Note: Anti-Adversarially Manipulated Attributions for Weakly and Semi-Supervised Semantic Segmentation

sgc

May 22, 2022

1 Introduction

This paper proposed a new manipulating method to extending the discriminative regions of a target object. This method is based on adversarial attack, but in a anti-adversarial manner that an image is perturbed along pixel gradients to increase the classification score, then a larger cover of the object area is generated.
And in case the irrelevant pixels(like background) be wrongly classified, authors apply regularization term to suppress the score of other classes.
Notably this method is post-hoc and can be used directly to improve the performance without modification, resulting in new SOTA performance.
Main contributions:
1, This paper proposed AdvCAM to identify more regions of a object.
2, This method can be implemented to other model without any modification.
3, Better performance than existing methods in both weakly supervised semantic segmentation and Semi-supervised semantic segmentation.

2 Method

2.1 Adversarial Climbing

Anti-adversarial attack is used to increase the score of classification:

$$x^t = x^{t-1} + \xi \nabla_{x^{t-1}} y_c^{t-1}$$

where t is the step index, x^t is the manipulated image, and y is the classification logit for class c.
And authors produce a localization map through:

$$\mathcal{A} = \frac{\sum_{t=0}^{T} CAM(x^t)}{max \sum_{t=0}^{T} CAM(x^t)}$$

2.2 How can Adversarial Climbing Improve CAMs?

2.2.1 Can non-discriminative features be enhanced?

Actually both discriminative regions and non-discriminative regions grow, but but enhances non-discriminative features more than discriminative ones, resulting in a denser CAM.

2.2.2 Are those enhanced features class-relevant from a human point of view?

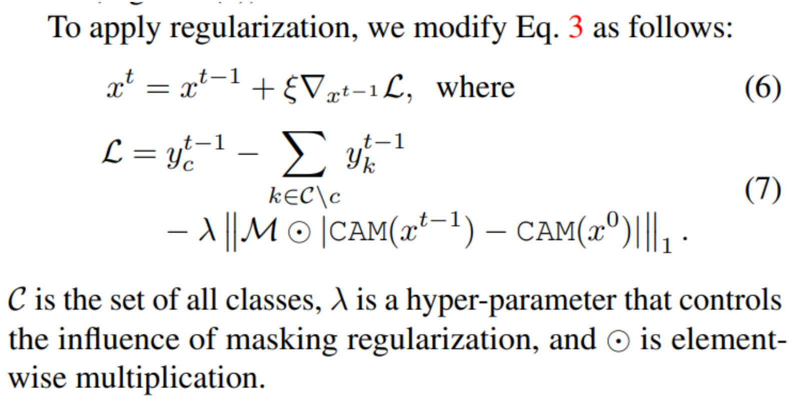
The loss landscape obtained by adversarial climbing is much more flatten than that obtained by adversarial attacking. Therefore, This method increases the attribution of features relevant to the class from a human point of view, resulting in a better CAM.

2.3 Regularization

Without regularization, adversarial climbing may make regions of irrelevant class to grow or increase the attribution scores of the regions that already have high scores. This paper address this by:
(i) suppressing the logit values associated with other classes
(ii) restricting high attributions on discriminative regions of the target object.
The second reason actually matters even though it may be hard to understand: Mainly for 2 reasons: 1) it prevents new regions from being additionally attributed to the classification score, and 2) if the maximum value of the attribution score increases during adversarial climbing, the normalized scores of the remaining area may decrease.
To solve that, this paper uses a restricting mask:

$$\mathcal{M} = \nVdash(CAM(x^{t-1}) > \tau)$$

As aforementioned, adversarial climbing enhance the non-discriminative features more than discriminative features but regularization makes this even more.



To apply regularization, we modify Eq. 3 as follows:

$$x^t = x^{t-1} + \xi \nabla_{x^{t-1}} \mathcal{L}, \quad \text{where} \tag{6}$$

$$\mathcal{L} = y_c^{t-1} - \sum_{k \in \mathcal{C} \setminus c} y_k^{t-1} - \lambda \left\| \mathcal{M} \odot \left| \texttt{CAM}(x^{t-1}) - \texttt{CAM}(x^0) \right| \right\|_1 . \tag{7}$$

$\mathcal{C}$ is the set of all classes, λ is a hyper-parameter that controls the influence of masking regularization, and $\odot$ is element-wise multiplication.

At last, authors also used seed refinement to improve the performance.

3 Results

For the quality of the mask, whichever refinement method is used, AdvCAM outperformed other methods by a large margin. For WSSS and SSSS, this method outperforms the existing(2021) method and other methods using auxiliary salient object mask supervision.

4 Ablation Study

The graph followed shows the results:

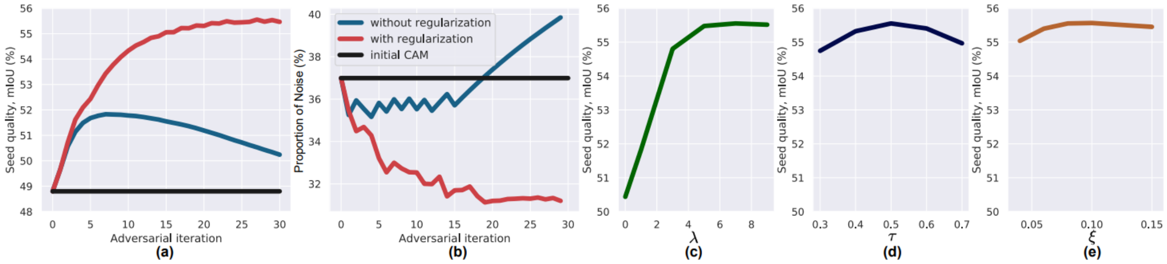


Figure 6: Effect of adversarial climbing and regularization on (a) the seed quality and (b) the proportion of noise. (c) Effect of the regularization coefficient λ. (d) Effect of the masking threshold τ. (d) Effect of the step size ξ.

Remarkably, τ is less sensitive than λ: varying τ between 0.3 and 0.7 produces less than 1% change in mIoU. And changes in step size ξ are not particularly significant.

5 Summary

This paper uses the adversarial manipulating to obtain a better localization of the target object. Notably this method can be implemented without any modification. When add AdvCAM to existing SOTA method, it achieved new SOTA performance on both WSSS and SSSS.

References