YOLOv3 论文笔记

1、解决什么问题

多尺度预测 （类 FPN）
更好的基础分类网络（类 ResNet）和分类器

2、使用什么方法

- bounding box 预测

使用维度聚类（dimension cluster）作为 anchor box 来预测边界框（bounding box）。
每个边界框 4 个参数（tx,ty,tw,th），如果边界框相对于图片左上角偏移（cx,cy）并且前面的边界框（bounding box prior？边界框先验？）大小为（pw, ph），那么对边界框的位置的预测为：

$$
\begin{aligned}
b_x &= \sigma(t_x) + c_x \\
b_y &= \sigma(t_y) + c_y \\
b_w &= p_w e^{t_w} \\
b_h &= p_h e^{t_h}
\end{aligned}
$$

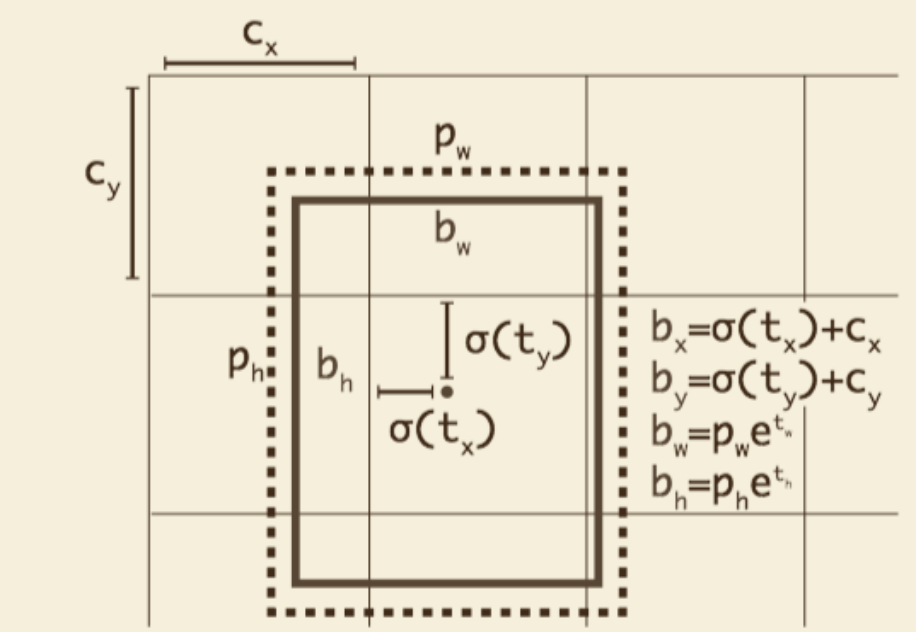


Figure 2. **Bounding boxes with dimension priors and location prediction.** We predict the width and height of the box as offsets from cluster centroids. We predict the center coordinates of the box relative to the location of filter application using a sigmoid function. This figure blatantly self-plagiarized from [15].

loss 使用的是均方误差（squared error）；

使用逻辑回归预测每个边框里面对象的分数，若某个边界框与真实值的相似度大于别的边界

框，那么这个分数就是 1。设置了一个阈值 0.5，当边界框不是与真实值重合得最好的但是大于该阈值时，不进行预测。只为每一个真实值（ground truth）分配一个边界框，如果没有将边界框分配给一个真实值，只会导致 objectness 的 loss 而不影响 coordinate 和 prediction 的 loss。

- 类别预测

每个框使用多个分类标签预测边界框可能包含的类。 这里分类的激活函数不使用 softmax，因为我们发现如果想要达到良好的性能，softmax 不是必要的，只是使用独立的逻辑分类器就可以达到。在训练期间，使用二元交叉熵损失函数进行类预测。

当迁移到更复杂的领域（如 Open Images Dataset [7]）时，这个公式会有所帮助。 在这个数据集中有许多重叠的标签（即女人和人）。 使用 softmax 假设每个盒子只有一个类别，而实际情况通常并非如此。 多标签方法可以更好地模拟数据。

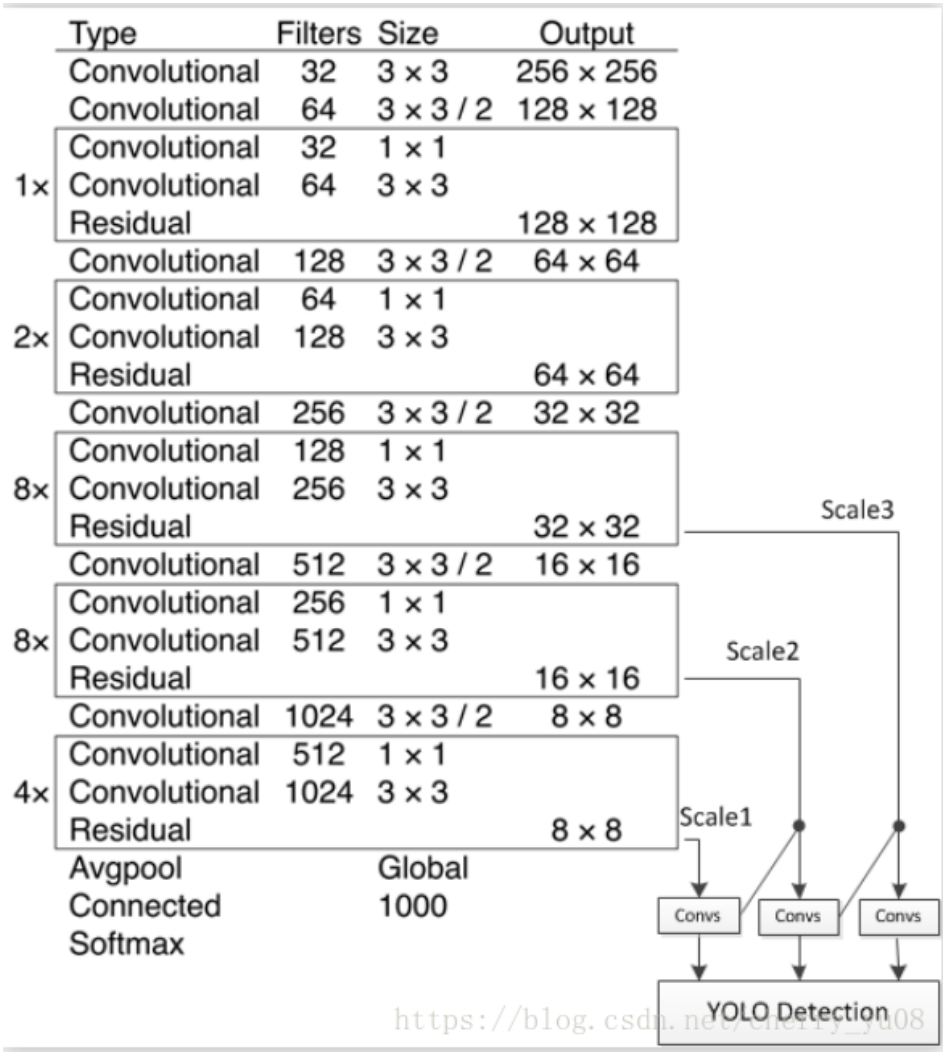
- 多尺度预测（Predictions Across Scales）

YOLOv3 在三个不同的尺度预测 box。使用 Open Image 数据集来提取这三个不同尺度的特征。

尺度 1: 在基础网络之后添加一些卷积层再输出 box 信息.

尺度 2: 从尺度 1 中的倒数第二层的卷积层上采样(x2)再与最后一个 16x16 大小的特征图相加，再次通过多个卷积后输出 box 信息.相比尺度 1 变大两倍.

尺度 3: 与尺度 2 类似,使用了 32x32 大小的特征图.



使用 k-means 聚类来确定边界框：任意选择 9 个 cluster 和三个尺度，然后在尺度上均匀的划分 cluster。在 COCO 数据集中的 9 个 cluster 为：

$$(10 \times 13), (16 \times 30), (33 \times 23), (30 \times 61), (62 \times 45), (59 \times 119), (116 \times 90), (156 \times 198), (373 \times 326).$$

这个 cluster 表示什么？

- 特征提取

特征提取使用的是另一个全新的网络，该网络是 YOLOv2、Darknet-19 和 Resnet 的混合体，一共有 53 个卷积层，所以称之为 Darknet53。

	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			

Table 1. **Darknet-53.**

Darknet53 比 Darknet19 更强大，比 Resnet101 和 Resnet102 效率更高：

Backbone	Top-1	Top-5	Bn Ops	BFLOP/s	FPS
Darknet-19 [15]	74.1	91.8	7.29	1246	**171**
ResNet-101[5]	77.1	93.7	19.7	1039	53
ResNet-152 [5]	**77.6**	**93.8**	29.4	1090	37
Darknet-53	77.2	**93.8**	18.7	**1457**	78

Table 2. **Comparison of backbones.** Accuracy, billions of operations, billion floating point operations per second, and FPS for various networks.

以 256*256 大小的图片测试，
Darknet-53 的性能与最先进的分类器相当，但浮点运算更少，速度更快，
Darknet-53 比 ResNet-101 更好，速度提高 1.5 倍。 Darknet-53 具有与 ResNet-152 相似的性能，速度提高了 2 倍。

- 训练

直接输入完整的图像进行训练，训练过程中加入了以下操作：
multi-scale
data augmentation
batch normalization
等等？
使用 Darknet 神经网络框架进行培训和测试。

3、相关工作

作者做了一些工作但是效果都不太好，比如对于偏移量 x，y 的预测使用线性方法而非回归方法等。

4、效果

- 对于 320*320 的输入，用时 22ms，达到 28.2mAP，准确率与 SSD 一样，但是速度快三倍。
- 在 Titan X 上运行达到 57.9 AP(50)用时 51ms，相对而言，RetinaNet 达到 57.5 AP(50)用时 198ms，一样的效果，YOLOv3 快 3.8 倍。

5、还存在什么问题

- YOLOv3 是一个非常强大的 detector，擅长为物体产生合适的 box。但是随着 IOU 阈值的增加，性能显着下降，表明 YOLOv3 努力让盒子与物体完美对齐。

- 对于小物体的检测效果以及有了提高，通过多尺度的预测提高了 APS 性能，但是在中型和大型物体上性能较差（为什么？）

作者在论文最后说的话，必须 po 出来

5. What This All Means

YOLOv3 is a good detector. It's fast, it's accurate. It's not as great on the COCO average AP between .5 and .95 IOU metric. But it's very good on the old detection metric of .5 IOU.

Why did we switch metrics anyway? The original COCO paper just has this cryptic sentence: "A full discussion of evaluation metrics will be added once the evaluation server is complete". Russakovsky et al report that that humans have a hard time distinguishing an IOU of .3 from .5! "Training humans to visually inspect a bounding box with IOU of 0.3 and distinguish it from one with IOU 0.5 is surprisingly difficult." [18] If humans have a hard time telling the difference, how much does it matter?

But maybe a better question is: "What are we going to do with these detectors now that we have them?" A lot of the people doing this research are at Google and Facebook. I guess at least we know the technology is in good hands and definitely won't be used to harvest your personal information and sell it to.... wait, you're saying that's exactly what it will be used for?? Oh.

Well the other people heavily funding vision research are the military and they've never done anything horrible like killing lots of people with new technology oh wait.....[1]

I have a lot of hope that most of the people using computer vision are just doing happy, good stuff with it, like counting the number of zebras in a national park [13], or tracking their cat as it wanders around their house [19]. But computer vision is already being put to questionable use and as researchers we have a responsibility to at least consider the harm our work might be doing and think of ways to mitigate it. We owe the world that much.

In closing, do not @ me. (Because I finally quit Twitter).