YOLO V3

YOLO v3 maintained its consistent detection speed and improved performance: The input image was a 320×320 image size and could be run in 22 ms. The mAP reached 28.2. This data is the same as SSD, but faster. Three times. On the TitanX, YOLO v3 can be completed in 51ms, and the AP50 value is 57.9. RetinaNet requires 198ms, AP50 is slightly lower, at 57.5.

We don't use softmax to do the classification. Instead, we use an independent logistic to do the classification. The advantage of this approach is that it can handle overlapping multi-label issues, such as Open Image Dataset. Among them, overlapping tags such as Woman and Person appear.

One of YOLO's weaknesses was the lack of multi-scale transformation. Using the ideas in FPN, v3 made predictions on three different scales. On the COCO, we predict three boxes for each scale, so there are a total of nine. Therefore, the size of the output feature map is N×N×[3×(4+1+80)].

Then we take the feature map from two layers, upsample 2x, and merge it with the feature map that was output earlier by element-wide. In this way, we can get more high-level semantic information from the later layers, and can also get fine-grained information (larger feature map, smaller receptive field) from the previous layer. Then it is followed by some processing, and finally get a feature map similar to the above size, only the spatial dimension becomes 2 times.

Using multi-scale prediction, v3's detection results for small targets have become significantly better. However, for medium and large targets, the performance is relatively poor. This is where further work needs to be improved.