only a good number of pixel in the pseudo segmentation masks generated from the bounding boxes are not correct, we need to only backpropagate the loss for a limited number of pixels that we are confident to belong to the given class, this is done using filing rates, for each class we calculate a number of filing rates per orientation (n = 3), so for each class we'll have 3 percentages of filling (say 20%, 15% and 35%), and then we can calculate the loss only for these top classses, so in our example the loss will be calculated for the top 20%, 15% and 35% most confident

pixels and summed, so first we get the filling rates for each class using the ground thruth box and the propsal of the segmentation netwrok (pixels that belong to the foreground):

$$FR_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{P_{proposal}(i)}{P_{box}(i)}$$

And then we calculate the losses for the top3 per class, the top 3 filling rates are generated usign k-means over all the filling rates of the classes in different poses, that are clustered into three clusters:

$$L_{fr} = \sum_{c=1}^{N} \sum_{sc}^{3} \sum_{i=1}^{top\left(FR_{(c,sc)}\right)} L_{(c,sc)}(i)$$

3 Experiments

They use Deeplab with segnet network, based on VGG16 and pretrained on ImageNet, they first train for 20k iteration without the FR loss and the BCM, and then fine tune for 5k more iterations

Modes	# GT	# Box	Methods	mIoU
Weak	-	10,582	BoxSup_{Box} [9]	52.3
			WSSL_{Box} [29]	52.5
			SDI_{Box} [21]	61.2
			Ours$_{Box}$	54.9
			BoxSup_{MCG} [9]	62.0
			WSSL_{CRF} [29]	60.6
			SDI_{M+G} [21]	65.7
			Ours$_{CRF}$	**66.8**
Semi	1,464	9,118	WSSL_{Box} [29]	62.1
			BoxSup_{MCG} [9]	63.5
			WSSL_{CRF}	65.1
			SDI_{M+G} [21]	65.8
			Ours$_{CRF}$	**67.5**
Full	10,582	-	DeepLab-LargeFOV [5]	69.8

Table 2. Weakly and Semi-supervised results on VOC2012 validation set. With only 1/10 labeled segments, our method can achieve comparable performance with the fully supervised model. Box: directly using rectangle proposals, M+G: using the combined labels with both MCG and GrabCut.

Modes	# GT	# Box	Methods	mIoU
Weak	-	10,582	SDI [21]	69.4
			Ours	**70.2**
Semi	1,464	9,118	**Ours**	**71.6**
Full	10,582	-	DeepLab-ResNet-101 [5]	74.5

Table 3. Results of ResNet-101 backbone on VOC2012 validation set. Our method outperforms the compared SDI [21] method, achieving comparable performance with the fully supervised one.

Methods	Units	mIoU
Baseline [29]	-	60.6
Ours	CM	63.4
	BGM	64.9
	BCM	65.6
	Global-loss	64.1
	FR-loss	65.8
	FR-loss(Refine)	66.3
	BCM + FR-loss(Refine)	**66.8**

Table 1. Evaluate the effectivenesses of BCM and FR-loss on VOC2012 validation set. All models are based on the same Deeplab VGG16-LargeFOV backbones. The performance is evaluated in terms of mean IoU (%). CM: class-wise masking without box supervision, BGM: box-driven global masking, Global-loss: all boxes adopt the same global filling rate of 0.6.