Consistency-CAM
Towards Improved Weakly supervised Semantic Segmentation

TL;DR
• We identify and propose three key improvements for high performing weakly supervised semantic segmentation (WSSS) tasks. The resulting Consistency-CAM framework attains superior performance on PascalVOC and MS-COCO datasets.

Introduction
• Typical pipelines for WSSS are trained in two stages.
  – 1. Train a classification network with global average pooling to obtain 2d class activation maps (CAMs). 2. Train a segmentation network using CAMs as pixel-level supervision.
  – Issue: CAMs are noisier than real labels and needs refinement using some regularization.
• Puzzle-CAM splits the image into multiple tiles and ensures that the CAM for the image matches the CAM obtained after stitching the individual CAMs.
  – However, pre-training using single-label prediction has a negative effect since image segmentation datasets have more than one class.
  – Also, GAP enforces the network to overrepresent the labeled objects in the feature maps
  – Lastly, Puzzle-CAM uses fixed tile sizes & the puzzle operation can be complemented with other transformation
• We address the above three issues highlighted using our Consistency-CAM pipeline.

Method
1. We train the backbone with MILe, which learns multi-label representations from singly labeled images.
   – Ensures the backbone is able to predict multiple classes per image.
2. We replace GAP with a differentiable noisy-or operation. This marks the presence of a class in an image independent to the number of pixels that belong to that class.
   – With noisy-or a class will be active with high probability even if only one pixel is activated for that class.
3. We propose a more general set of transformations.
   – We use a consistency or reconstruction loss to ensure that the CAM is robust to a diverse set of augmentations as well as the puzzle operation in with different tile sizes.

• We pretrain the backbone for multi-label classification.
• We change the GAP operation by a noisy-or operation.
• We propose a more general set of augmentations for CAM refinement.
• Finally, these three improvements result in better performance on COCO and Pascal.
• Our method improves Puzzle-CAM by several points.
• We also see that training the backbone with multi-label learning is beneficial.

Quantitative Results

Conclusions

References