Research paper Summary

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

This paper is all about region proposal network-based object detection. It is an enhanced version of Fast R-CNN. Instead of using selective search to make object proposals the method introduces a CNN based region proposal network. The proposed method in the paper is a unified network for object detection. This model uses the idea of an anchor boxes, and using this anchor boxes of different sizes the model get huge performance boost on object detection. It achieves improved performance on numerous datasets, including as PASCAL VOC 2012 and MS COCO, with just 300 proposals per image, and works at a frame rate of 5fps on a GPU. They named this object detection system Faster R-CNN, because it is much faster than the previous fast R-CNN. The mAP on PASCAL VOC 2007 test dataset is 69.9%.

The very first step of this algorithm is region proposal network. In region proposal images feed into convolution layer and output a feature map. Then a small layer of CNN is applied on this extracted feature map, and it's generate a box proposal. It uses the idea of anchor boxes for finding the region where the object are presents. Anchor boxes are basically a set of predefined bounding boxes with some height and width. For using the different sizes of anchor boxes it can find the different sizes objects.

The training of this Faster R-CNN model is divided into four steps. First the region proposal network is trained on the object detection dataset to make region proposals. After the model trained to make proposals Fast R-CNN is comes to play. They trained this model on the region proposal network. In the third step they use the detector network to initialize training of the region proposal network and fine-tune the layers unique to the region proposal network. In the final step the shared Conv layer is fixed and the fully connected layers of the Fast R-CNN is fine-tuned.

One drawback of Faster R-CNN is that the RPN is trained where all anchors in the mini-batch, of size 256, are extracted from a single image. Because all samples from a single image may be correlated (i.e. their features are similar), the network may take a lot of time until reaching convergence.