Note:Weakly Supervised Semantic Segmentation by Pixel-to-Prototype Contrast

Sinkoo

April 10, 2022

1 Introduction

This paper proposed a novel weakly-supervised pixel-to-prototype contrastive learning method for WSSS. This method based on 2 main ideas: (i)features should retain semantic consistency across different views of an image; and (ii) pixels sharing the same label should have similar representations in the feature space.
In this method, a prototype is a embedding of a category which is estimated from pixel-wise feature embeddings with the top activations in the CAMs.. And the core idea is pulling pixels together to their positive prototypes and pushing them away from their negative prototypes.
This method can be seamlessly incorporated into existing WSSS models without any changes to the base networks.

2 Method

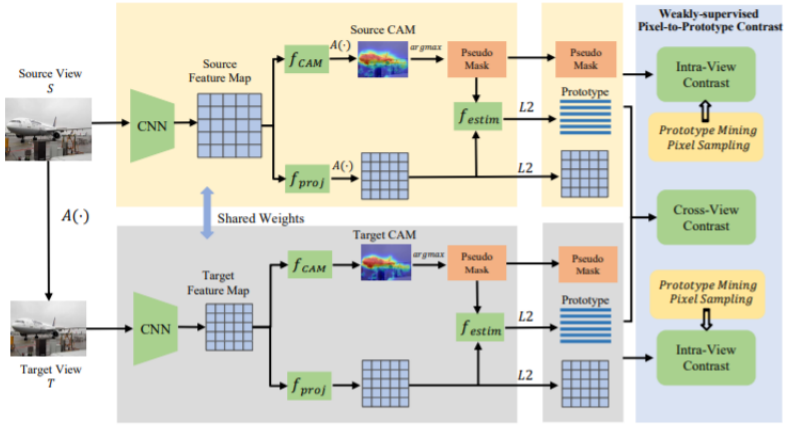

Figure 2. The overall pipeline of our proposed pixel-to-prototype contrast for WSSS. $A(\cdot)$ is a spatial transformation for augmenting training samples. f_{CAM}, f_{proj} are implemented by 1×1 convolutional layer followed by ReLU. f_{est} represents the prototype estimation process and $\boldsymbol{p}^{\{S,T\}}$ represent the generated prototypes. *L2* denotes per-pixel L2 normalization. The *argmax* function is conducted per-pixel along the channel dimension and returns the index of the maximum value.

The total loss consists of two parts: $L^{contrast} = \alpha L^{cross} + \beta L^{intra}$

2.1 Preliminary

in this paper, 1*1*C filter is used to compute GAMs.

2.2 Pixel-to-Prototype Contrast

Given the CAM, use a pixel-wise argmax function to generate pseudo mask y determines the category of each pixel.
The pixel-to-prototype contrast:

$$\mathcal{F}(\boldsymbol{v}_i;\, \boldsymbol{y}_i;\, \mathcal{P}) = -\log \frac{\exp(\boldsymbol{v}_i \cdot \boldsymbol{p}_{\boldsymbol{y}_i}/\tau)}{\sum_{\boldsymbol{p}_c \in \mathcal{P}} \exp(\boldsymbol{v}_i \cdot \boldsymbol{p}_c/\tau)}$$

2.3 Prototype Estimation

Choosing the ones with top K confidences to estimate the prototype:

$$\boldsymbol{p}_c = \frac{\sum_{i \in \Omega_c} \boldsymbol{m}_{c,i} \boldsymbol{v}_i}{\sum_{i' \in \Omega_c} \boldsymbol{m}_{c,i'}}$$

where Ω_c is set of chosen pixels of class c, and each pixel has CAM value $m_{c,i}$. K is a hyper-parameter. And compute prototypes across the entire training batch.

3 Cross-view Contrast

Given source view S, get a target view T through spatial transformation. Then these two views are encoded through a pre-trained CNN backbone and later processed to CAMs. Apply same transformation to the feature map and the CAM generated by S.

Cross Prototype Contrast Prototypes P' from the other view can be used to supervise the current view.

$$\mathcal{L}^{\text{cp}} = \frac{1}{|I|} \sum_{i \in I} \mathcal{F}(\boldsymbol{v}_i; \boldsymbol{y}_i; \mathcal{P}')$$

Cross CAM Contrast Pseudo labels $y_i^{'}$ from the other view can be used to supervise current view, too.

$$\mathcal{L}^{\text{cc}} = \frac{1}{|I|} \sum_{i \in I} \mathcal{F}(\boldsymbol{v}_i; \boldsymbol{y}_i'; \mathcal{P})$$

And then the total cross-view contrastive loss is : $L^{cross} = L^{cp} + L^{cc}$ for both S and T

4 Intra-view Contrast

Intra-view Contrast

$$\mathcal{L}^{\text{intra}} = \frac{1}{|I|} \sum_{i \in I} \mathcal{F}(\boldsymbol{v}_i; \boldsymbol{y}_i; \mathcal{P})$$

But trivially introducing L^{intra} could cause performance degeneration probably because the pseudo labels are inaccurate.

Semi-hard Prototype Mining Use semi-hard prototype mining: for each pixel, first collect the top 60% hardest negative prototypes, from which then choose 50% as the negative samples to compute the intra-view contrastive loss. Definition of "harder": For pixel i, view the prototypes except p_{y_i} with dot products to pixel feature embedding v_i closer to 1 to be harder.

Hard Pixel Sampling When computing Intra-view loss: for each class, half of the pixels are randomly sampled and half are the hard ones.
"harder pixel" is the opposite of "harder prototype" for we need to pay more attention to pixels far away.

5 Results

This method outperforms all existing methods in a large margin, achieving new state-of the-art performance on PASCAL VOC 2012 benchmark.

6 Ablation Study

1, Effectiveness of each component: Cross Prototype and Cross CAM improved the performance. But when applying Intra View, the performance slightly drops both in training and validation, Proto mining and Pixel Sampling mitigate this issue.
2, Choices of K: In this paper, K is set as 32.
3. Spatial Transformation: This paper used scaling, but using multiple method seldom helps.

7 Summary

This paper proposed a new method to implement Pixel-Prototype contrast loss to help in WSSS. And get to SOTA. This method mainly generate 2 views and use them to caculate the Cross loss and Intra loss and use both of them to show the performance of this model.