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

Link

arXiv

Summary

- This paper further improves R-CNN based object detection algorithm. Instead of using a separate module for generating region proposals, they incorporate it in the CNN itself. For this they introduce a Region proposal Network (RPN) that learns to generate region proposals and objectness score (how likely is the region to be a non-background object) based on CNN features using some additional convolution operations. A fast R-CNN model then classifies and does bounding box regressions on these regions. Since feature computation is shared between region proposals and classification network, the test time is reduced significantly.

- To generate region proposals a small network is slid over the output features of the last shared convolution layer. It takes 3×3 feature map and produces fixed dimensional (512-d for VGG) feature. This is implemented as a 3×3 convolution. This is followed by a two sibling fully connected layers-one that produces the bounding box and another that predicts the probability of the region being an object. Since these also need to be slid over the entire feature map, they are implemented as two 1×1 convolution layer.

- To account for scale and translation invariance, at each sliding window location they simultaneously propose multiple multiple regions. They used 3 scales and 3 aspect ratios resulting in 9 anchors at each sliding position. This ensures the network can detect image at different scale, aspect ratio or location without incurring much extra computational cost since the same feature is shared for all anchors.

- One interesting point is even though we use different scale or aspect ratios to generate proposals the loss does not depend on them (scales/ratios). One way to understand it is that, we use a fixed sized feature map to predict the different sized objects. Anchor boxes tell the regression model what the expected shape of the object is. But the loss function actually looks at the predicted value and original ground truth value. For top-left x-coordinate of the proposal box,if predicted box, anchor box and ground truth boxes are respectively x, x_a and x^* respectively, then the regression loss is $L_reg(t_x, t^*_x)$ where $t_x = (x - x_a)/w_a, t^*_x = (x^* - x_a)/w_a$ and w_a is anchor box width. Expanding the loss gives $||(x - x_a)/w_a - (x^* - x_a)/w_a||^2_2 = ||(x - x^*)/w_a||^2_2$.

- The faster R-CNN achieved state of the art result on PASCAL VOC, MS COCO and ILSVRC competition.