Note: Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals

sgc

June 5, 2022

1 Introduction

Previous works on unsupervised semantic segmentation mainly focus on building proxy works, thus only concentrate on limited domain of the image. This paper made a first attempt to build a pixel level prior, and then use a contrastive framework to learn embeddings.

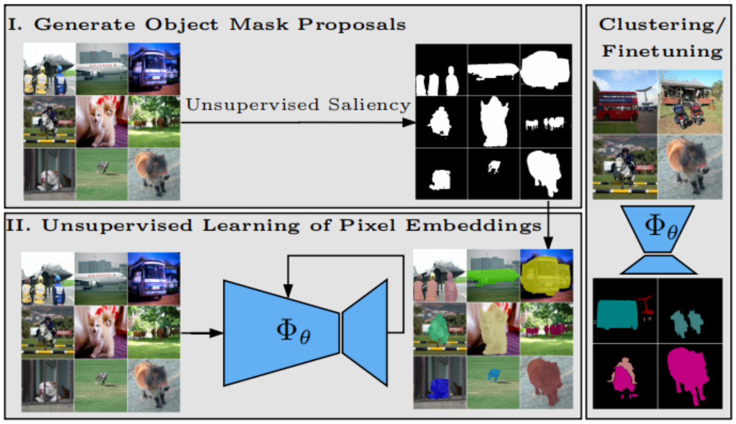
I. Generate Object Mask Proposals
Unsupervised Saliency
II. Unsupervised Learning of Pixel Embeddings
Φ_θ
Clustering/
Finetuning
Φ_θ

Figure 1. We learn pixel embeddings for semantic segmentation in an unsupervised way. First, we predict object mask proposals using unsupervised saliency. Second, we use the obtained masks as a prior in a self-supervised optimization objective. Finally, the pixel embeddings can be clustered or fine-tuned to a semantic segmentation of the image.

2 Method

The proposed method is based on divide-and-conquer strategy. It first look for images regions where pixels are likely to belong together. MaskContrast consists of 2 steps: first determine a prior by identifying objects in the images that can be grouped together, and then employ the obtained prior in a contrastive loss.

2.1 Mining Object Mask Proposals

The saliency estimation is use to generate the object mask proposals.

2.2 MaskContrast: Learning Pixel Embeddings by Contrasting Salient Objects

The goal is ti learn a pixel embedding function $\Phi_\theta : \mathcal{X} \to \mathcal{Z}$ parameterized by a neural network with weights θ, that maps each pixel i in an image to a point z_i on a D-dimensional normalized hyper-sphere.

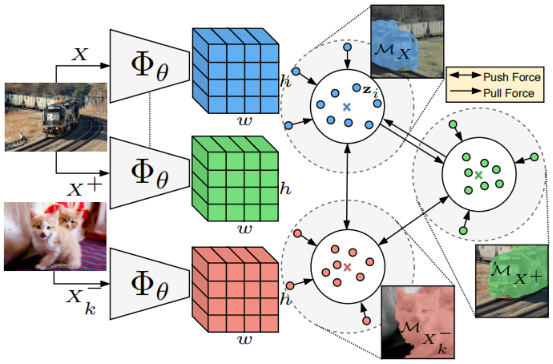

Figure 2. **MaskContrast** learns pixel embeddings for unsupervised semantic segmentation in the following way. We use a saliency estimator to generate positive pairs of object-centric crops (X, X^+) and negative pairs X_k^-. The model Φ_θ is trained to maximize the agreement between embeddings of pixels belonging to the objects in X, X^+, while minimizing the agreement with pixels from objects in X_k^-.

Learning Image-Level Representations. Positive views contain the same object and they are generated by applying augmentations. Negative views never contain the same object.

Learning Pixel-Level Representations. The optimization is derived from a pull- and push-force.

$$\mathcal{L}_i = -\log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_{\mathcal{M}_{X+}}/\tau\right)}{\sum_{k=0}^{K} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_{\mathcal{M}_{X_k^-}}/\tau\right)}.$$

where the mean pixel embedding $z_{\mathcal{M}_n}$ of an object mask M_n be defined as:

$$\mathbf{z}_{\mathcal{M}_n} = \frac{1}{|\mathcal{M}_n|} \sum_{i \in \mathcal{M}_n} \mathbf{z}_i.$$

3 Summary

This work presented a general two-step framework based upon a mid-level visual prior for tackling unsupervised semantic segmentation. This work does not concentrate on low level information and get great performance.

References