SUMMARY

Paper: Rich feature hierarchies for accurate object detection and semantic segmentation

Ocean Monjur

June 19, 2022

1 Summary

1.1 Overview

The paper proposes R-CNN a two-stage object detector, whose base characteristics also enable it to be used as a segmentation algorithm. R-CNN was one of the initial works in the field of object detection using Deep Learning. The purpose of this paper was to utilize deep learning in object detection and beat the state of art models of the time like Over feat.

1.2 R-CNN architecture and Training

Most Object Detectors before R-CNN used a sliding window approach over the whole image which slows down the process and makes it practically unusable. But the R-CNN,
1. Used a Region Proposal Algorithm (Selective search) to select (around 2000) regions of interest where the classifier is run.
2. They tested using VGG and AlexNet as their backbone feature extractors.
3. Finally they outputted the class using SVM from the extracted features and predicted the bounding boxes.
For training, the feature extractor (AlexNet) was pretrained using the ImageNet dataset, and the output softmax layer was replaced with the SVM which was fined tuned on the Pascal VOC dataset one SVM per class. The input images are warped to 227*227 to match the expected input of the AlexNet architecture.

1.3 Results

Some interesting points on the results, the authors show when they extract from the pool layers compared to the fc6 and fc7 layers the performance drops yes but ever so slightly (44.2 vs 46.2 in fc6 on the VOC 2007). The best performance they recorded came from the O-Net (VGG) backbone which achieved an mAP of 66.

1.4 Pros and Cons

At the time of publication, the R-CNN was the benchmark for Object Detection, it was relatively fast and was the most accurate Object detector of its time.
Some of its cons were addressed in the subsequent papers like Fast R-CNN and Faster R-CNN. One of the key disadvantages of the R-CNN was that the calculations weren't shared between the convolutions, which slows down the architecture with repetitive calculation.
The warping of images to fit the correct input resolution, causes some information to be lost, and multi stage training (one for classification another for detection fine tuning) slows down the R-CNN significantly.