Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing

(2018)

Huang, Zilong Wang, Xinggang Wang, Jiasi Liu, Wenyu Wang, Jingdong

Notes

1 Introduction

In weakly supervised segmentation, it is quite hard to generate pseudo labels for segmentation using only image level labels, given that the localization maps produced by methods such as CAN using the classification scores are not very precise and are only over the descriminative parts of the image, the authors propose to use seeded region growing, that start from a very limited number of labeled pixel in the image, and using some similarity measure like the RGB, intensity of texture values, will grow the segmentation region.

The seed they propose to use will be generated by the classification scores using CAN, and then these seeds will be grown using the output of the segmentation and a given similarity measure, and the loss will only be calculated for the labeled pixels, this method is an online method, the region keep growing with more training given that the fetures used in the similarity become more and more representative, and this give very accuracte masks at the end.

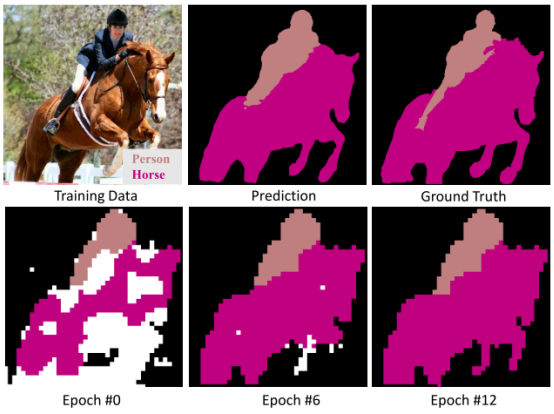

Figure 1. The top row orderly shows a training image with the image-level labels, the segmentation result of our proposed method only using image-level supervision, and the ground truth. Our segmentation result is very close to the ground truth annotated by human. The bottom row shows the dynamic supervision in several epochs during the training of the proposed weakly-supervised semantic segmentation network. (The black represents background and the white represents unlabeled/ignore pixels).

2 Method

The first step is to generate the seed, this is done using the classification scores with CAM, the classification scores are applied to the conv7 to generate heatmaps for each object class, then the seed region are obtain after a hard threshold, using silency map, all the pixel with low silency are considered as background.

Now after having the localization maps, we use them as seed region, the regions are then grown from these seed points to adjacent unlabeled points depending on a region similarity criterion, the similarity criterion used on the output segmentation probabilities, in a region with class C, if the adjacent pixel probability to belong to the class C is maximal and above a threshold, we assign to it label C:

$$P\left(H_{u,c}, \theta_c\right) = \begin{cases} \text{TRUE } H_{u,c} \geq \theta_c \text{ and } c = \arg\max_{c'} H_{u,c'} \\ \text{FALSE ohterwise} \end{cases}$$

And this is done in an iterative way, for each pixel the 8 connectivity neighboors are visited, note that there is two threshlods, one for the background and one for the foreground:

1: if $P\left(H_{u,c}, \theta_c\right)$ then
2: the pixel at u is labeled as c;
3 : else
4 : the pixel at u keeps unlabeled state.
5 : end if

Seeding loss now for the loss, given that some pixel will not be assigned to any class, we need to calculate the loss to match only the seeded cures given the classification network, and considering the unbalenced distribution of pixels, we normalize of foreground for each class and the background seperately:

$$\ell_{\text{seed}} = -\frac{1}{\sum_{c\in\mathcal{C}} |S_c|} \sum_{c\in\mathcal{C}} \sum_{u\in S_c} \log H_{u,c} - \frac{1}{\sum_{c\in\overline{\mathcal{C}}} |S_c|} \sum_{c\in\overline{\mathcal{C}}} \sum_{u\in S_c} \log H_{u,c}$$

and finnaly the total loss is two folds, the seeding loss and te boundary loss that uses a CRF to encourage segmentation map to math up with object boudaries:

$$\ell = \ell_{\text{seed}} + \ell_{\text{boundary}}$$

3 Experiments

They use VGG16, SGD with 0.0005 weight decay and 0.9 momentum and a batch of 20, the initial learning rate is 5e-4 and is divided by 10 each 2000 iterations, the 20% heat map values are considered as forground and the saliency threshold in 0.06 to obtain the background pixels, for the similarity threshold, foreground is 0.85 and the background is 0.99.

Table 2. Comparison of mIoU using different settings of our approach on VOC 2012 val set

Method	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
baseline	82.5	67.5	23.2	65.7	29.7	47.5	71.8	66.8	76.7	23.3	51.7	26.2	69.7	54.2	63.2	57.2	33.7	64.5	33.5	48.7	46.1	52.5
+BSL	82.4	71.9	29.1	67.7	32.4	49.8	75.5	67.9	74.7	22.8	54.9	26.6	64.3	55.7	64.7	56.0	35.0	67.7	32.7	50.2	45.8	53.6
+DSRG	86.6	70.5	28.8	70.6	34.7	55.7	74.9	70.1	80.2	24.1	63.6	24.8	76.6	64.1	64.9	72.3	38.5	68.7	35.8	51.8	51.9	57.6
+Retrain	87.5	73.1	28.4	75.4	39.5	54.5	78.2	71.3	80.6	25.0	63.3	25.4	77.8	65.4	65.2	72.8	41.2	74.3	34.1	52.1	53.0	59.0

Table 1. Comparison of weakly-supervised semantic segmentation methods on VOC 2012 val and test set

Method	**Training**	**Val**	**Test**
Supervision: Image-level Labels			
(* methods implicitly use pixel-level supervision)			
(† methods implicitly use box supervision)			
SN_B* [34]	10k	41.9	40.6
MIL-seg* [23]	700k	42.0	43.2
TransferNet* [10]	70k	52.1	51.2
AF-MCG* [24]	10k	54.3	55.5
GuidedSeg† [19]	20k	55.7	56.7
Supervision: Image-level Labels			
MIL-FCN [22]	10k	25.7	24.9
CCNN [21]	700k	35.3	35.6
MIL-bb [23]	700k	37.8	37.0
EM-Adapt [20]	10k	38.2	39.6
DCSM [29]	10k	44.1	45.1
BFBP [27]	10k	46.6	48.0
STC [35]	50k	49.8	51.2
SEC [14]	10k	50.7	51.7
AF-SS [24]	10k	52.6	52.7
Combining Cues [26]	10k	52.8	53.7
AE-PSL [33]	10k	55.0	55.7
DCSP [4]	10k	58.6	59.2
Supervision: Image-level Labels			
DSRG (VGG16)	10k	**59.0**	**60.4**[1]
DSRG (Resnet101)	10k	**61.4**	**63.2**[2]