

图解YOLO



晓雷

529 人赞了该文章



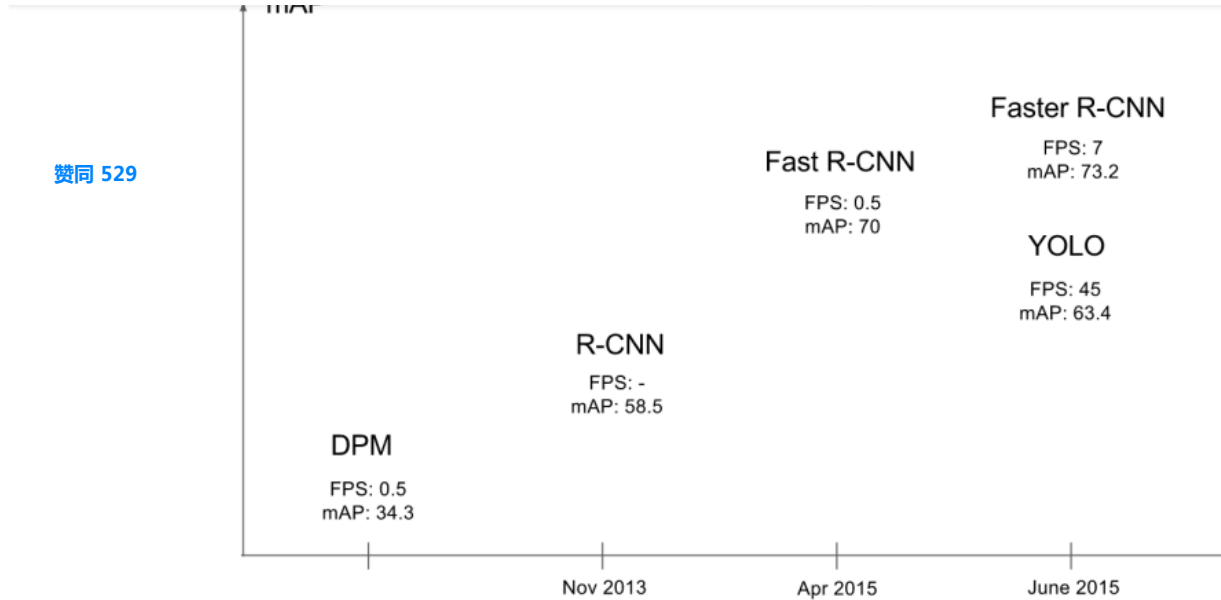
首发于
晓雷机器学习笔记

关注他

YOLO核心思想：从R-CNN到Fast R-CNN一直采用的思路是proposal+分类（proposal 提供位置信息，分类提供类别信息）精度已经很高，但是速度还不行。YOLO提供了另一种更为直接的思路：直接在输出层回归bounding box的位置和bounding box所属的类别(整张图作为网络的输入，把 Object Detection 的问题转化成一个 Regression 问题)。

YOLO的主要特点：

- 速度快，能够达到实时的要求。在 Titan X 的 GPU 上 能够达到 45 帧每秒。
- 使用全图作为 Context 信息，背景错误（把背景错认为物体）比较少。
- 泛化能力强。



大致流程：

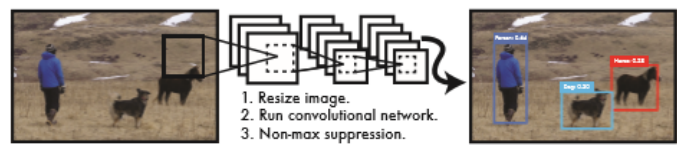


Figure 1: The YOLO Detection System. Processing images with YOLO is simple and straightforward. Our system (1) resizes the input image to 448×448 , (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence.

- 1. **Resize**成 448×448 ，图片分割得到 7×7 网格(cell)
- 2. **CNN提取特征和预测**：卷积不负责提特征。全链接部分负责预测：a) $7 \times 7 \times 2 = 98$ 个bounding box(bbox) 的坐标 $x_{center}, y_{center}, w, h$ 和是否有物体的confidence。b) $7 \times 7 = 49$ 个cell所属20个物体的概率。
- 3. **过滤**bbox (通过nms)



网络设计：

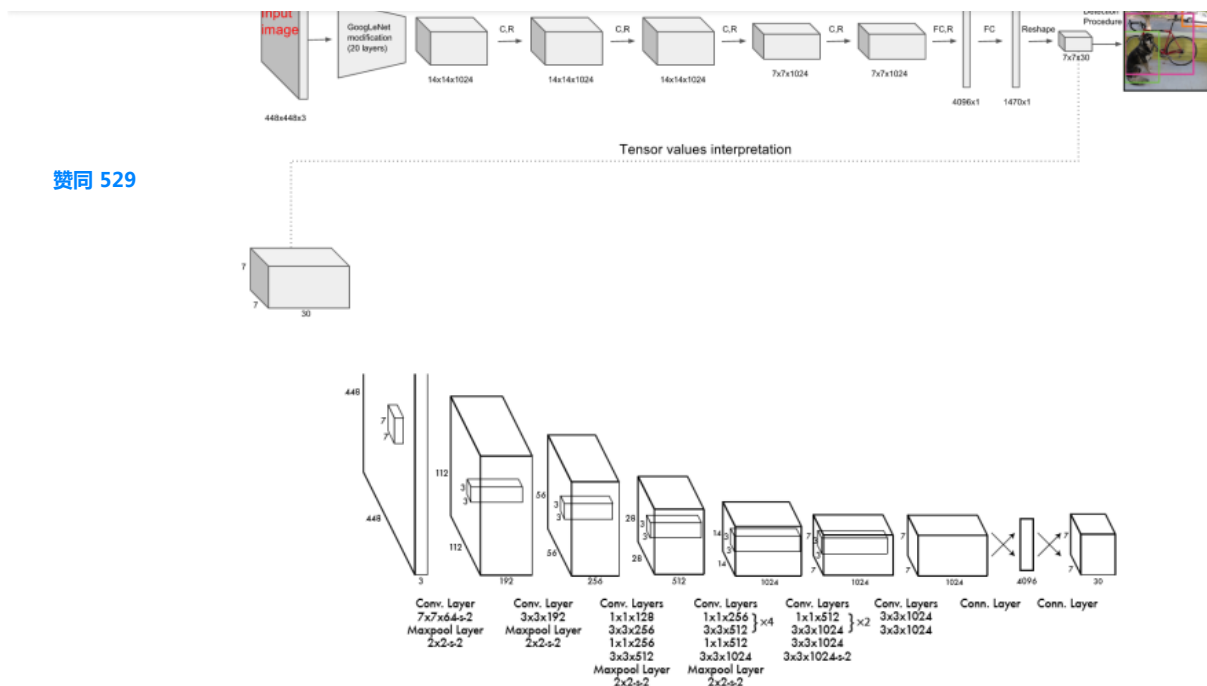


Figure 3: The Architecture. Our detection network has 24 convolutional layers followed by 2 fully connected layers. Alternating 1×1 convolutional layers reduce the features space from preceding layers. We pretrain the convolutional layers on the ImageNet classification task at half the resolution (224×224 input image) and then double the resolution for detection.

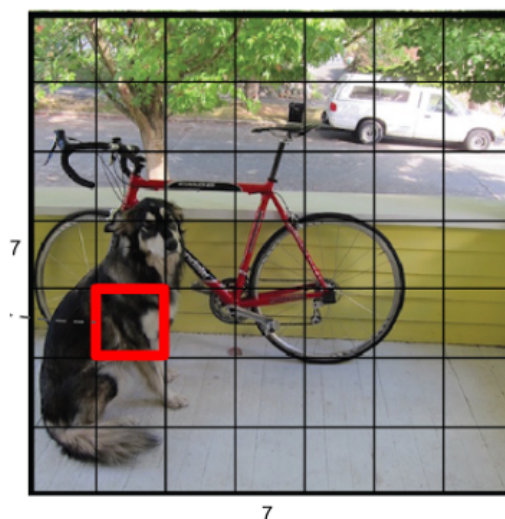
网络结构借鉴了 GoogLeNet。24个卷积层，2个全链接层。（用1×1 reduction layers 紧跟3×3 convolutional layers 取代GooleNet的 inception modules ）

训练：

预训练分类网络：在 ImageNet 1000-class competition dataset上预训练一个分类网络，这个网络是Figure3中的前20个卷积网络+average-pooling layer+ fully connected layer（此时网络输入是 224×224 ）。

训练检测网络：转换模型去执行检测任务，《Object detection networks on convolutional feature maps》提到说在预训练网络中增加卷积和全链接层可以改善性能。在他们例子基础上添加4个卷积层和2个全链接层，随机初始化权重。检测要求细粒度的视觉信息，所以把网络输入也又 224×224 变成 448×448 。见Figure3。

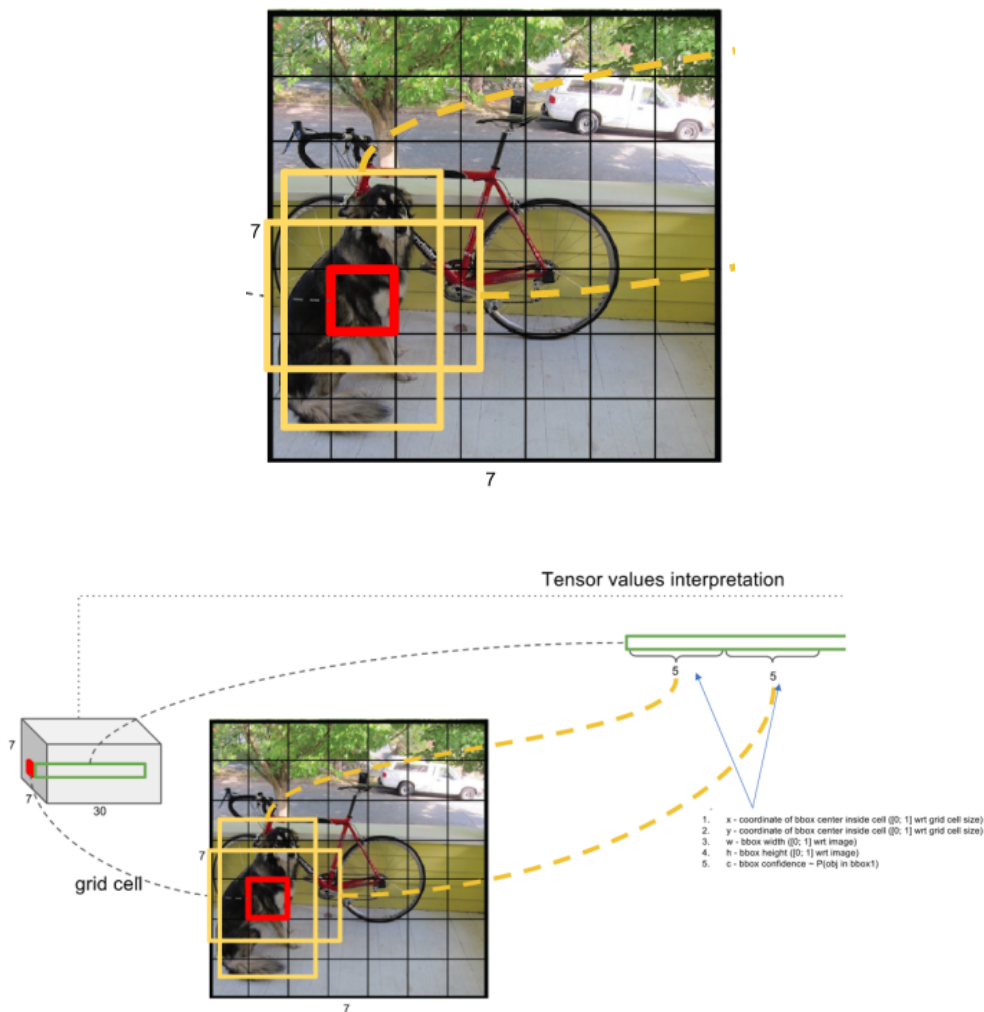
- 一幅图片分成7x7个网格(grid cell), 某个物体的中心落在这个网格中此网格就负责预测这个物体。



- 最后一层输出为 $(7 \times 7) \times 30$ 的维度。每个 $1 \times 1 \times 30$ 的维度对应原图 7×7 个 cell 中的一个， $1 \times 1 \times 30$ 中含有类别预测和 bbox 坐标预测。总得来讲就是让网格负责类别信息，bounding box 主要负责坐标信息(部分负责类别信息：confidence 也算类别信息)。具体如下：

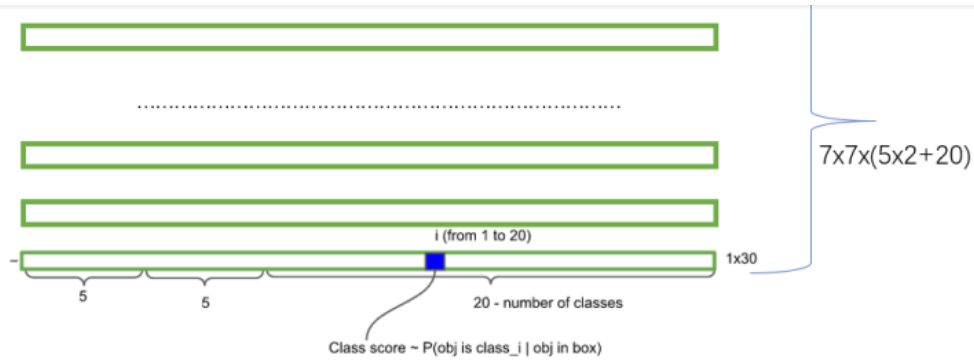
赞同 529

- 每个网格 ($1 \times 1 \times 30$ 维度对应原图中的 cell) 要预测 2 个 bounding box (图中黄色实线框) 的坐标 $(x_{center}, y_{center}, w, h)$ ，其中：中心坐标的 x_{center}, y_{center} 相对于对应的网格归一化到 0-1 之间， w, h 用图像的 width 和 height 归一化到 0-1 之间。每个 bounding box 除了要回归自身的位置之外，还要附带预测一个 confidence 值。这个 confidence 代表了所预测的 box 中含有 object 的置信度和这个 box 预测的有多准两重信息：confidence = $Pr(Object) * IOU_{pred}^{truth}$ 。其中如果有 ground true box (人工标记的物体) 落在一个 grid cell 里，第一项取 1，否则取 0。第二项是预测的 bounding box 和实际的 ground truth box 之间的 IOU 值。即：每个 bounding box 要预测 $x_{center}, y_{center}, w, h, confidence$ ，共 5 个值，2 个 bounding box 共 10 个值，对应 $1 \times 1 \times 30$ 维度特征中的前 10 个。

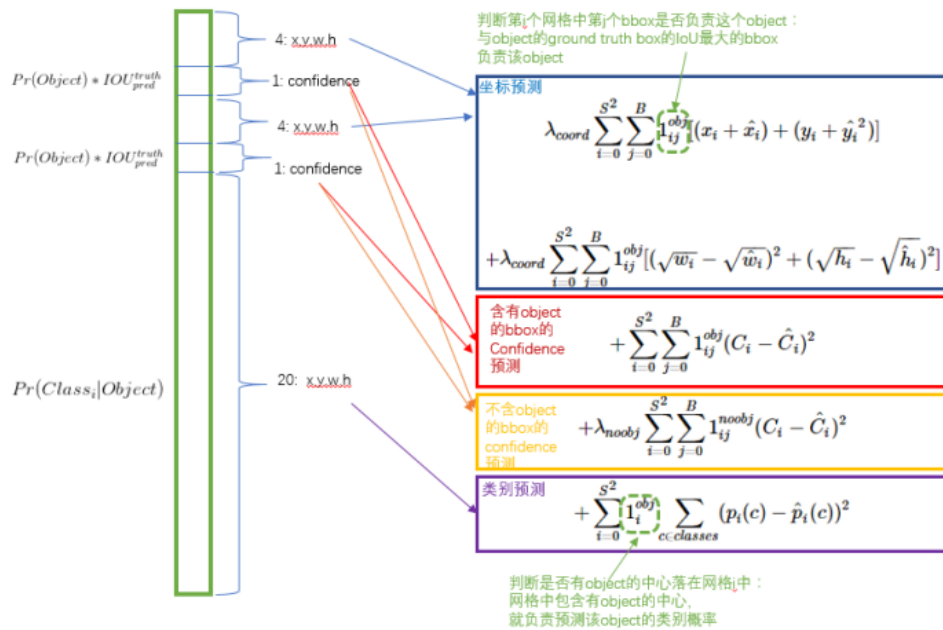


- 每个网格还要预测类别信息，论文中有 20 类。7x7 的网格，每个网格要预测 2 个 bounding box 和 20 个类别概率，输出就是 $7 \times 7 \times (5 \times 2 + 20)$ 。(通用公式： $S \times S$ 个网格，每个网格要预测 B 个 bounding box 还要预测 C 个 categories，输出就是 $S \times S \times (5 \times B + C)$ 的一个 tensor。注意：class 信息是针对每个网格的，confidence 信息是针对每个 bounding box 的)

赞同 529



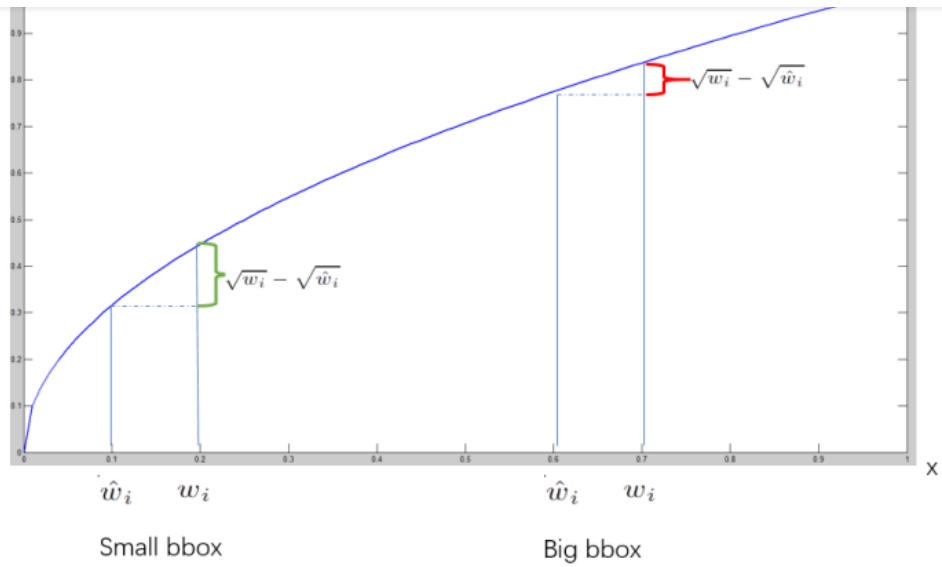
损失函数设计：



损失函数的设计目标就是让坐标 (x,y,w,h)，confidence，classification 这个三个方面达到很好的平衡。简单的全部采用了sum-squared error loss来做这件事会有以下不足：a) 8维的 localization error和20维的classification error同等重要显然是不合理的；b) 如果一个网格中没有object（一幅图中这种网格很多），那么就会将这些网格中的box的confidence push到0，相比于较少的有object的网格，这种做法是overpowering的，这会导致网络不稳定甚至发散。解决方案如下：

- 更重视8维的坐标预测，给这些损失前面赋予更大的loss weight，记为 λ_{coord} ，在pascal VOC训练集中取5。（上图蓝色框）
- 对没有object的bbox的confidence loss，赋予小的loss weight，记为 λ_{noobj} ，在pascal VOC训练集中取0.5。（上图橙色框）
- 有object的bbox的confidence loss（上图红色框）和类别的loss（上图紫色框）的loss weight正常取1。
- 对不同大小的bbox预测中，相比于大bbox预测偏一点，小bbox预测偏一点更不能忍受。而sum-square error loss中对同样的偏移loss是一样。为了缓和这个问题，作者用了一个比较取巧的办法，就是将box的width和height取平方根代替原本的height和width。如下图：small bbox的横轴值较小，发生偏移时，反应到y轴上的loss（下图绿色）比big box(下图红色)要大。

赞同 529



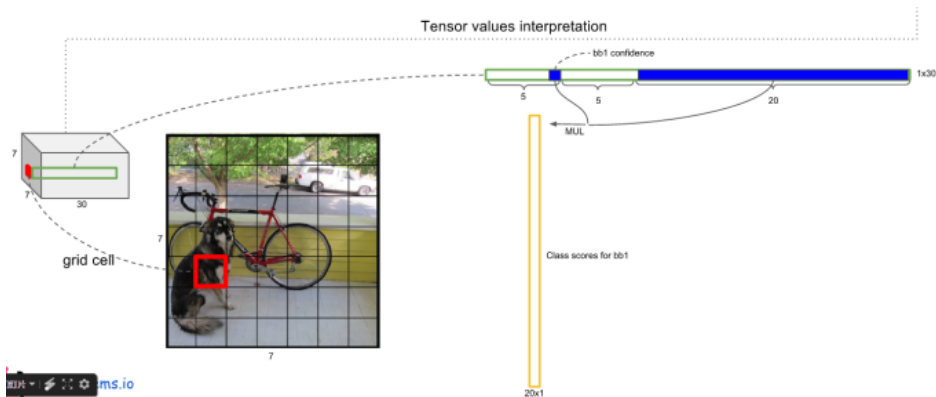
- 一个网格预测多个bounding box，在训练时我们希望每个object (ground true box) 只有一个bounding box专门负责（一个object 一个bbox）。具体做法是与ground true box (object) 的IOU最大的bounding box 负责该ground true box(object)的预测。这种做法称作bounding box predictor的specialization(专职化)。每个预测器会对特定（ sizes,aspect ratio or classed of object ）的ground true box预测的越来越好。（个人理解：IOU最大者偏移会更少一些，可以更快速的学习到正确位置）

测试：

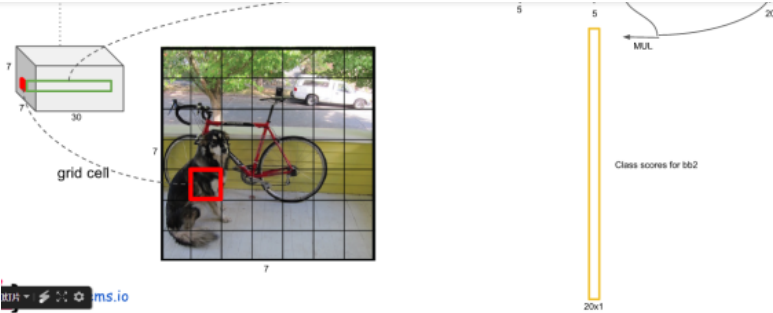
Test的时候，每个网格预测的class信息($Pr(Class_i|Object)$)和bounding box预测的confidence信息($Pr(Object) * IOU_{pred}^{truth}$)相乘，就得到每个bounding box的class-specific confidence score。

$$Pr(Class_i|Object) * Pr(Object) * IOU_{pred}^{truth} = Pr(Class_i) * IOU_{pred}^{truth}$$

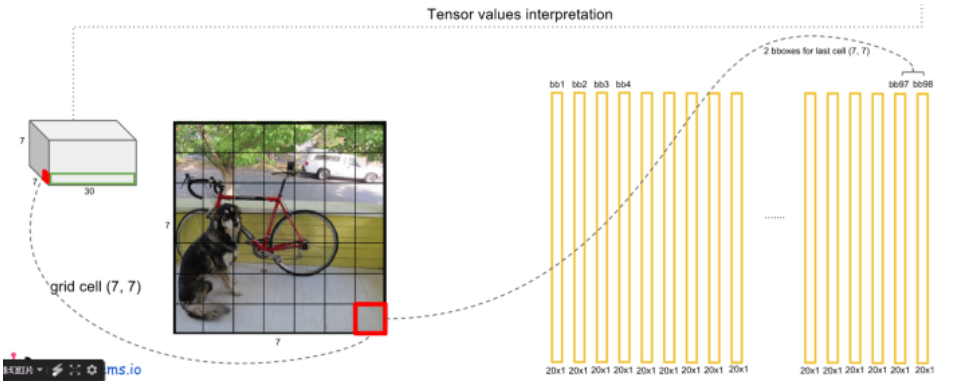
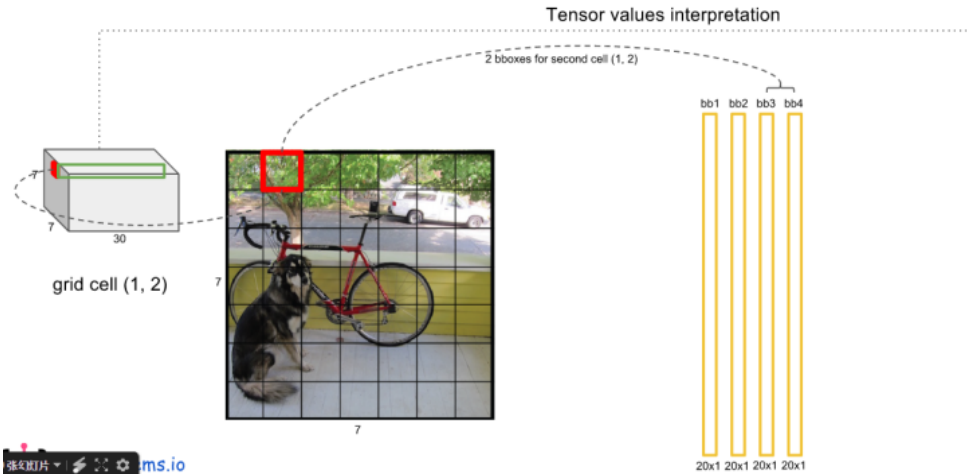
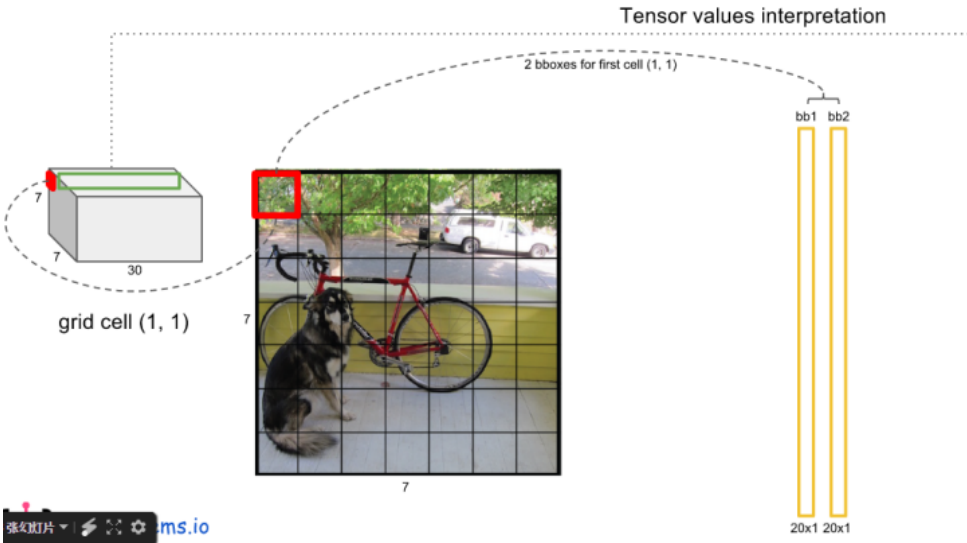
- 等式左边第一项就是每个网格预测的类别信息，第二三项就是每个bounding box预测的confidence。这个乘积即encoded了预测的box属于某一类的概率，也有该box准确度的信息。



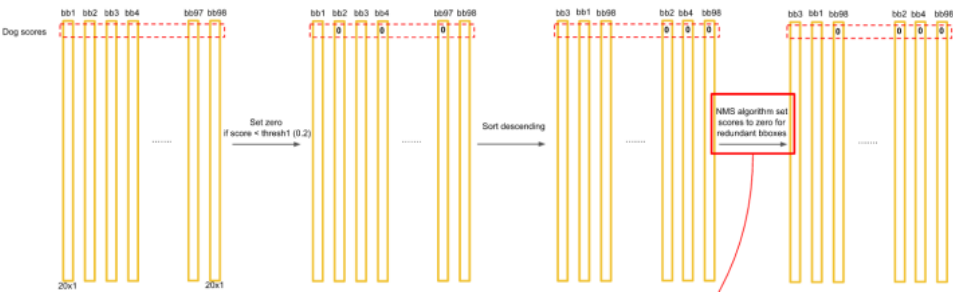
赞同 529



- 对每一个网格的每一个bbox执行同样操作： $7 \times 7 \times 2 = 98$ bbox（每个bbox既有对应的class信息又有坐标信息）



赞同 529



缺陷：

- YOLO对相互靠的很近的物体（挨在一起且中点都落在同一个格子上的情况），还有很小的群体检测效果不好，这是因为一个网格中只预测了两个框，并且只属于一类。
- 测试图像中，当同一类物体出现的不常见的长宽比和其他情况时泛化能力偏弱。
- 由于损失函数的问题，定位误差是影响检测效果的主要原因，尤其是大小物体的处理上，还有待加强。

本文图片很多来自PPT: deepsystems.io
内容主要参考如下博客：

- [RCNN学习笔记\(6\)：You Only Look Once\(YOLO\):Unified, Real-Time Object Detection](#)
- [You Only Look Once: Unified, Real-Time Object Detection](#)

画图不易，如果觉得文章不错欢迎点赞支持一下。

编辑于 2017-01-22

「如果觉得文章有帮助就打赏一杯咖啡吧~」

赞赏

3 人已赞赏



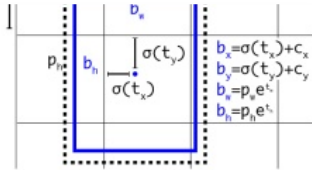
神经网络 深度学习（Deep Learning） 目标检测

文章被以下专栏收录

晓雷机器学习笔记


推荐阅读

赞同 529



YOLO2

晓雷 发表于晓雷机器学习...



Tensorflow实现YOLO1

KOD Chen

YOLOC
这篇博
及如何
算法。
object
结构简
detect

Mirac

98 条评论

切换为时间排序

写下你的评论...


😊

精选评论 (1)


 白色源代码
楼主加油(ง ٩_٩)ง
👍 3 查看回复

1 年前

评论 (98)

 月夜
刚看完上一篇，这一篇又写好了。楼主真高产，高水平。
👍 2

1 年前

 晓雷 (作者) 回复 月夜
最近刚好工作需要就总结了一下。高水平谈不上，只是画了点图，内容也是整理别人的。只算是低水平搬运工哈哈
👍 2

1 年前

 no speaking 回复 晓雷 (作者)
图是怎么画的呢
👍 赞


3 个月前

 白色源代码
楼主加油(ง ٩_٩)ง
👍 3


1 年前

 晓雷 (作者) 回复 白色源代码
谢谢！☺
👍 赞

1 年前

 RTMLD
楼主，有个问题想问一下。文章里目标的中心怎么确定？
👍 1

1 年前

 晓雷 (作者) 回复 RTMLD
你说的中心是不是指图像中人工标注的物体中心：一幅图像中人工标注框的中心就是该目标的中心。（一幅图像会标注一到多个目标位置信息）

1 年前

赞同 529



jiarenfy

1 年前

重点没讲，当然翻译的不错.....(๑`o`)๓

👍 5



Solomon

1 年前

楼主有keras/tf代码实现吗？原版的darknet看不太懂

👍 1



器鸢

1 年前

楼主，很想问下：在训练过程中，网格所预测的两个黄色的bounding boxes是怎样生成的？

👍 1



晓雷 (作者) 回复 器鸢

1 年前

YOLO中两个bbox是人为选定的(2个不同长宽比)的box，Faster RCNN也是人为选定的(9个不同长宽比和scale)，YOLOv2是统计分析ground true box的特点得到的(5个)。

👍 6



器鸢 回复 晓雷 (作者)

1 年前

好的，明白了。谢谢楼主！

👍 赞



soy肉泥worn欧诺 回复 晓雷 (作者)

1 年前

论文当中有指出吗？我看过后以为是网络输出的，，，

👍 赞



hui zhou 回复 晓雷 (作者)

1 年前

论文中并没有提人为选定两个不同尺度的box啊

👍 1



lemon xin

1 年前

您好，每个网格预测两个bounding box，这两个bounding box没有初始坐标值，怎么给这两个bounding box给标签呢？faster-rcnn系列的anchor都是有初始坐标值的，可以和groundtruth比较得出标签值。

👍 1



晓雷 (作者) 回复 lemon xin

1 年前

每个cell对应的2个bounding box是手动选定固定尺度(比例和面积)的，即宽高(w1,h1)(w2,h2)都可以计算出来。再加上cell本身的中心点坐标(x,y)。(x,y,w1,h1)(x,y,w2,h2)可以完全确定两个bounding box。

👍 赞



拉链 回复 晓雷 (作者)

1 年前

您这个有点看不懂。。我理解的是网络不是直接能输出4个坐标值么，就这相当于给定了“预测的bbox”的坐标值，直接拿这组坐标值和ground-truth的坐标值进行对比进而进行训练优化等后续操作。这样的理解对吗？

👍 赞

查看全部 7 条回复



Fate

1 年前

改写了一个caffe版的yolo9000，有需要的童鞋可以参考^_^
github.com/choasUp/caff...

👍 4



杨思达zzzz

1 年前

赞同 529



晓雷 (作者) 回复 杨思达zzzz

1 年前

感谢支持~哈哈

👍 赞



杨思达zzzz

1 年前

非常感谢您的分享，感觉您写的很仔细很用心

👍 赞



薛冬毅

1 年前

很棒的文章！我有一个问题，标记中心的grid用于预测bbx，但是同样在一个物体范围内的其它grid呢？如何训练他们？

👍 赞



何志 回复 薛冬毅

6 个月前

.我也想知道

👍 赞



mary

1 年前

楼主，YOLO输出某个class的confidence值大于1，请问是怎么回事？

👍 1



soy肉泥worn欧诺

1 年前

以前没看过目标检测的文章，我想问下，目标方程中的监督信息从何而来，直接用的groundtruth中的box吗？

👍 1



拉链

1 年前

文中写到“一个网格预测多个bounding box，在训练时我们希望每个object (ground true box) 只有一个bounding box专门负责 (一个object 一个bbox) ”。这一段文字该如何理解呢？？什么叫“一个bbox专门负责1个object”？？那另1个bbox如何处理？？麻烦能不能详细解释下这段文字

👍 1



麦田守望者 回复 拉链

1 年前

文中默认一个grid cell最多含有1个物体，但是一个grid cell却含有2个bbox，然后作者只让“IoU最大的那个bbox”对该物体的预测负责，这体现在YOLO的loss函数上。使用这种loss来训练的的目的是为了让grid cell的不同bbox适应不同尺度比例的物体

👍 5



甜豆的爱豆 回复 麦田守望者

4 个月前

训练阶段，一个对应了物体的cell，它训练出来的2个bounding box，会不会为了减小误差而都向ground truth靠近，导致这2个bounding box没有差异，那么要2个完全相同的bounding box有什么用呢？还望楼主能指点一下~

👍 赞



夏康力 回复 甜豆的爱豆

29 天前

我觉得不会，我的理解是这两个box只会有一个来负责预测目标，另一个的confidence标签是0

👍 赞



肖大大

1 年前

楼主，我觉得你对深度网络的理解很是到位，想向你学习学习，有些细节问题搞得不是很清楚，可否加我QQ：405631964，具体探讨探讨，如果可以，也愿意有偿咨询；

👍 1

赞同 529

该怎样注明引用来源呢？谢谢

👍 2

李鼎

见过的几篇中写的最好的了，对我帮助很大

👍 1

子猫言

loss设计那里的公式打错了0.0

👍 1

陌上花香自来

回复 子猫言

对，前面两个都是平方

👍 赞

XUGeorge

损失函数的图片有错误

👍 1

1

2

3

下一页

https://zhuanlan.zhihu.com/p/24916786

12/12