Note: FreeSOLO: Learning to Segment Objects without Annotations[1]

Sinkoo

April 17, 2022

1 Introduction

To deal with expensive pixel annotation or box annotations containing strong localization information, this paper proposed FreeSOLO based on their recent works of SOLO. FreeSOLO consists of 2 major parts: Free Mask and Self-supervised SOLO.
Contributions:
1, Free Mask approach leverage the specific design of SOLO to extract coarse object masks and semantic embeddings in an unsupervised manner.
2, Self-supervised SOLO takes the coarse masks and semantic embeddings from Free Mask and trains the SOLO instance segmentation model and designed to handle noisy.
3, FreeSOLO presents a simple and effective framework that demonstrates unsupervised instance segmentation successfully for the first time.

2 Method

Background(SOLO) The model consists of two branches, a category branch and a mask branch. The category branch predicts the semantic categories. The mask branch generates S^2(Number of grids) sized masks, one corresponding to each grid cell.

2.1 Overview of FreeSOLO

FreeSOLO does not require any annotations, only unlabeled image is enough.

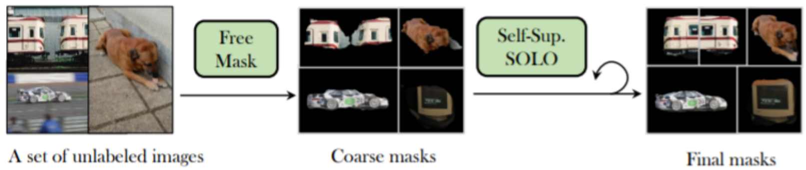

Figure 2. Overview of FreeSOLO. Unlabeled images are first input to Free Mask to generate coarse object masks. The segmentation masks as well as their associated semantic embeddings are used to train a SOLO-based instance segmentation model via weak supervision. We use self-training to improve object mask segmentation.

And the well-trained model can serve as a strong pre-trained model for downstream tasks.

2.2 Free Mask

Extract the dense feature maps $I \in R^{H*W*E}$ from a backbone model trained via self-supervision. Quires $Q \in R^{H'*W'*E}$ is generated by downsampling I. Keys K is I itself. The score maps $S \in R^{H*W*(H'W')}$ is obtained:

$$S_{i,j,q} = sim(Q_q, K_{i,j})$$

where $Q_q \in R^E$ is the q^{th} query, and $K_{i,j} \in R^E$ is the key at spatial location (i, j).
The score maps are then normalized as soft masks by shifting the scores to the range [0, 1].
The final output is $M = NMS(Maskness(Norm(S)))$.

Self-supervised pre-training Free Mask uses a pretrained backbone via self-supervision as the starting point. The authors leverage the self-supervised model pretrained with dense correspondence. It optimizes a pairwise (dis)similarity loss at the level of local features between two views of the input image.

Pyramid queries When constructing the queries Q from I, a pyramid queries method is designed to generate masks for instances at different scales. First set a list of scale factors, *e.g.*, [1.0, 0.5, 0.25], when downsampling I, this obtains a list of Q at different scales from large to small. All pyramid queries are flattened and concatenated together as the final Q.

Maskness score

$$maskness = \frac{1}{N_f} \sum_{i}^{N_f} p_i$$

. where N_f denotes the number of foreground pixels of the soft mask p, *i.e.*, the pixels that have values greater than threshold τ.

2.3 Self-Supervised SOLO

Learning with coarse masks Use the masks as weak annotations. This paper first project the predicted masks and the coarse masks on to a x-axis and a y-axis and use Dice loss.
And they proposed 2 methods:

1, max:

$$\begin{aligned}\mathcal{L}_{max\_proj} &= \mathcal{L}(\max_x(\boldsymbol{m}), \max_x(\boldsymbol{m}^*)) \\ &+ \mathcal{L}(\max_y(\boldsymbol{m}), \max_y(\boldsymbol{m}^*)),\end{aligned}$$

2, average:

$$\begin{aligned}\mathcal{L}_{avg\_proj} &= \mathcal{L}(\text{avg}_x(\boldsymbol{m}), \text{avg}_x(\boldsymbol{m}^*)) \\ &+ \mathcal{L}(\text{avg}_y(\boldsymbol{m}), \text{avg}_y(\boldsymbol{m}^*)),\end{aligned}$$

where m and m^* are predicted and coarse masks.
They also employ a pairwise affinity loss $L_{pairwise}$
The total loss for mask prediction is : $L_{mask} = \alpha L_{avg-proj} + L_{max-proj} + L_{pairwise}$

Self-training Input unlabeled images into the instance segmenter and collect their predicted object masks. The low confidence predictions are removed and the remaining ones are treated as a new set of coarse masks. Again the model trains an instance segmenter with the unlabeled images and the new masks. Performing self-training once already brings clear improvements and more iterations do not provide additional gains.

Semantic representation learning

References

[1] X. Wang, Z. Yu, S. De Mello, J. Kautz, A. Anandkumar, C. Shen, and J. M. Alvarez, "Freesolo: Learning to segment objects without annotations," *arXiv preprint arXiv:2202.12181*, 2022. (document)