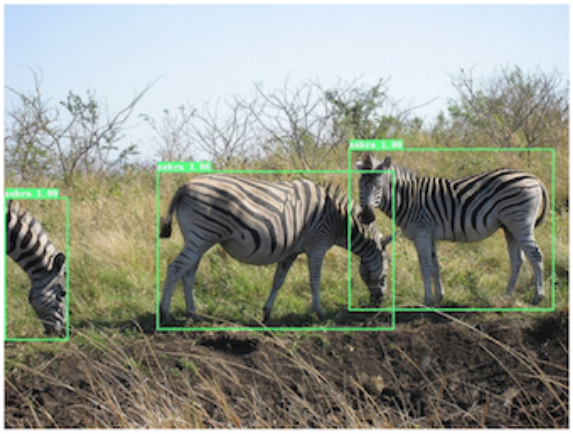
对计算机而言，能够“看到”的是图像被编码之后的数字，但它很难理解高层语义概念，比如图像或者视频帧中出现的目标是人还是物体，更无法定位目标出现在图像中哪个区域。目标检测的主要目的是让计算机可以自动识别图片或者视频帧中所有目标的类别，并在该目标周围绘制边界框，标示出每个目标的位置，如 下图所示。

(a) 分类：动物或者斑马



(b) 检测：准确检测出每个斑马在图上出现的位置

目标检测发展历程

在上一节中我们学习了图像分类处理基本流程，先使用卷积神经网络提取图像特征，然后再用这些特征预测分类概率，根据训练样本标签建立起分类损失函数，开启端到端的训练，如 下图所示。

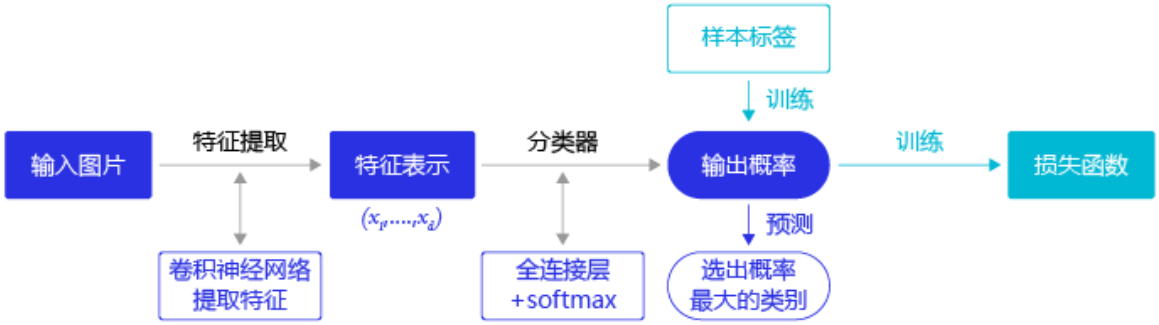


图2：图像分类流程示意图

为了解决这个问题，结合图片分类任务取得的成功经验，我们可以将目标检测任务进行拆分。假设我们现在有某种方式可以在输入图片上生成一系列可能包含物体的区域，这些区域称为候选区域，在一张图上可以生成很多个候选区域。然后对每个候选区域，可以把它单独当成一幅图像来看待，使用图像分类模型对它进行分类，看它属于哪个类别或者背景（即不包含任何物体的类别）。

上一节我们学过如何解决图像分类任务，使用卷积神经网络对一幅图像进行分类不再是一件困难的事情。那么，现在问题的关键就是如何产生候选区域？比如我们可以使用穷举法来产生候选区域，如图3所示。

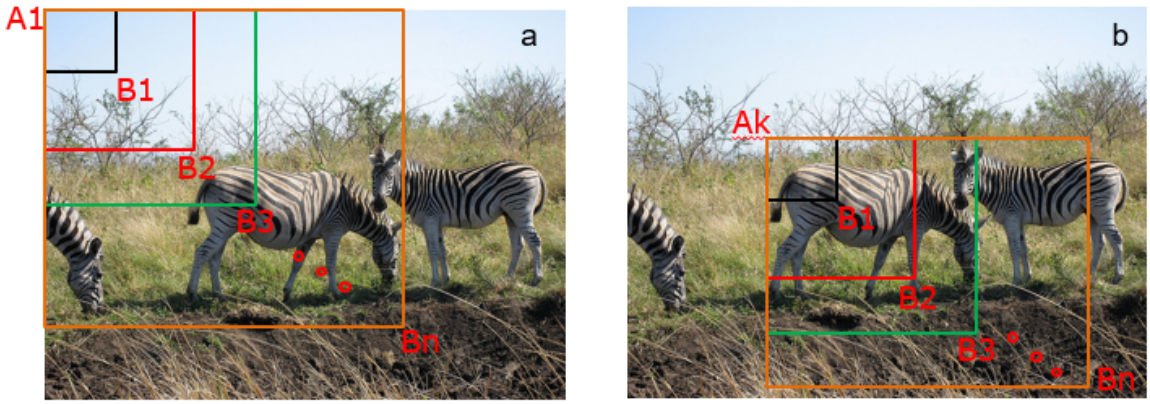


图3：候选区域

- 如图3（a）所示：A在图片左上角位置，B遍历除A之外的所有位置，生成矩形框A1B1, ..., A1Bn, ...
- 如图3（b）所示：A在图片中间某个位置，B遍历A右下方所有位置，生成矩形框AkB1, ..., AkBn, ...

当A遍历图像上所有像素点，B则遍历它右下方所有的像素点，最终生成的矩形框集合{AiBj}将会包含图像上所有可以选择的区域。

只要我们对每个候选区域的分类足够的准确，则一定能找到跟实际物体足够接近的区域来。穷举法也许能得到正确的预测结果，但其计算量也是非常巨大的，其所生成的总候选区域数目约为$\frac{W^2H^2}{4}$，假设$H=W=100$，总数将会达到2.5×10^7个，如此多的候选区域使得这种方法几乎没有什么实用性。但是通过这种方式，我们可以看出，假设分类任务完成的足够完美，从理论上来讲检测任务也是可以解决的，亟待解决的问题是如何设计出合适的方法来产生候选区域。

R-CNN，Fast-RCNN的系列算法分成两个阶段，先在图像上产生候选区域，再对候选区域进行分类并预测目标物体位置，它们通常被叫做两阶段检测算法。 YOLO算法则只使用一个网络同时产生候选区域并预测出物体的类别和位置，所以它们通常被叫做单阶段检测算法。

目标检测基础概念

在介绍目标检测算法之前，先介绍一些跟检测相关的基本概念，包括边界框、锚框和交并比等。

边界框（bounding box）

检测任务需要同时预测物体的类别和位置，因此需要引入一些跟位置相关的概念。通常使用边界框（bounding box，bbox）来表示物体的位置，边界框是正好能包含物体的矩形框，如 **图4** 所示，图中3个人分别对应3个边界框。

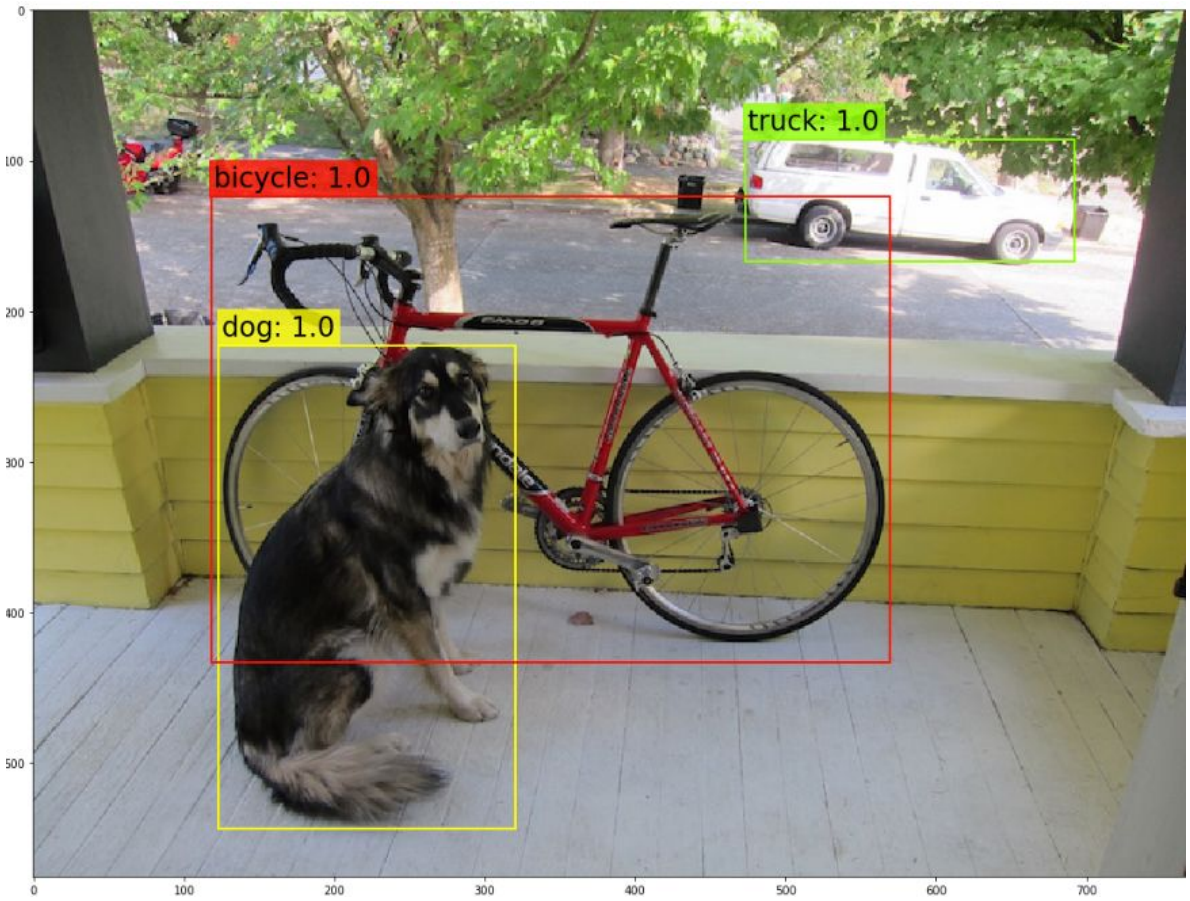
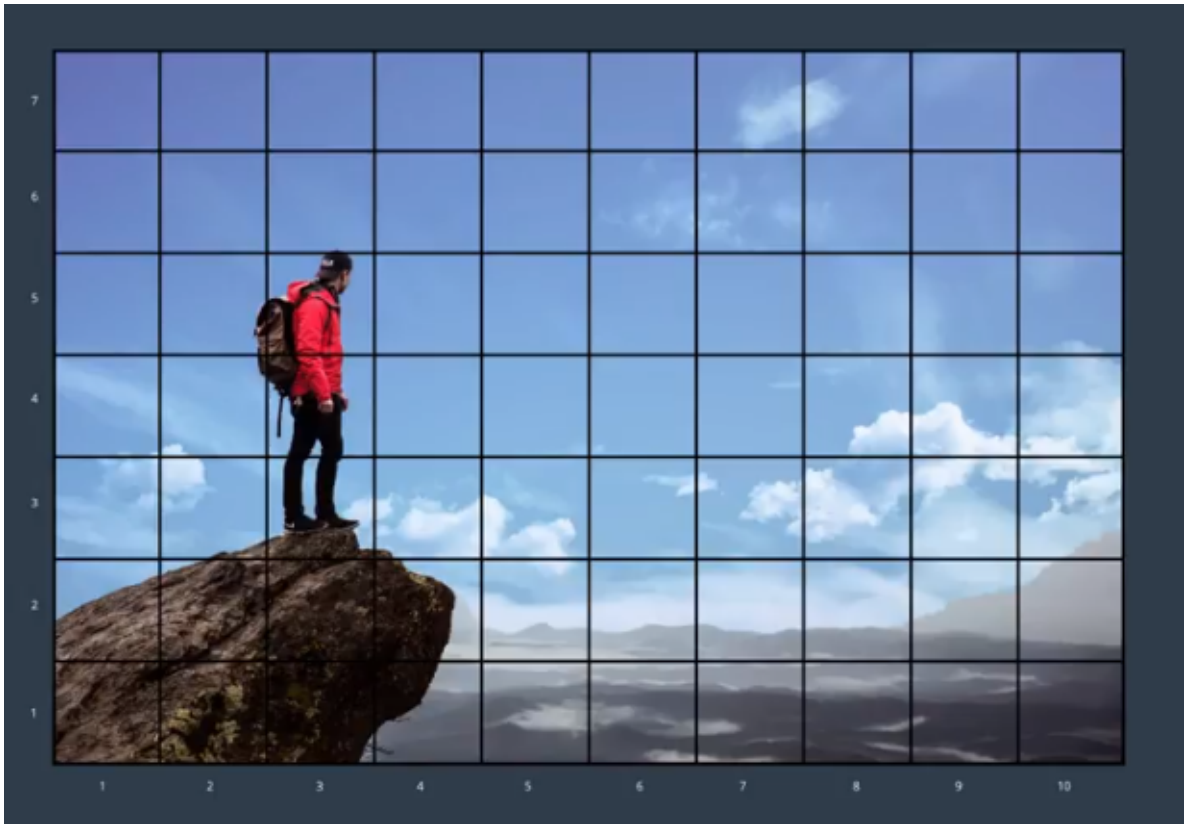


图4：边界框

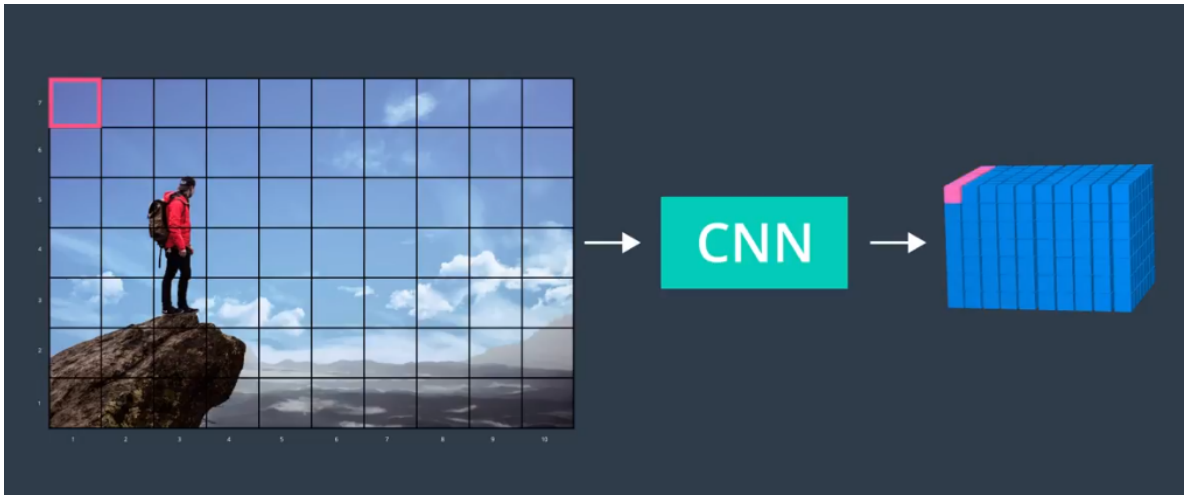
通常有两种格式来表示边界框的位置：

1. $xyxy$，即(x_1, y_1, x_2, y_2)，其中(x_1, y_1)是矩形框左上角的坐标，(x_2, y_2)是矩形框右下角的坐标。图4中3个红色矩形框用xyxy格式表示如下：
2. $xywh$，即(x, y, w, h)，其中(x, y)是矩形框中心点的坐标，w是矩形框的宽度，h是矩形框的高度。

在检测任务中，训练数据集的标签里会给出目标物体真实边界框所对应的(x_1, y_1, x_2, y_2)，这样的边界框也被称为真实框（ground truth box），模型会对目标物体可能出现的位置进行预测，由模型预测出的边界框则称为预测框（prediction box）。



要完成一项检测任务，我们通常希望模型能够根据输入的图片，输出一些预测的边界框，以及边界框中所包含的物体的类别或者说属于某个类别的概率，例如这种格式: $[C, L, P, x_1, y_1, x_2, y_2]$，其中$L$是类别标签，$P$是物体属于该类别的概率。



交并比（IoU）

假设两个矩形框A和B的位置分别为：

A：$[x_{a1}, y_{a1}, x_{a2}, y_{a2}]$

B：$[x_{b1}, y_{b1}, x_{b2}, y_{b2}]$

假如位置关系如 **图6** 所示：

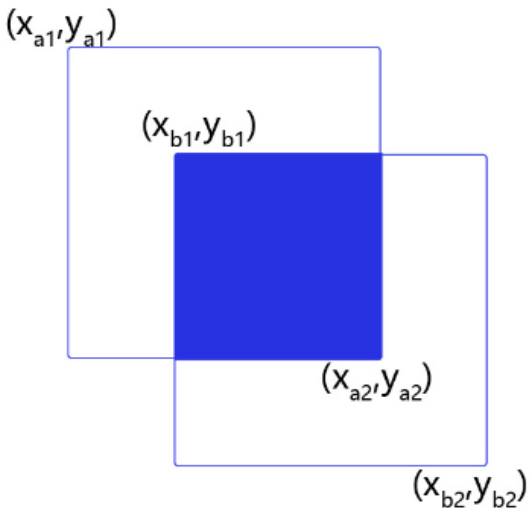


图6：计算交并比

计算交并比：

$IoU = \frac{intersection}{union}$

思考：

两个矩形框之间的相对位置关系，除了上面的示意图之外，还有哪些可能，上面的公式能否覆盖所有的情形？

为了直观的展示交并比的大小跟重合程度之间的关系，**图7** 示意了不同交并比下两个框之间的相对位置关系，从 IoU = 0.95 到 IoU = 0.

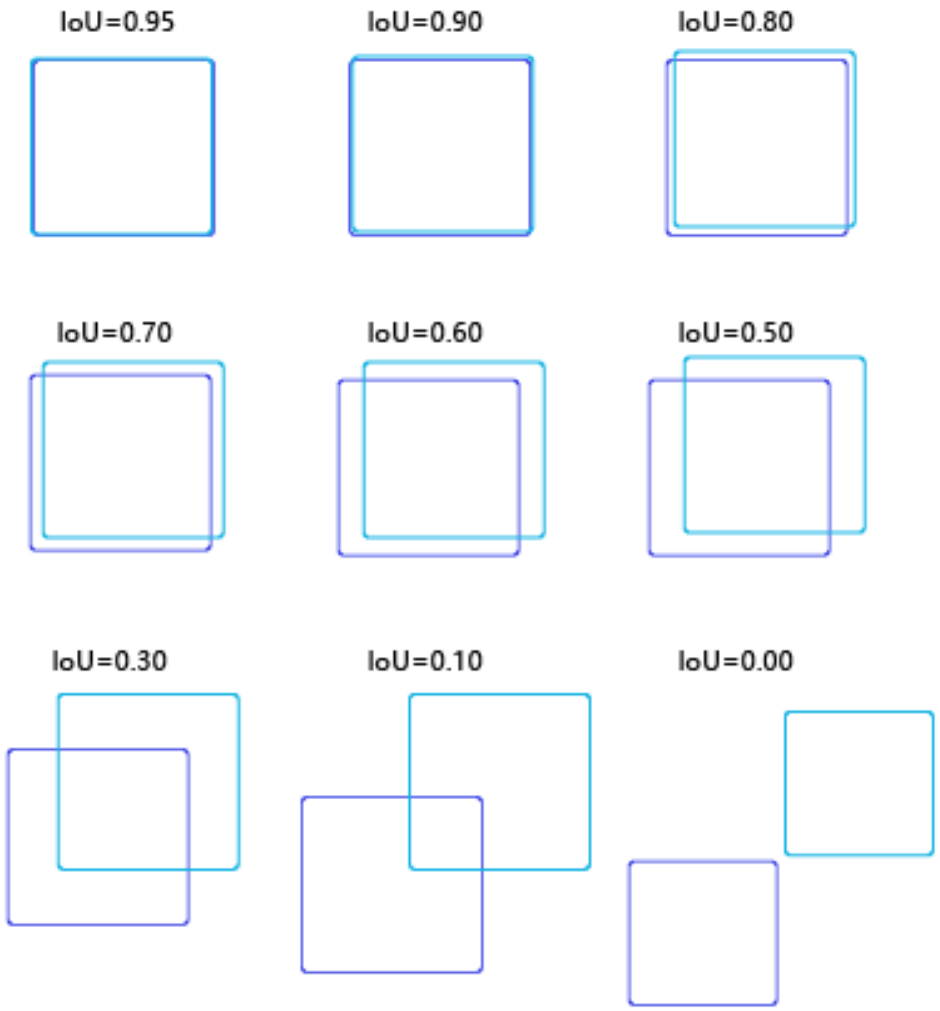
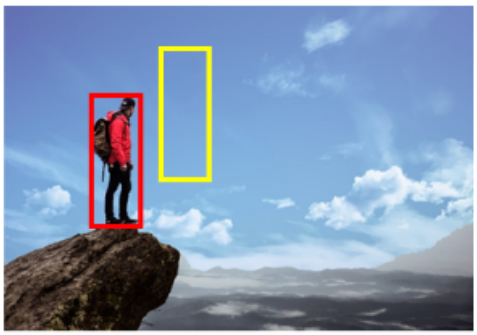


图7：不同交并比下两个框之间相对位置示意图

问题：

1. 什么情况下两个矩形框的IoU等于1？
2. 什么情况下两个矩形框的IoU等于0？

$$IOU = \frac{2000}{2000} = 1$$



$$IOU = \frac{0}{4000} = 0$$

损失函数

1. 是否包含物体的损失函数
2. 物体类别的损失函数
3. 物体位置的损失函数

非极大值抑制算法（non maximum suppression, NMS），这个算法不单单是针对Yolo算法的，而是所有的检测算法中都会用到。NMS算法主要解决的是一个目标被多次检测的问题，如图11中人脸检测，可以看到人脸被多次检测，但是其实我们希望最后仅仅输出其中一个最好的预测框，

那么可以采用NMS算法来实现这样的效果：首先从所有的检测框中找到置信度最大的那个框，然后挨个计算其与剩余框的IOU，如果其值大于一定阈值（重合度过高），那么就将该框剔除；然后对剩余的检测框重复上述过程，直到处理完所有的检测框。Yolo预测过程也需要用到NMS算法。