FCOS: A Simple and Strong Anchor-Free Object Detector – Critique

This is a critique of the above-mentioned paper done as a part of the assignment for CS 771. It follows the format described in the assignment writeup

Ramapriya Ranganath, Ruthvika Reddy Loka, Mahitha Pillodi

Summary

This paper introduces FCOS (Fully Convolutional One-Stage Object Detection), a novel object detection model that distinguishes itself from prevailing models like Faster R-CNN, YOLOv3, and RetinaNet by employing a single-stage methodology. Unlike the conventional two-stage approaches utilized by existing state-of-the-art models, FCOS streamlines the detection process by discarding the necessity for predefined anchor boxes. This departure from anchor-based computations aligns FCOS with the elegance of per-pixel prediction akin to FCN (Fully Convolutional Networks) in semantic segmentation. Moreover, FCOS demonstrates that a simpler FCN-based detector surpasses the performance of its anchor-based counterparts.

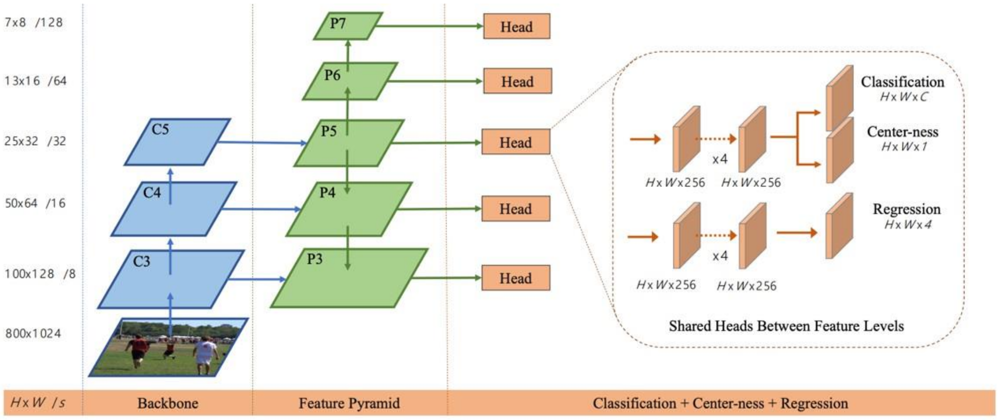

Figure 2 – The network architecture of FCOS, where C3, C4, and C5 denote the feature maps of the backbone network and P3 to P7 are the feature levels used for the final prediction. $H \times W$ is the height and width of feature maps. '/s' ($s = 8, 16, ..., 128$) is the down-sampling ratio of the feature maps at the level to the input image. As an example, all the numbers are computed with an 800×1024 input.

Experiments

- FCOS employs focal loss for classification, IOU loss for regression, and binary cross-entropy loss for image center-ness. It utilizes the ResNet-50 architecture as its backbone, pretrained on ImageNet. The training involves 90k iterations with SGD, a batch size of 16, and a learning rate of 0.01. Input images are resized, with the shorter side set to 800px and the larger side less than or equal to 1333px. During inference, the image undergoes the network's forward pass to obtain bounding boxes and class scores, followed by Non-Maximum Suppression (NMS) with a threshold of 0.6 to filter overlapping boxes.

- The authors conducted various experiments, including comparisons with RetinaNet, investigating the impact of center sampling and FPN, analyzing the significance of center-ness in crowded scenarios, and performing an ablation study removing group normalization. FCOS outperformed RetinaNet on the COCO dataset, achieved real-time performance with ResNet-50, and demonstrated a speed/accuracy tradeoff with DLA-34 surpassing CenterNet by 1.7% AP.

Major Contributions:

- FCOS is a single-stage, proposal-free, and anchor-free detector, enhancing efficiency and simplicity. It has significantly fewer parameters than anchor-based detectors, making it suitable for resource-constrained environments.
- The use of Feature Pyramid Networks (FPNs) ensures scale-invariant properties, improving accuracy in detecting objects of varying sizes. Sharing heads between different feature levels enhances parameter efficiency and overall model performance.
- A strategy to reduce low-level detections distant from the object center involves predicting center-ness, normalizing the distance from the object center to a location.
- Strengths and Weaknesses:

Strengths:

- FCOS simplifies computations by eliminating anchor boxes, leading to faster training and testing.
- Customizable for various tasks like instance segmentation, text spotting, and keypoint detection.
- Unifies FCN-related tasks with detection, facilitating the application of ideas from semantic segmentation.

Weaknesses:

- FCOS may not perform well with low-resolution images and exhibits limited accuracy.
- Sensitive to starting predictions, requiring careful initialization or pre-training.
- Lacks refinement compared to two-stage detectors, making it less reliable for complex datasets.

Improvement Suggestions:

- Address the performance issues with low-resolution images by incorporating varying scales during training and inference.
- Enhance context understanding by introducing additional context branches or increasing input image sizes.
- Explore ensemble or cascade methods for object detection to potentially improve model performance.