IIT4204 SW Project Final Report

2021849341 Se Ho Kwak

1. Introduction

Semantic segmentation is a computer vision task that not only aims to classify what an object is on an image, but also identifies where that object is in the image. Semantic segmentation has many important use-cases, such as in self-driving cars, where vehicles need to locate and identify traffic lights, signs, humans, and other vehicles on the road. It is also useful in medical image diagnosis when detecting anomalies in CT scans.

With weakly-supervised learning, concrete target labels for each input are not available, but rather weaker, lower-quality labels are used. For the weakly-supervised semantic segmentation(WSSS) task, common methods are to use cheaper labels such as bounding boxes, scribbles, or image-level labels. These labels are used to generate pseudo-masks which are then used to train the segmentation model. Hence it is important to be able to generate high quality pseudo-masks. In this paper, we outline the methods used to train our classification model and analyze the quality of the generated CAMs and pseudo-masks.

2. Related Work

There are many previous works dedicated to the segmentation task using a wide range of methods including contrastive learning, an unsupervised learning approach which utilizes self-created pseudo-labels [1], or consistency regulation, a semi-supervised learning method using partially labeled and unlabeled data [2]. However, because we are interested in weakly-supervised semantic segmentation, we focus largely on recent works that implement weakly-supervised learning. For instance, [3] uses random walks to iteratively refine CAMs. Other works [4] erase certain regions to guide networks to learn more non-distinctive parts of objects. Others such as [5] learn activation maps by learning from two images containing the same objects.

3. Methodology

We implement the following pipeline for a general WSSS task that uses image-level labels. First, a classification model is trained, then pseudo-masks are generated using class activation mapping, and finally the segmentation model is trained using the pseudo-masks generated previously. For the purpose of this course project, we do not train the segmentation model, but only train the classification model and generate pseudo-masks using CAMs.

For our dataset, we use the PASCAL VOC 2012 dataset which includes over 9,000 labeled images with 20 classes. We train two separate classification networks on data labelled with five classes and ten classes.

For the classification model, we use a residual network, or ResNet as introduced in [6]. ResNets are deep neural networks that utilize residual blocks which contain skip connections between layers. These skip connections help solve the problem of vanishing gradients when going from layer to layer. There are many widely-used pretrained models such as ResNet-50, ResNet-101, or ResNeSt-269. Although the latter two have been shown to outperform ResNet-50, due to limitations in time and computational resources in comparison to the gain in performance, we use ResNet-50 as our backbone for the classification network.

Because CAMs have a tendency to highlight the most discriminative parts of images, we implement the Hide and Seek method from [4]. Hide and seek is a data augmentation technique that drops parts of training images randomly, forcing the network to learn other relevant features when the most discriminative parts are hidden.

4. Results

We use a mean Intersection over Union(mIoU) as the measure to analyze the quality of the pseudo-masks generated using CAMs. The mIoU is defined to be the mean area of overlap between the ground truth labels and the predicted masks divided by the union of the two.

After training our classification network for 20000 iterations on the baseline model, the respective measures for five classes are represented in Table 1 below.

Table 1.

IDX	Name	IoU	Prec	Recall
-	mean	47.20	75.67	56.71
0	background	73.37	78.82	91.38
1	car	52.14	76.10	62.34
2	cat	37.47	89.14	39.27
3	chair	30.54	58.57	38.96
4	dog	38.55	82.28	42.03
5	person	51.12	69.11	66.26

After dropping out patches of images in our training dataset using the Hide and Seek method, we were able to improve our results which are represented in Table 2 for five classes and Table 3 for ten classes. Once the CAMs were generated, because the predicted masks had seemed to create large, general pictures, by lowering the threshold for CAMs from 0.2 to 0.1, we are able to further improve our results for the CAMs. The final results are shown in Table 4 for five classes and Table 5 for ten classes.

Table 2.

IDX	Name	IoU	Prec	Recall
-	mean	51.30	76.84	61.46
0	background	74.94	81.05	90.86
1	car	56.64	78.42	67.10
2	cat	43.17	88.73	45.68
3	chair	33.98	59.59	44.16
4	dog	42.54	83.09	46.57
5	person	56.50	70.13	74.41

Table 3.

IDX	Name	IoU	Prec	Recall
-	mean	47.29	71.43	58.55
0	background	73.28	80.74	88.79
1	car	44.32	70.84	54.21
2	cat	41.55	88.21	44.00
3	chair	29.56	48.26	43.27
4	dog	43.34	80.50	48.42
5	person	53.97	68.02	72.31
6	bus	57.94	89.50	62.17
7	motorbike	53.51	75.08	65.06
8	sofa	38.25	55.52	55.14
9	train	43.50	61.85	59.45
10	tvmonitor	40.96	67.22	51.18

Table 4.

IDX	Name	IoU	Prec	Recall
-	mean	56.98	70.53	74.99
0	background	73.20	87.59	81.67
1	car	60.00	68.02	83.57
2	cat	59.86	82.03	68.89
3	chair	38.57	50.87	61.46
4	dog	54.43	73.65	67.59
5	person	55.83	61.02	86.78

Table 5.

IDX	Name	IoU	Prec	Recall
-	mean	50.59	65.80	69.22
0	background	71.00	84.75	81.40
1	car	50.92	64.17	71.14
2	cat	42.03	84.22	45.62
3	chair	30.63	42.02	53.06
4	dog	53.82	77.25	63.96
5	person	54.88	61.60	83.40
6	bus	69.31	83.33	80.46
7	motorbike	57.66	68.04	79.09
8	sofa	38.74	50.70	62.16
9	train	44.21	51.36	76.05
10	tvmonitor	43.25	56.34	65.05

5. Conclusion

For our weakly-supervised semantic segmentation task, we are able to achieve up to 56.98% mIoU for this WSSS task in five classes and up to 50.59% mIoU for ten classes. When implementing the Hide and Seek data augmentation technique, we are able to increase the mIoU for five classes from 47.20% to 51.30%. We are further able to increase this mIoU for both five classes and ten classes by tuning the threshold for generating CAMs.

References

[1] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. Technologies, 2021.

[2] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In ECCV, 2020.

[3] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In CVPR, 2018.

[4] Krishna Kumar Singh, Hao Yu, Aron Sarmasi, Gautam Pradeep, and Yong Jae Lee. Hide-and-Seek: A Data Augmentation Technique for Weakly-Supervised Localization and Beyond. In ICCV, 2017.

[5] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Cian: Cross-image affinity net for weakly supervised semantic segmentation. arXiv preprint arXiv:1811.10842, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. arXiv:1512.03385, 2015