Review of “Rich feature hierarchies for accurate object detection and semantic segmentation”

Riyad Bin Rafiq

1. Paper summary

The paper [1] proposed a new object detection algorithm that achieved a mean average precision (mAP) of 53.3% at VOC 2012. Before this, the best approaches were complicated ensemble systems combining multiple low-level image features. In this paper, the authors generated a number of region proposals from the input image and then classified and localized specific objects with a convolutional neural network (CNN). So, the network was named as R-CNN: Regions with CNN features. The authors also provided the visualization of learned features by the network and obtained satisfactory performance in segmentation tasks with minor modifications.

2. Contribution

2.1 Layer-wise learning and performance

In my opinion, the paper contributed mostly in giving the intuition of layer-by-layer learned features and performance. Top regions of pool5 (max pooled output of the network's final convolutional layer) were demonstrated where the network learns on specific representative samples such as dog face, dot array, red blob detector, etc. Moreover, experiments were done to understand the importance of certain layers for detection performance. For example, pool5, fc6 and fc7 layers were analyzed with and without fine-tuning on the dataset. In this way, the authors suggested removing the layers (with parameters) without degrading the mAP score.

2.2 Tackling data unavailability in detection

Labeled data is scarce for training a large CNN in an object detection task. The paper showed how supervised training on the ImageNet dataset followed by fine-tuning on a small (PASCAL) dataset effectively solved the issue. The authors achieved 8% detection improvements after fine-tuning.

3. Critique

3.1 Region proposals

In the paper, around 2000 region proposals were generated from an input image and those proposals were classified later by a CNN. But it was not mentioned how that number, 2000, was chosen. Was it chosen arbitrarily or followed any method? If the region proposals increased, would the performance improve? In my opinion, there should be a brief discussion of region proposals associated with performance.

3.2 Complicated method

The system is a bit complex as it contains three modules to detect an object. The first module generates region proposals from the input image, the second module including large CNN extracts feature vectors from each proposal and finally, the third module includes a set of SVMs that detect the class that belongs to the image with a specific localization. In this scenario, if we consider training a R-CNN, 2000 region proposals do forward propagation in a CNN for 2000 times and this takes a lot of time and storage (storing the features). That's why the test time of R-CNN is also slow.

Reference

1. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. pp. 580–587.