目标检测中的评价体系

# 摘要

在训练YOLO v2的过程中，系统会显示出一些评价训练效果的值，如Recall，IoU等等。为了怕以后忘了，现在把自己对这几种度量方式的理解记录一下。

这一文章首先假设一个测试集，然后围绕这一测试集来介绍这几种度量方式的计算方法。

关键词：召回率（Recall），精确率（Precision），平均正确率（AP），交除并（IoU）

# 一、数据集（大雁与飞机）

假设现在有这样一个测试集，测试集中的图片只由大雁和飞机两种图片组成，如下图所示：



假设你的分类系统最终的目的是：能取出测试集中所有飞机的图片，而不是大雁的图片。

现在做如下的定义：

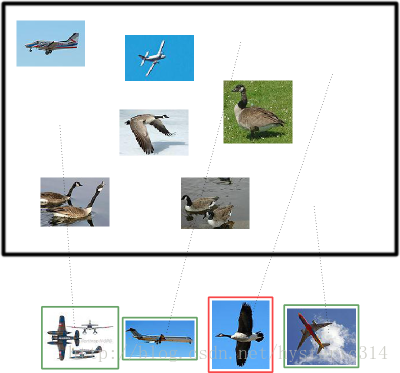
True positives : 飞机的图片被正确的识别成了飞机。

True negatives: 大雁的图片没有被识别出来，系统正确地认为它们是大雁。

False positives: 大雁的图片被错误地识别成了飞机。

False negatives: 飞机的图片没有被识别出来，系统错误地认为它们是大雁。

假设你的分类系统使用了上述假设识别出了四个结果，如下图所示：



那么在识别出的这四张照片中：

True positives : 有三个，画绿色框的飞机。

False positives: 有一个，画红色框的大雁。

没被识别出来的六张图片中：

True negatives : 有四个，这四个大雁的图片，系统正确地没有把它们识别成飞机。

False negatives: 有两个，两个飞机没有被识别出来，系统错误地认为它们是大雁。

# 二、Precision 与 Recall

Precision其实就是在识别出来的图片中，True positives所占的比率：

gfjh

其中的n代表的是(True positives + False positives)，也就是系统一共识别出来多少照片 。

在这一例子中，True positives为3，False positives为1，所以Precision值是 3/（3+1）=0.75。

意味着在识别出的结果中，飞机的图片占75%。

Recall 是被正确识别出来的飞机个数与测试集中所有飞机的个数的比值：

DGHFD

Recall的分母是(True positives + False negatives)，这两个值的和，可以理解为一共有多少张飞机的照片。

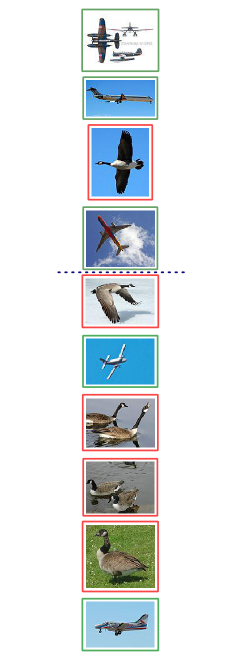
在这一例子中，True positives为3，False negatives为2，那么Recall值是 3/（3+2）=0.6。

意味着在所有的飞机图片中，60%的飞机被正确的识别成飞机.。

# 三、调整阈值

你也可以通过调整阈值，来选择让系统识别出多少图片，进而改变Precision 或 Recall 的值。

在某种阈值的前提下（蓝色虚线），系统识别出了四张图片，如下图中所示：

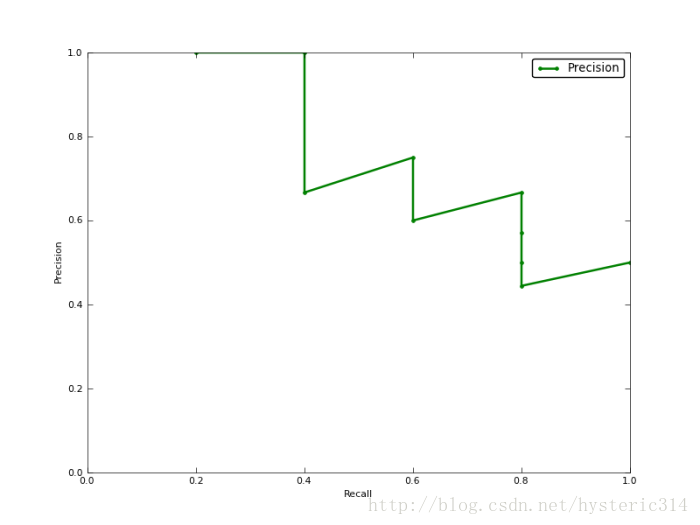


分类系统认为大于阈值（蓝色虚线之上）的四个图片更像飞机。

我们可以通过改变阈值（也可以看作上下移动蓝色的虚线），来选择让系统识别能出多少个图片，当然阈值的变化会导致Precision与Recall值发生变化。比如，把蓝色虚线放到第一张图片下面，也就是说让系统只识别出最上面的那张飞机图片，那么Precision的值就是100%，而Recall的值则是20%。如果把蓝色虚线放到第二张图片下面，也就是说让系统只识别出最上面的前两张图片，那么Precision的值还是100%，而Recall的值则增长到是40%。

# 四、Precision-recall 曲线

如果你想评估一个分类器的性能，一个比较好的方法就是：观察当阈值变化时，Precision与Recall值的变化情况。如果一个分类器的性能比较好，那么它应该有如下的表现：被识别出的图片中飞机所占的比重比较大，并且在识别出大雁之前，尽可能多地正确识别出飞机，也就是让Recall值增长的同时保持Precision的值在一个很高的水平。而性能比较差的分类器可能会损失很多Precision值才能换来Recall值的提高。通常情况下，文章中都会使用Precision-recall曲线，来显示出分类器在Precision与Recall之间的权衡。



具体网址：

<http://blog.csdn.net/syoung9029/article/details/56276567>