Note: Cross Language Image Matching for Weakly Supervised Semantic Segmentation[1]

sgc

April 23, 2022

1 Introduction

This paper proposed a framework CLIMS(Cross Language Image Matching) for WSSS. THe core idea is that: introduce natural language supervision to activate more complete object regions and suppress closely-related open background.

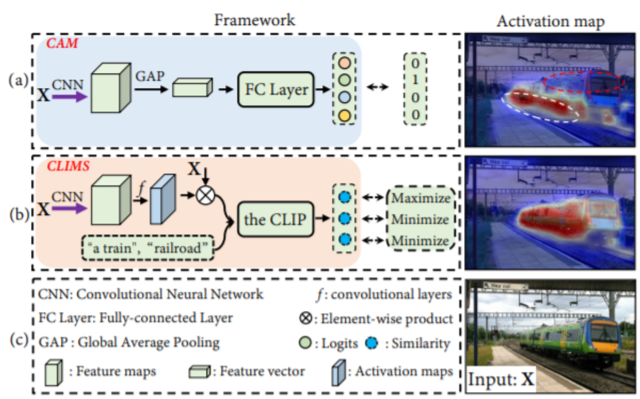

Figure 1. (a) Conventional CAM solution. (b) The proposed CLIMS. The problem of false-activation of irrelevant background, e.g., railroad and ground, and underestimation of object contents usually exist in conventional CAM method. To solve this problem, we propose a novel text-driven learning framework, CLIMS, which introduces natural language supervision, i.e., an open-world setting, for exploring complete object contents and excluding irrelevant background regions. Best viewed in color.

The ClIMS consists of 2 mian parts: a backbone and an evaluator based on 3 Losses(i.e. Object region and Text label Matching loss (L_{OTM} to maximize), Background region and Text label Matching loss (L_{BTM} to minimize), and Co-occurring Background Suppression loss (L_{CBS} to minimize).
Main contribution:
1, Proposed CLIMS to utilize image-text information to WSSS tasks. And it

outperformed previous SOTA.
2, Designed 3 useful loss.

2 Method

2.1 Cross Language Image Matching Framework

The CAMS is generated similar to conventional CAM by deplete the GAP and take use of sigmiod:

$$P_k(h, w) = \sigma(W_k^T Z(h, w))$$

(K denotes the number of classes.)
The text image prompt label is showed in the picture. They share the structure : The photo of . But the L co-occurring L backgrounds are manually pre-defined(shortcoming). And $X\dot{P}_k$ means masking out the foreground object.

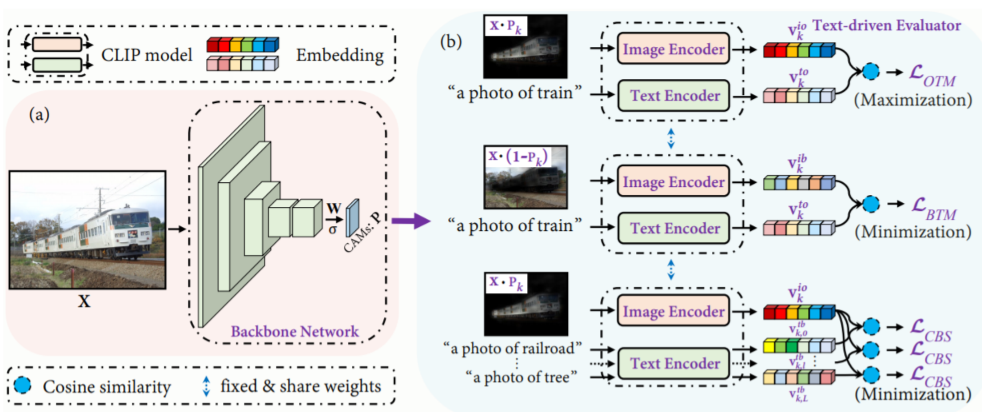

Figure 2. An overview of the proposed Cross Language Image Matching framework for WSSS, i.e., CLIMS. (a) The backbone network for predicting initial CAMs. σ denotes the sigmoid activation function. $\mathbf{W}$ denotes the weight matrix of convolutional layers. (b) The text-driven evaluator. It consists of three CLIP-based loss functions, i.e., object region and text label matching loss $\mathcal{L}_{OTM}$, background region and text label matching loss $\mathcal{L}_{BTM}$, and co-occurring background suppression loss $\mathcal{L}_{CBS}$. Best viewed in color.

2.2 Object region and Text label Matching

L_{OTM}:

$$\mathcal{L}_{OTM} = -\sum_{k=1}^{K} y_k \cdot \log(s_k^{oo}),$$

$$s_k^{oo} = \text{sim}(\mathbf{v}_k^{io}, \mathbf{v}_k^{to}),$$

2.3 Background region and Text label Matching

L_{BTM}:

$$\mathcal{L}_{BTM} = -\sum_{k=1}^{K} y_k \cdot \log(1 - s_k^{bo}),$$

$$s_k^{bo} = \text{sim}(\mathbf{v}_k^{ib}, \mathbf{v}_k^{to}),$$

When L_{BTM} is minimized, fewer target object pixels are reserved in $X\dot{(}1 - P_k)$ and more target object contents are recovered in $X\dot{P}_k$.

2.4 Co-occurring Background Suppression

L_{CBS}:

$$\mathcal{L}_{CBS} = -\sum_{k=1}^{K}\sum_{l=1}^{L} y_k \cdot \log(1 - s_{k,l}^{ob}),$$

$$s_{k,l}^{ob} = \text{sim}(\mathbf{v}_k^{io}, \mathbf{v}_{k,l}^{tb}),$$

where l denotes the different co-occurring background.

2.5 Area Regularization

A pixel-level area regularization term to constraint the size of activation maps to ensure that the irrelevant backgrounds are excluded in the activation map P_k:

$$\mathcal{L}_{REG} = \frac{1}{K}\sum_{k=1}^{K} S_k, \quad \text{where} \quad S_k = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} \mathbf{P}_k(h, w).$$

2.6 Overall Training Objective

$$\mathcal{L} = \alpha\mathcal{L}_{OTM} + \beta\mathcal{L}_{BTM} + \gamma\mathcal{L}_{CBS} + \delta\mathcal{L}_{REG},$$

where $\alpha, \beta, \gamma, \sigma$ are hyper-parameters.

3 Results

Compared with conventional CAM and recent method, Proposed CLIMS generate activation maps with more complete object and less class-related background regions.
In segmentation tasks, this proposed method yields better performance than recent methods and even achieved competitive performance compared to methods with additional saliency map (obtained from a fully supervised model).

4 Ablation Study

LOSS Ablation study on Loss:

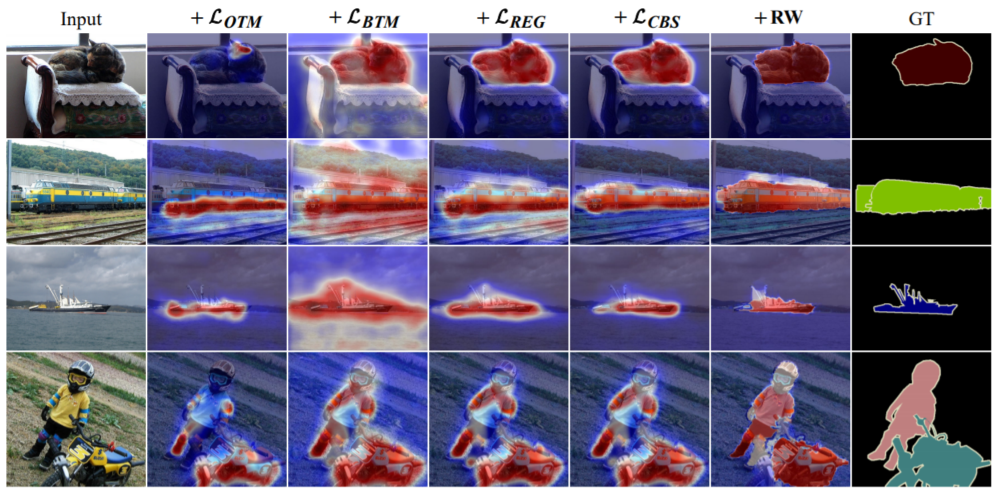

Figure 3. Initial CAMs generated by the proposed CLIMS using different combinations of loss functions. Input images are shown in column 1. Columns 2 to 5 present the generated CAMs using $\mathcal{L}_{OTM}$, $\mathcal{L}_{OTM} + \mathcal{L}_{BTM}$, $\mathcal{L}_{OTM} + \mathcal{L}_{BTM} + \mathcal{L}_{REG}$, and $\mathcal{L}_{OTM} + \mathcal{L}_{BTM} + \mathcal{L}_{REG} + \mathcal{L}_{CBS}$, respectively. **RW** denotes the refinement of PSA [2]. Best viewed in color.

By only using L_{OTM}, only the discriminative object parts are activated. Addition of L_{BTM} increased the the size of activated regions. L_{REG} efficiently constrains the size of activated regions and L_{CBS} significantly excludes the class-related background.

Class-related Background Based on the embeddings of background regions and text descriptions, L_{CBS} can effectively exclude these co-concurring backgrounds from the activated regions of foreground objects.

5 Summary

This paper proposed CLIMS to introduce natural language supervision for WSSS. The design of 4 loss function improved the performance in different aspect. The use of extra information in a main tendency in WSSS.

References

[1] J. Xie, X. Hou, K. Ye, and L. Shen, "Cross language image matching for weakly supervised semantic segmentation," *arXiv preprint arXiv:2203.02668*, 2022. (document)