Assignment 1 – Report

Introduction

The Mask R-CNN is a state-of-the-art model for object instance segmentation that was introduced by He, Gkioxari, Dollar and Girshick in 2018 in a paper titled “Mask R-CNN”. It extends the Faster R-CNN model by generating pixel-level masks for each detected object such that instance segmentation can be done. In this report, we will discuss the method and result of the Mask R-CNN paper.

Method

Compared to Faster R-CNN the Mask R-CNN model works in two stages. In the first stage, a Region Proposal Network (RPN) is used to assign regions of interest (RoIs). In the second stage, the Fast R-CNN network is used to extract features from RoIs to classify them. In addition to that, a branch is added that predicts segmentation masks for each instance in a pixel-to-pixel manner.

Architecturally, the Mask R-CNN model can be divided into three parts: the convolutional backbone network, the region proposal network (RPN) and the Faster R-RCNN network with an additional branch. These networks run once per image to return a set of region proposals. The backbone is usually a pre-trained CNN such as ResNet or VGG, that extracts features from the input image, whereas the RPN generates a set of rectangular candidate regions, that may have an object. The additional branch to the Faster R-CNN network is a fully convolutional network (FCN) that takes the last convolutional layer of the Fast R-CNN network to generate a binary mask for each instance.

The multi-task loss function used in Mask R-CNN for training combines the losses from the classification, bounding box regression, and mask prediction tasks: the classification loss penalizes incorrect predictions of the object category, the bounding box regression loss penalizes errors in the predicted bounding box coordinates and the mask segmentation loss penalizes errors in the predicted binary mask for each object instance. The total loss is defined as follows:

$L = L_{cls} + L_{box} + L_{mask}$

Result

The authors of the Mask R-CNN paper tested the model on several benchmark datasets, such as COCO and Cityscapes. Compared to previous models, it was able to outperform in both object detection and instance segmentation. Especially its accuracy and time efficiency was notable due to its ability to perform instance segmentation in real-time. This is possible due to the mask branch sharing the same backbone as the RPN and Faster R-CNN network, such that computation is facilitated. Not only the larger backbone network but also a more efficient RoIAlign operation, and the inclusion of skip connections between the backbone and the mask branch have been attributed to the model’s success.

Also, the use of pre-trained convolutional networks for feature extraction significantly improves the model's overall performance on object detection and instance segmentation tasks. Even though the Mask R-CNN model is highly successful the authors of the paper acknowledged that there are limitations to it, such as the reliance on large amounts of labeled data needed for training.

Conclusion

In conclusion, we can say that the Mask R-CNN model has deemed itself a highly effective and powerful model for object instance segmentation. Compared to its predecessor the model not only allows object detection but also accurate instance segmentation which is a significant improvement.

References:

He, K., Gkioxari, G., Dollar, P. and Girshick, R. (2018). Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*