Note:Regional Semantic Contrast and Aggregation for Weakly Supervised Semantic Segmentation

Sinkoo

April 10, 2022

1 Introduction

Learning semantic segmentation from weekly-labeled data(e.g., image tags only) is challenging since it is hard to infer dense object regions from sparse semantic tags. And recent methods use only one image or limited number of images information to generate object localization map. To alleviate this, this paper proposed Regional semantic Contrast and Aggregation(RCA). RCA is equipped with a memory bank to store object pattern appearing in training data.
RCA explores semantic relations of regions in each mini-batch and the memory bank from two novel perspectives:
1, Semantic contrast: For each pseudo region, semantic contrast enforces the network to pull its embedding close to memory embeddings of the same category and push apart those of different.
2, Semantic aggregation: allows the model to gather dataset-level contextual knowledge to yield more meaningful object representations. This is achieved via a nonparametric attention module which summarizes memory representations for each image independently.

2 Method

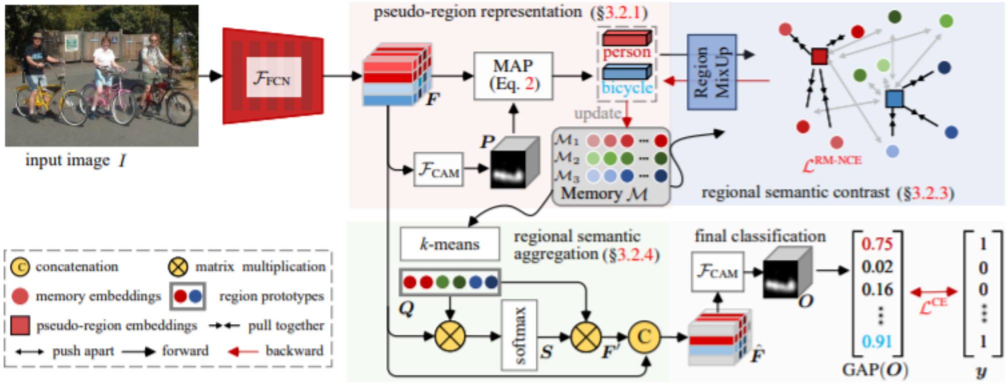

Figure 2. Detailed illustration of **regional semantic contrast and aggregation**. See §3 for more details.

2.1 Regional Semantic Contrast and Aggregation

2.1.1 Pseudo-Region Representation

$$\boldsymbol{F} = \mathcal{F}_{\text{FCN}}(\boldsymbol{I}) \in \mathbb{R}^{W\times H\times D}, \quad \boldsymbol{P} = \mathcal{F}_{\text{CAM}}(\boldsymbol{F}) \in \mathbb{R}^{W\times H\times L}. \quad (1)$$

For the l-th category that appears in I (i.e.,$y_l = 1$), its region-level semantic information is summarized to a compact embedding vector $f_l \in R^D$by masked average pooling (MAP).(y is image-level label vector where $y_l = 1$ represents objects if l-th category exists in the image. And D is number of channels)

$$\boldsymbol{f}_l = \frac{\sum_{x=1,y=1}^{W,H} \boldsymbol{M}_l(x,y)\boldsymbol{F}(x,y)}{\sum_{x=1,y=1}^{W,H} \boldsymbol{M}_l(x,y)} \in \mathbb{R}^D, \qquad (2)$$

where $M_l = \mathbf{1}(P_l > \mu)$ is a binary mask highlighting only strongly-activated pixels of class l in its activation map. The threshold μ is set to the mean value of P_l.

2.1.2 Pseudo-Region Memory Bank

The memory bank M consists of L dictionaries, i.e.,$M = \{M_1, M_2, ..., M_L\}$, each for one category.
Update: $m_l \leftarrow \gamma m_l + (1-\gamma)f_l$
update m_l when the l-th class appears in I (i.e.,$y_l = 1$) and its classification score is higher than a threshold μ

2.1.3 Regional Semantic Contrast (RSC)

Region-aware contrastive loss:

$$\mathcal{L}_l^{\mathrm{NCE}}(\boldsymbol{f}_l, y_l) = \frac{1}{|\mathcal{M}_l|} \sum_{\boldsymbol{m}_l^+ \in \mathcal{M}_l} -\log \frac{e^{\mathrm{sim}(\boldsymbol{f}_l, \boldsymbol{m}_l^+)/\tau}}{e^{\mathrm{sim}(\boldsymbol{f}_l, \boldsymbol{m}_l^+)/\tau} + \sum_{\boldsymbol{m}_l^- \in \mathcal{M} \setminus \mathcal{M}_l} e^{\mathrm{sim}(\boldsymbol{f}_l, \boldsymbol{m}_l^-)/\tau}}, \quad (4)$$

where $sim(i,j) = \frac{i*j}{|i|_2 |j|_2}$
But labels are week and noisy.
In this paper, assume that regions l and $l^{`}$ are from different categories. The embedding of the mixed region is computed as $\hat{f}_l = \omega f_l + (1-\omega) f_{l^-}$
And a new region mixup caontrastive loss:

$$L_l^{RM-NCE} = \omega L_l^{NCE}(\hat{f}_l, y_l) + (1-\omega) L_l^{NCE}(\hat{f}_l, y_{l^-})$$

Encourage the network to learn relative similarities for mixed regions, regularizing the model to learn robust representations from label-imperfect samples.

2.1.4 Regional Semantic Aggregation (RSA)

Exploit dataset-level context cues in the memory bank for enhancing semantic understanding. And directly aggregating large-scale representations is computationally expensive. So, for each class l, do k-means clustering over all features in M_l to obtain K prototype vectors (i.e., class centroids), organized in a matrix form $Q_l \in R^{K \times D}$
Then, calculate its affinity matrix S with the prototypical representation Q as follows: $S = softmax(F \otimes Q^T) \in R^{(WH)*(LK)}$
F' denotes an enriched feature representation of F: $F' = S \otimes Q$, and further reshaped into $R^{(WH) \times D}$. Then concatenate F' and F to get $\hat{F}$. Here, $\hat{F}$ not only encodes intra-image local contexts in F, but also captures inter-image global contexts in $F2$.

2.1.5 Class Activation Map Prediction

Final activation maps O:

$$O = F_{CAM}(\hat{F})$$

2.2 Detailed Network Architecture

Detailed Architecture is in Figure2. The loss function of this classifier is :

$$L = \sum_I \alpha_1 L^{RM-NCE} + \alpha_2 L^{CE}(GAP(P), y) + L^{CE}(GAP(O), y)$$

The first term L^{RM-NCE} is the region mixup contrastive loss, which is computed as the average loss of all regions appearing in I. The second one is an auxiliary cross-entropy loss L^{CE} for supervising the intermediate CAM prediction P, while the third loss is the main cross-entropy loss imposing on the final CAM prediction O.

3 Ablation Study

1, Semantic Contrast and Semantic Aggregation actually helped.
2, region mixup indeed helps the model learn more robust representations from noisy data.
3, Memory Updating Coefficient γ : it is beneficial to update the memory in a relatively slow speed, but not too slow. In this paper, the optimal value is 0.99.
4, Prototype Number K: . As seen from the table, RCA shows stable performance when K is in 10100.
5, Memory Size: By storing only 100 or 500 region embeddings per class, the performance only degrades very slightly, indicates that the model is scalable to lager scale datasets.

4 Summary

this paper proposed a method to use inter-image information to help to learn semantic segmentation using image-level labels by keeping a memory bank storing storing massive historical features and exploiting semantic relations between memory and samples as additional supervisory signals (by semantic contrast) or holistic contextual cues (by semantic aggregation) to improve network learning and inference. RCA(with backbone network) sets a new state-of-the-art.