YOLOV3

整理：李英杰

[paper]https://pjreddie.com/media/files/papers/YOLOv3.pdf

论文特点：在YOLO上复现一些别的论文的思想。

- 对YOLO进行了一些更新。新的网络比之前要大，但仍然很快。320x320的输入22ms，mAP为28.2，与SSD一样准确，但比它快三倍。When we look at the old .5 IOU mAP detection metric YOLOv3 is quite good. It achieves 57.9 AP50 in 51 ms on a Titan X, compared to 57.5 AP50 in 198 ms by RetinaNet, similar performance but 3.8×faster.

改进：

- 1：使用残差模型
- 2：采用FPN架构

Bounding Box Prediction：

- 每个bounding box有四个坐标，tx，ty，tw，th，如果cell相对左上角的 偏移量为（cx，cy），且bounding box priors宽和高分别为pw，ph，预测为

the output feature map. The network predicts 5 coordinates for each bounding box, t_x, t_y, t_w, t_h, and t_o. If the cell is offset from the top left corner of the image by (c_x, c_y) and the bounding box prior has width and height p_w, p_h, then the predictions correspond to:

$$
\begin{aligned}
b_x &= \sigma(t_x) + c_x \\
b_y &= \sigma(t_y) + c_y \\
b_w &= p_w e^{t_w} \\
b_h &= p_h e^{t_h} \\
Pr(\text{object}) * IOU(b, \text{object}) &= \sigma(t_o)
\end{aligned}
$$

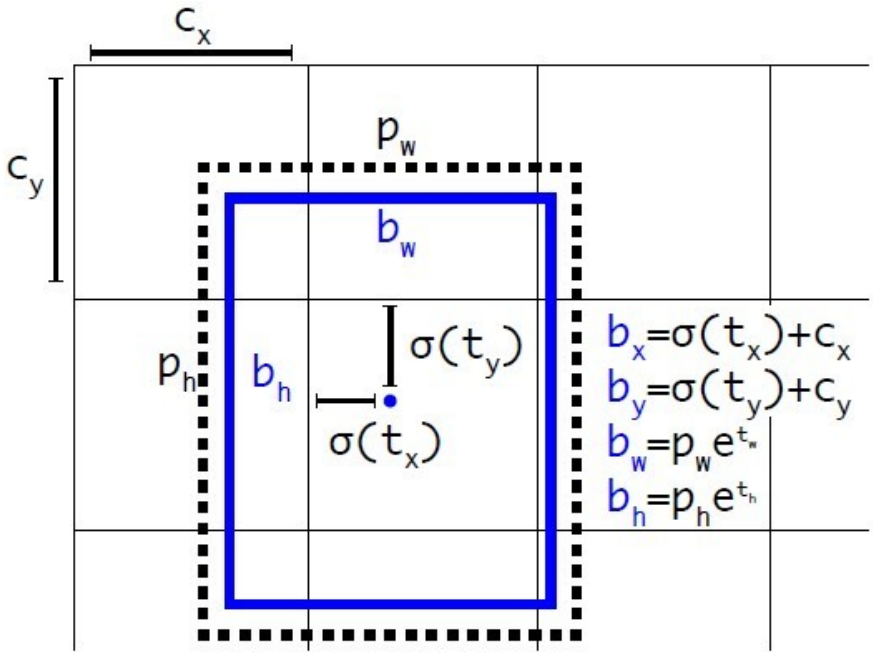

Figure 3: Bounding boxes with dimension priors and location prediction. We predict the width and height of the box as offsets from cluster centroids. We predict the center coordinates of the box relative to the location of filter application using a sigmoid function.

http://blog.csdn.net/hrsstudy

- 训练时使用平方误差之和。 YOLOv3使用**逻辑回归**（logistic regression）的方法给每个bounding box预测一个对象分数。如果bounding box prior比任何一个其他的bounding box prior重叠ground truth都多，这个值应该是1。如果bounding box不是最佳的但是重叠部分比某个阈值高就忽略此次预测，而是按照Faster R-CNN中的方法进行，阈值使用0.5。作者系统只为每个ground truth对象分配一个bounding box prior。如果一个bounding box prior没有分配给ground truth对象，则不会对坐标或者类别预测造成损失，仅会对对象造成损失。

Class Prediction：

每个bounding box预测框中可能包含的物体类别时使用多标签分类（multilabelclassification）。我们没有使用softmax，因为我们发现没有必要，而是使用独立的logistic classifiers。训练时使用二元交叉熵来进行类别预测。

当我们应用到如Open Image Dataset更复杂的领域时，这个方法很有用。此数据集中有很多重叠标签。使用softmax是在每个框只有一个类别的假设下，而通常情况并不是这样。多标签方法能更好地模拟数据。

Predictions Across Scales

- **YOLOv3预测三种不同scale的box。** 系统使用类似金字塔网络的概念从这些尺度中提取特征。在基本的特征提取器中加入了几个卷积层。其中最后一层预测一个三维张量，它由bounding box，objectness，class的预测编码得到。在COCO的实验中，每个scale预测三个boxes，因此张量为NxNx[3*(4+1+80)]，4个bounding box offsets，1个objectness prediction，和80个class predictionss。
- **然后从对两层的特征映射进行2x的上采样。再对原来网络的特征映射和上采样后的特征映射进行合并。** 这种方法使我们能够从上采样的特征和早期特征映射的细化信息中获得更有意义的语义信息。 然后，我们再添加一些卷积层来处理这种组合的特征映射，一种类似于张量的类似张量，只不过是原来的两倍。
- **采用相同的设计来预测最终尺寸的方框。** Thus our predictions for the 3rd scale benefit from all the prior computation as well as finegrained features from early on in the network.
- **使用k-means clustering来确定bounding box priors。** 选择了9个cluster和3个scales，然后在整个scales上均匀分割clusters。在COCO数据集上，9个clusters是：(10×13),(16×30),(33×23),(30×61),(62×45),(59× 119),(116×90),(156×198),(373×326).

Feature Extractor

- YOLOv3的特征提取器是一个残差模型，因为包含53个卷积层，所以称为Darknet-53，
  从网络结构上看，相比Darknet-19网络使用了残差单元，所以可以构建得更深。

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

图14：YOLOv3所用的Darknet-53模型

- 采用FPN架构（Feature Pyramid Networks for Object Detection）来实现多尺度检测。YOLOv3采用了3个尺度的特征图（当输入为416 * 416时）：(13 * 13)，(26 * 26),(52 * 52）。
  VOC数据集上的YOLOv3网络结构如下图所示：

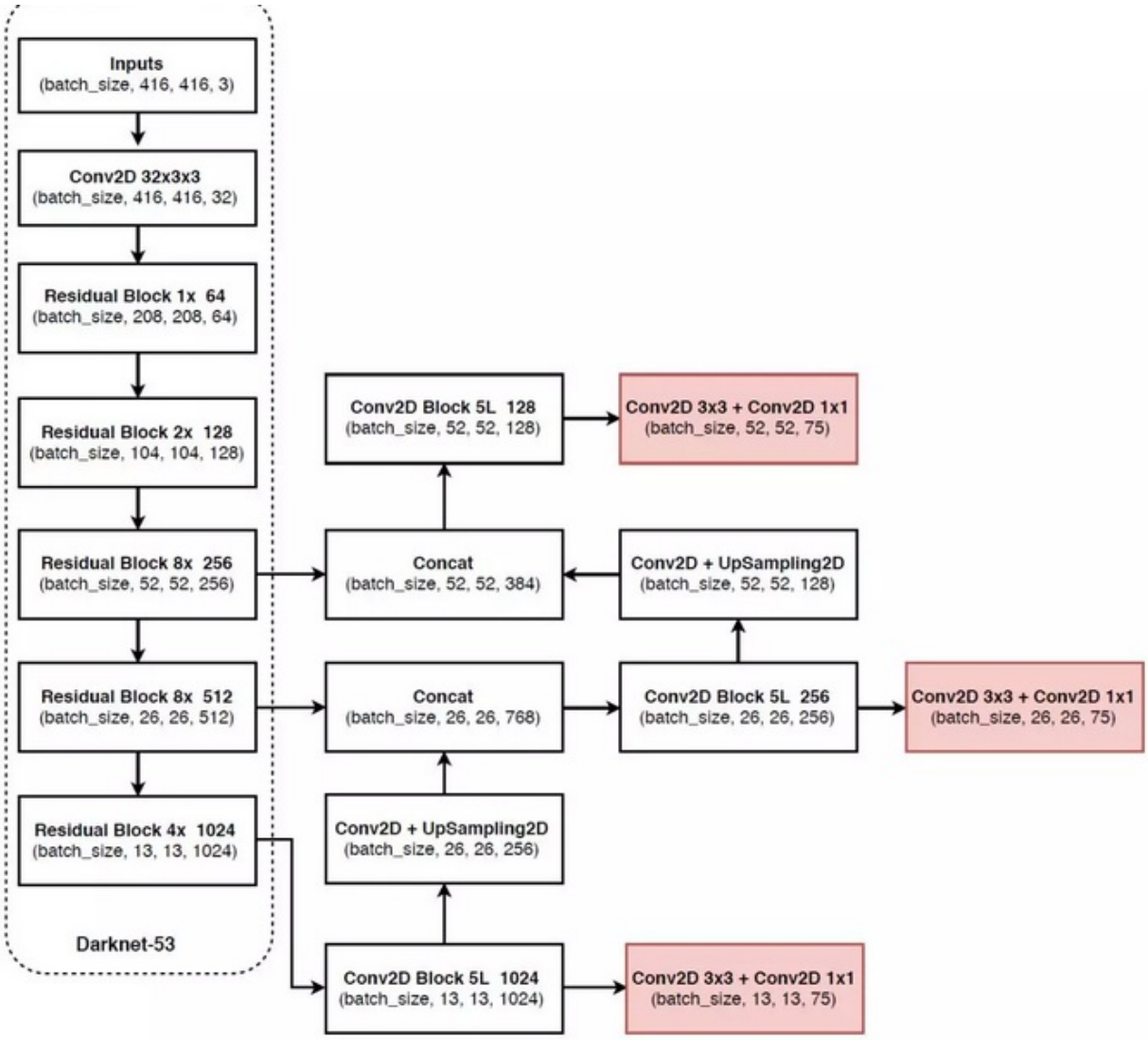

图15 YOLOv3网络结构示意图（VOC数据集）

其中红色部分为各个尺度特征图的检测结果。YOLOv3每个位置使用3个先验框，所以使用k-means得到9个先验框，并将其划分到3个尺度特征图上，尺度更大的特征图使用更小的先验框，和SSD类似。

Traing

作者训练的时候仍然采用完整的图片进行训练，使用不同的scale、大量的data argumentation、batch normalization等很多trick。

实验结果：

YOLOv3性能很好。在COCO数据集上，它的mAP与SSD及其变体相当，但比它快三倍。与其它模型如RetinaNet相比还是稍显落后。

然而，当我们考察“旧”的检测度量时IOU = 0.5（即图表中的AP50）时的mAP，YOLOv3非常强大。 它几乎与RetinaNet相当，并且远高于SSD变体。这表明YOLOv3是一款非常强大的检测器，能够较好地为目标预测边界框。 随着IOU阈值增加，性能下降明显。

老版本的YOLO在检测小物体上有困难，使用了多尺度预测后，YOLOv3有相对高的APs值。但是它在中等或较大目标上的性能较差。要解决这个问题还需要进一步研究。

画出AP50度量下的准确率和速度曲线，可以看出YOLOv3与其它检测系统相比性能更好。

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods							
Faster R-CNN+++ [3]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [6]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [4]	Inception-ResNet-v2 [19]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [18]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
One-stage methods							
YOLOv2 [13]	DarkNet-19 [13]	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [9, 2]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [2]	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet [7]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet [7]	ResNeXt-101-FPN	**40.8**	**61.1**	**44.1**	**24.1**	**44.2**	51.2
YOLOv3 608 × 608	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9

Table 3. I'm seriously just stealing all these tables from [7] they take soooo long to make from scratch. Ok, YOLOv3 is doing alright. Keep in mind that RetinaNet has like 3.8× longer to process an image. YOLOv3 is much better than SSD variants and comparable to state-of-the-art models on the AP_{50} metric.

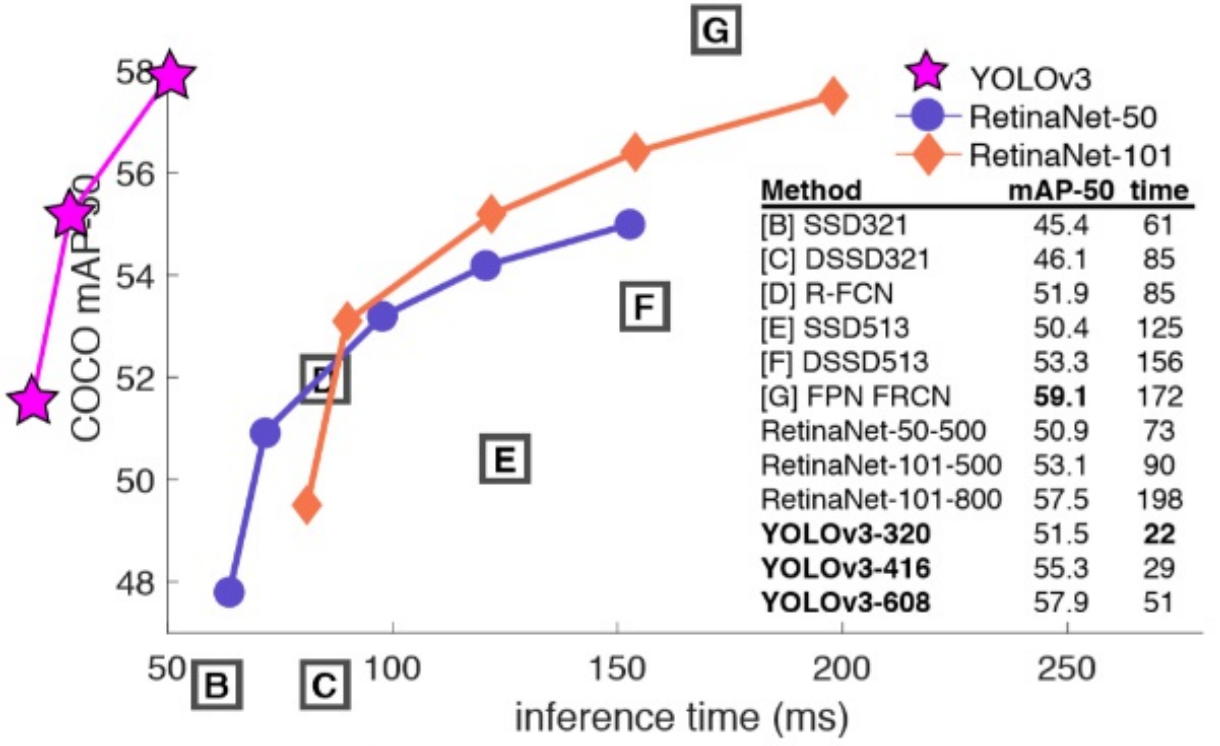

Figure 3. Again adapted from the [7], this time displaying speed/accuracy tradeoff on the mAP at .5 IOU metric. You can tell YOLOv3 is good because it's very high and far to the left. Can you cite your own paper? Guess who's going to try, this guy → [14].

Things We Tried That Didn't Work

Anchor box x,y offset predictions.

尝试使用常规的anchor box预测机制，即使用线性激活函数将x,y的偏移量预测为边界框宽度或高度的倍数，这个方法降低了模型的稳定性，效果不佳。

Linear x,y predictions instead of logistic.

尝试使用线性激活函数直接预测x,y的偏移量来代替logistic activation，这会使用mAP有一些下降。

Focal loss.

尝试使用focal loss，这使mAP下降约2个百分点。YOLOv3可能对focal loss试图解决的问题已经很鲁棒，因为它有单独的objectness predictions和conditional class predictions。可能是对多数样例，没有来自class predictions的误差，不太确定。

Dual IOU thresholds and truth assignment.

Faster R-CNN训练时使用了两个IOU阈值。如果和真实值重叠超过0.7则为正例，0.3到0.7忽略，0.3以下为反例。作者尝试了类似的策略但效果不好。

结论：

从YOLO的三代变革中可以看到，在目标检测领域比较好的策略包含：设置先验框，采用全卷积做预测，采用残差网络，采用多尺度特征图做预测。