YOLO v3详解

全卷积神经网络
YOLO仅是用卷积层，所以它是全卷积网络（FCN）。它具有75个卷积层，具有跳过连接和上采样层。不使用任何形式的池化，使用具有步幅为2的卷积层来下采样特征图。这有助于防止由于池化导致低级特征的丢失。

网络通过称为网络步幅的因子对图像进行下采样。例如，如果网络的步幅为32，则尺寸为416 x 416的输入图像将产生尺寸为13 x 13的输出。一般而言，网络中任何层的步幅等于该层的输出的尺寸比网络的输入图像的尺寸小的倍数。

解析输出
典型地，（对于所有目标检测器都适用）卷积层学习的特征会被传递到进行检测预测（边界框的坐标，类标签等）的分类器/回归器。

在YOLO中，预测是通过使用1×1卷积的卷积层完成的。

现在，首先要注意的是我们的输出是一个特征图。由于我们使用了1 x 1卷积，所以预测图的大小恰好是之前的特征图的大小。在YOLO v3（以及它的后续版本）中，这个预测图的每个单元格可以预测固定数量的边界框。

深度方面，我们在特征图中有（B x（5 + C））个条目。 B表示每个单元可以预测的边界框的数量。根据该论文，这B个边界框中的每一个可能专门检测某种目标。每个边界框都有5 + C个属性，分别描述每个边界框的中心坐标，尺寸，目标分数（objectness score）和C个类的置信度。 YOLO v3为每个单元格预测3个边界框。

如果目标的中心位于单元格的感受野中，则希望特征图中的单元格可以通过其中一个边界框来预测目标。

锚框
预测边界框的宽度和高度可能是有意义的，但实际上，这会导致训练期间的梯度变得不稳定。因此，现在大多数目标检测器预测对数空间变换，或简单地预测与预定义的默认边界框（称作锚）之间的偏移。

然后，对锚框进行这些变换以获得预测。 YOLO v3有三个锚，可以为每个单元预测三个边界框。

中心坐标
注意我们正在通过一个sigmoid函数来预测中心坐标，它迫使输出的值压缩在0和1之间。为什么会这样呢?

通常情况下，YOLO不预测边界框中心的绝对坐标。它预测的偏移：
相对于负责预测目标的网格单元的左上角。
通过特征图中的单元的维度，即1，进行归一化。

例如狗的图像。如果预测的中心是（0.4,0.7），那么这意味着中心位于13×13特征图上的（6.4,6.7）。（因为红色单元的左上坐标是（6,6））。

但是如果预测的x，y坐标大于1，会发生什么情况，比如（1.2,0.7）。这意味着它的中心位于（7.2,6.7）。注意现在中心位于红色单元，即第7排的第8个单元的右侧。这打破了YOLO背后的理论，因为如果我们假设红色框负责预测狗，狗的中心必须位于红色单元中，而不是位于红色单元旁。

因此，为了解决这个问题，将输出通过一个sigmoid函

数，该函数把输出缩小至0到1的范围内，有效地将中心保持在预测的网格单元中。
预测的结果bw和bh通过图像的高度和宽度进行归一化。因此，如果包含狗的框的预测值bw和bh分别为0.3和0.8，则13 x 13特征图上的实际宽度和高度为（13 x 0.3,13 x 0.8）。

目标分数
目标分数表示边界框内包含目标的概率。红色及其相邻网格的目标分数应该接近1，而位于角落的网格接近0。

目标分数也使用sigmoid函数来压缩数值，因为它被定义为一个概率。

类别置信度
类别置信度表示检测目标属于特定类别（狗，猫，香蕉，汽车等）的概率。在YOLOv3之前的版本，YOLO对类别置信度使用softmax。

但是，该设计选择已经在YOLO v3中被舍弃了，作者选择使用sigmoid。原因是对类别分数进行softmax意味着类别之间是互相排斥的。简而言之，如果一个目标属于一个类，那么它就不能属于另一个类。这对于COCO数据库来说是正确的，而我们将在COCO数据库上训练检测器。

但是，当存在“女性”和“人”这样的类别时，这种假设可能不成立。这就是作者们避免使用Softmax激活的原因。

在不同尺度上进行预测
YOLO v3可以进行3种不同尺度的预测。检测层在分别具

有步幅32,16,8的三种不同尺寸的特征图上进行检测。这意味着，在输入416 x 416的情况下，我们在尺寸为13 x 13,26 x 26和52 x 52上进行检测。

网络对输入图像进行下采样直到第一个检测层，它使用具有步幅为32的层的特征图进行检测。然后，将层上采样2倍，并与具有相同特征图尺寸的前一层的特征图连接。现在在具有步幅16的层上进行另一次检测，重复相同的上采样过程。并且在步幅8的层处进行最终检测。

在每个尺度上，每个单元使用3个锚来预测3个边界框，锚的总数为9（在不同尺度上的锚是不同的）
作者声称这有助于YOLO v3更好地检测小的物体，小型物体检测是困扰YOLO早期版本的问题。 上采样可以帮助网络学习有助于小物体检测的细粒度特征。

输出处理
对于尺寸为416×416的图像，YOLO预测（（52×52）+（26×26）+ 13×13））×3 = 10647个边界框。 但是，在我们的图像中，只有一个物体——一只狗。

目标置信度的阈值处理
首先，我们根据目标分数过滤边界框框。 通常，具有低于阈值分数的框会被删除。

非最大值抑制
NMS的作用是解决同一图像的的多重检测问题。 例如，红色网格单元的3个边界框可能检测到同一个框，或者相邻单元可能检测到相同的目标。