Note: Instance Semantic Segmentation Benefits from Generative Adversarial Networks

sgc

June 20, 2022

1 Introduction

Using GAN to generate segmentation masks suffer from losses that are based on averaging pixel-wise differences. Authors tackle this problem by using a GAN loss merged with Mask R-CNN. They designed a new complementary GAN loss to make Mask R-CNN trainable in the GAN framework. Feature maps instead of real images are used in this model. And for GAN, it is trained on Precise RoI pooling instead of 4 coordinates.

2 Method

2.1 Objective

Like conventional GAN loss, the authors introduce L_{adv}^{Gb}**and**L_{adv}^{Gm} stand for generator loss of the bounding box and the mask, respectively. And the total loss for generator is:

$$L_{Generator} = L_{cls} + L_{bbox} + L_{mask} + L_{adv}^{G_b} + L_{adv}^{G_m},$$

The loss for discriminator:

$$L_{Discriminators} = L_{adv}^{D_b} + L_{adv}^{D_m}.$$

2.2 Network architecture

2.2.1 Box head

Generator :Just utilize the box head of Mask R-CNN as the box generator, this generator takes in RoI features, and outputs class and bounding box prediction.

Discriminator : The feature map of the predicted boxes and its ground truth as fake and real samples to the discriminator.

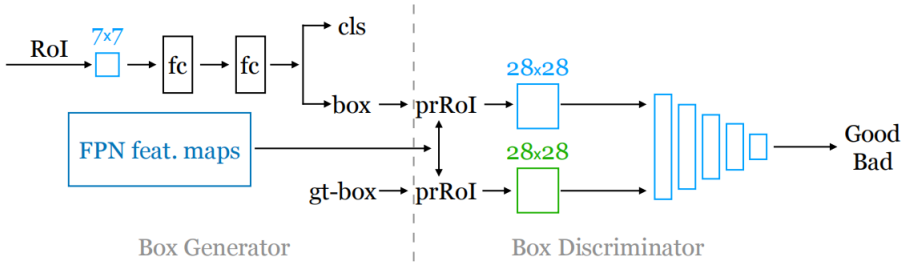

Figure 2: Box head architecture. We propose an adversarial training framework for object detection by extending the current box head architecture in Mask R-CNN with an FPN backbone [11]. We design a new discriminator network, which takes in bounding box predictions from the box head and their ground truths, then extracts the corresponding regions of interest (RoI) from the FPN feature maps using Precise RoI Pooling (prRoI) [15]. Finally, the discriminator outputs a score between $[0, 1]$ that represents how good the bounding box prediction is. There are 5 convolution layers in the discriminator network: each convolution layer is followed by a BatchNorm and a LeakyReLU layer. Note we only use convolution layers with the correct stride and kernel size to down-sample the bbox prediction from its original spatial dimension to our desired output shape.

2.2.2 Mask head

Generator :Assemble Box head

Discriminator :feed two inputs to the mask discriminator, the binary mask prediction and its ground truth, both multiplied element-wise with the RoI features.

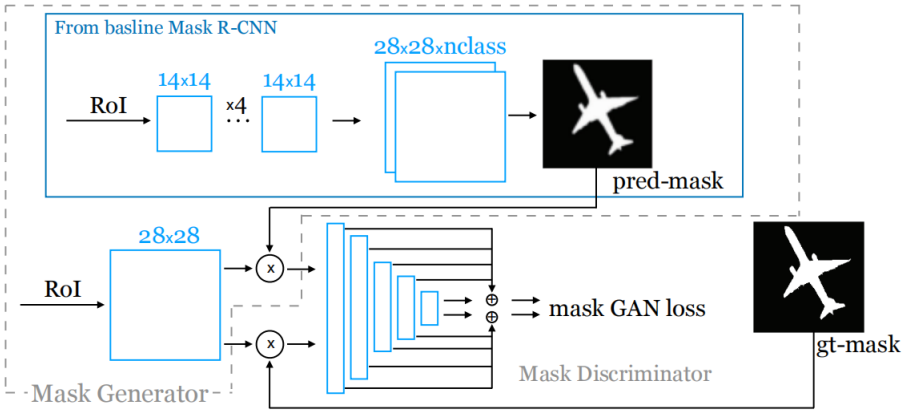

Figure 3: Mask head architecture. We utilize the mask head of Mask R-CNN [11] as the generator, and build another discriminator network for the mask segmentation task. During training, we first pick the binary mask predictions of true objects (pred-mask) and their ground truths (gt-mask), calculate the pixel-wise multiplication between these masks and the corresponding regions of interest (RoI), and send the results to the discriminator as inputs. Outputs from each layer of the discriminator are flatten then concatenated and used for final GAN loss computation. Similar to the box head, the mask head also has 5 convolution layers in the discriminator network, where each convolution layer is

3 Summary

This paper introduced designs to implement mask RCNN to GAN training, makes the framework more generalizable for learning domain-agnostic datasets. Experiments show that this GAN outperforms original Mask RCNN.

References