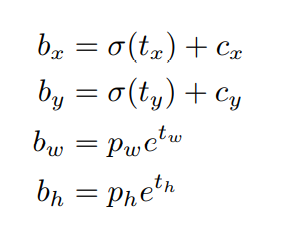
**YOLOv3: An Incremental Improvement**

*Joseph Redmon Ali Farhadi*

* Bounding box coordinates conversion is same as that in v2



* Similar to v2, we have class probabilities per bounding box.
* If an anchor has IoU > 0.5 with a ground-truth box, the anchor gets assigned to that ground-truth box. If no anchor is assigned to a ground-truth box, the object doesn’t have any coordinates or class predictions related loss terms; it has the loss for objectness only.
* For class predictions, use logistic function rather than softmax. Use binary cross entropy loss for class predictions.
* Softmax implies that an object can belong to only one class; i.e. all classes are mutually exclusive. So, it is better to use logistic loss function if you don’t want that constraint.
* Similar to v2, K-means is used to find the anchors
* V3 has 3 anchors, so it predicts 3 boxes per cell.
* For COCO dataset, the output shape becomes N \* N \* [ 3 \* (4+1+80) ], since there are 80 classes.
* The last conv. layer from the classification model that you retain, while converting the model for detection task, generates a conv. feature map. In YOLO v2, we add few conv. layers on top of this, and these layers predict the bounding boxes. However, in v3 there are some changes.

In v3 also, we add some conv. layers onto the conv. feature map to predict bounding boxes. But we do this **two more times** at different feature map scales.

We take the feature map from the 2 layers previous and upsample it by 2\* (i.e. 2 times). We also take feature map from some earlier layer and then concatenate it with the upsampled feature map. Now, this becomes our new conv. feature map for prediction. We add few conv. layers that take this feature map as input and predict bounding boxes.

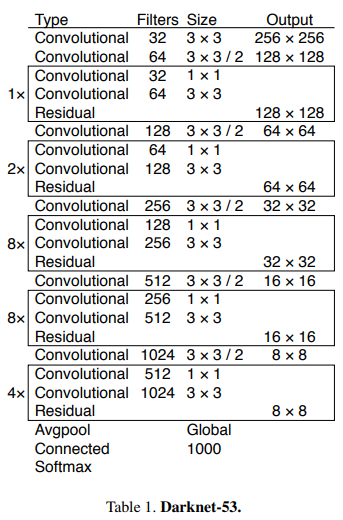
We do similar thing one more time to generate third set of bounding box predictions.

So, basically there are three sets of predictions in YOLO v3, each at different scale.

* V2 used Darkenet-19 as the base classification network. V3 comes up with another, which is hybrid of Darknet-19 and residual network.

So, this network uses 3\*3 and 1\*1 conv. layers, but it also has residual connections.

It has 53 conv. layers, hence called Darknet-53



* Training is similar to v2
* With IoU threshold 0.5, V3 is at par with RetinaNet and better than SSD. If this threshold is increased, mAP drops. So, it means YOLOv3 still struggles to get the bounding boxes perfectly aligned.
* However, YOLO v3 is significantly fast compared to other detection models, including RetinaNet.