Yolo 论文阅读笔记

整理：李英杰

[论文地址：]https://arxiv.org/abs/1612.08242

提出背景

- **之前的检测系统使用分类器进行检测，当检测某一个目标时，首先训练该目标的分类器，然后在一张测试图的不同位置和不同尺寸的boundingbox上使用该分类器去进行评估。如deformable parts models(DPM)（可变形部件模型）使用一个滑窗（sliding window）在整张图像上均匀滑动，用分类器评估是否有物体。之后的RCNN系列, 首先使用 region proposal在整张图像中生成潜在的bounding boxes ，然后用一个分类器 对 bounding boxes进行分类。分类之后，优化proposal boxes，消除重复的框，基于场景 ,中其他目标重新对边框进行打分，这些复杂的网络由于每部分独立的训练导致运行缓慢，难以优化。**

Yolo的介绍

- 把目标检测重新设计为回归问题来处理，直接通过整张图片的所有像素得到bounding box coordinates和 class probabilities。通过YOLO，每张图像只需要看一眼就能得出图像中都有哪些物体和这些物体的位置。

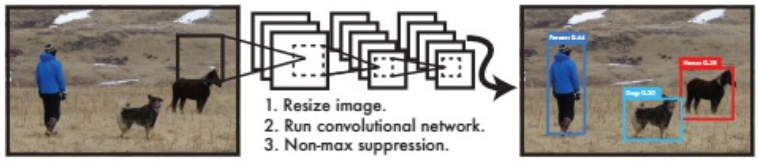
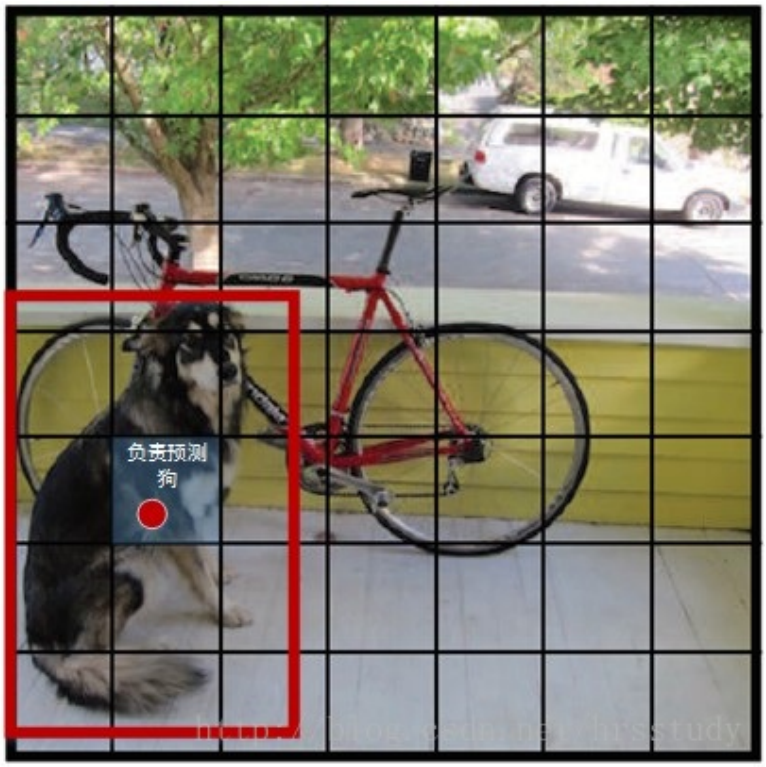


Figure 1: The YOLO Detection System. Processing images with YOLO is simple and straightforward. Our system (1) resizes the input image to 448 × 448, (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence.

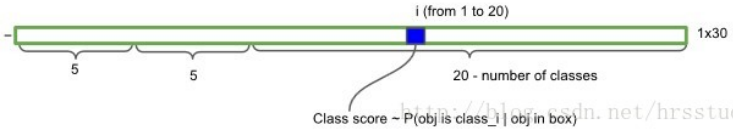
- 如图所示，使用YOLO来检测物体，其流程是非常简单明了的：
  - 1、将输入图像resize到448 * 448
  - 2、运行神经网络，得到bounding boxes坐标、box中包含物体的置信度和class probabilities
  - 3、非极大值抑制，筛选Boxes

Unified Detection

- 将输入图像划分为s*s网格如果一个目标的中心落入一个grid cell中，grid cell就负责检测该目标。



- 每一个grid cell预测**B**个框的bounding boxes和confidence scores(置信值)，confidence scores反映了模型对框的预测即这个框是否包含了一个目标和他对框的坐标预测的准确程度。定义：**confidence=Pr(object) * IOUpred/true**。当grid cell中没有目标时，confidence scores为0，否则，confidence scores等于IOUtrue pred。预测框和真实框的IOU）
- 其中，每一个bounding box 包含5个预测值（X , Y ,W, H, confidence），（X.Y）表示了框的中心与grid cell边界的相对值。（W，H）表示了框的宽，高相对于整幅图像width,height的比例。Confidence代表了预测框和真实框的IOU。
- 每一个grid cell还预测C个条件类概率**Pr(Classi|Object)** 。即grid cell包含目标的前提下，该目标属于某一类的概率。我们只预测每个grid cell的一系列（C个）类概率，而不考虑框B数量。



> attention !
>
> > conditional class probability信息是针对每个网格的。
> > confidence信息是针对每个bounding box的。

- 在测试阶段，将每个栅格的conditional class probabilities与每个 bounding box的 confidence相乘：

$$\Pr(\text{Class}_i|\text{Object}) * \Pr(\text{Object}) * \text{IOU}_{\text{pred}}^{\text{truth}} = \Pr(\text{Class}_i) * \text{IOU}_{\text{pred}}^{\text{truth}}$$

系统模型作为一个回归问题，把图像划分为s * s的单元格，每一个单元格预测B个bounding boxes，confidence for those boxes。C个条件类概率。这个预测被编码成S * S * (B*5+C)维的向量。

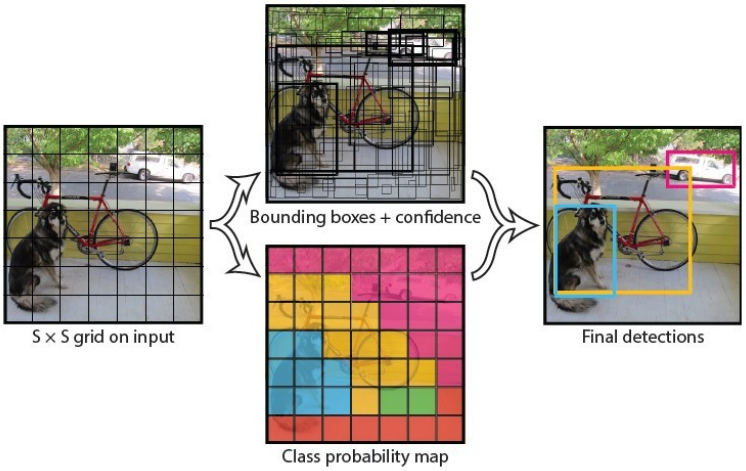
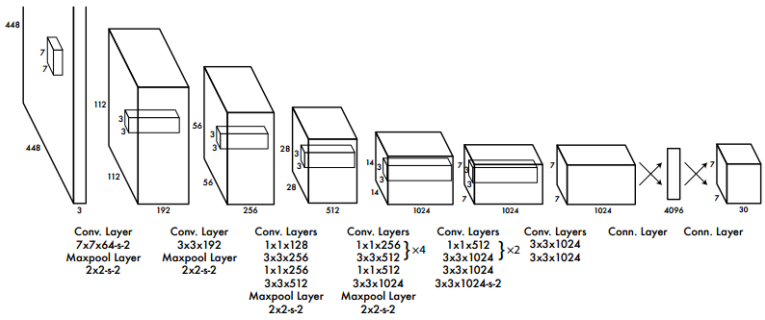


Figure 2: The Model. Our system models detection as a regression problem. It divides the image into an $S \times S$ grid and for each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities. These predictions are encoded as an $S \times S \times (B * 5 + C)$ tensor.

设计网络

- YOLO检测网络用卷积网络设计模型并且在PASCAL VOC检测数据集进行评估。网络的初始卷积层从图像抽取特征同时全连接层预测输出的类概率和坐标。网络有24个卷积层以及两层全连接层。



YOLO网络借鉴了GoogLeNet分类网络结构。不同的是，YOLO未使用inception module，而是使用1x1卷积层（此处1x1卷积层的存在是为了跨通道信息整合）+3x3卷积层简单替代。

- 同时：
  - 论文中还训练了一个Fast YOLO版本用于突破快速目标检测的记录，由9层卷积，层中的过滤器更少，除了尺寸的不同，所有的训练和测试参数在YOLO和Fast YOLO相同。
  - 本论文使用的 S=7，即将一张图像分为7×7=49个grid cell每一个grid cell预测B=2个boxes（每个box有x,y,w,h,confidence，5个预测值），同时C=20（PASCAL数据集中有20个类别）。预测结果为S * S * (B*5+C)=7 * 7 * （5 * 2+20）维向量。

训练

训练分类器

- 首先利用ImageNet 1000-class的分类任务数据集Pretrain卷积层。使用上述网络中的前20 个卷积层，加上一个 average-pooling layer，最后加一个全连接层，作为 Pretrain 的网络。
  训练大约一周的时间，使得在ImageNet 2012的验证数据集Top-5的精度达到 88%，这个结果跟

GoogleNet 的效果相当。

训练检测器

- 将Pretrain的结果的前20层卷积层应用到Detection中，并加入剩下的4个卷积层及2个全连接。 同时为了获取更精细化的结果，将输入图像的分辨率由 224* 224 提升到 448* 448。 将所有的预测结果都归一化到 0~1。最终的输出层使用线性激活函数，其他所有层使用leaky rectified linear 激活函数。

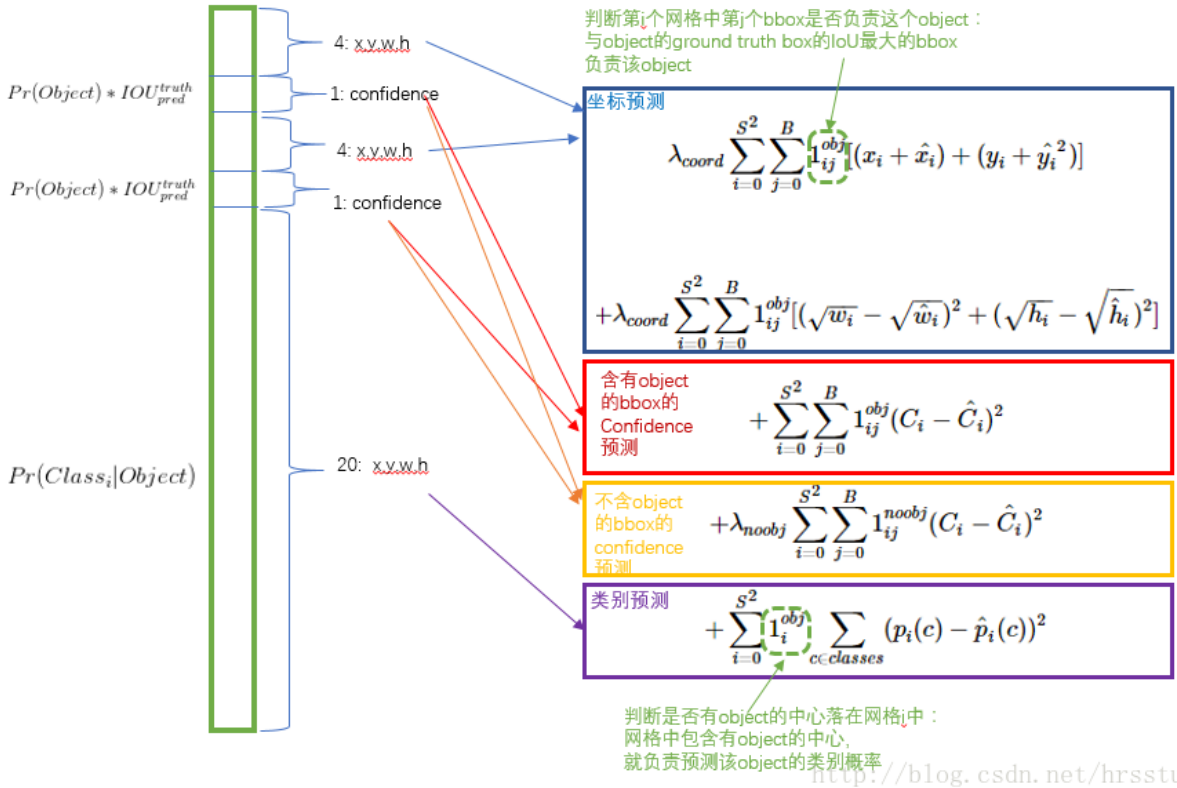
该激活函数为：

$$\phi(x) = \begin{cases} x, & \text{if } x > 0 \\ 0.1x, & \text{otherwise} \end{cases}$$

- 避免过拟合和提高精度：
  - 1 在第一个全连接层后面接了一个 ratio=0.5 的 Dropout 层。
  - 2数据填充以及在HSV颜色空间中调整图像的曝光率和饱和度。

损失函数

- 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。
- 不足之处为：
  - 缺点一：将定位误差和可能不理想的分类误差同等看待
  - 缺点二：如果一些grid cell中没有object（一幅图中这种grid cell很多），那么就会将这些grid cell中的bounding box的confidence 置为0，相比于较少的有object的grid cell，这些不包含物体的栅格对梯度更新的贡献会远大于包含物体的栅格对梯度更新的贡献，这会导致网络不稳定甚至发散。



- 解决方法：
  - 增加了bounding box coordinates的损失权重，减少不包含目标框的confidence预测损失的权重，

    objects. We use two parameters, λ_{coord} and λ_{noobj} to accomplish this. We set $\lambda_{coord} = 5$ and $\lambda_{noobj} = .5$.

  - 对不同大小的bbox预测中，相比于大bbox预测偏一点，小box预测偏相同的尺寸对IOU的影响更大。而sum-square error loss中对同样的偏移loss是一样。为了解决这个问题，就是将box的width和height取平方根代替原本的height和width。
  - 在 YOLO中，每个grid cell预测多个bounding box，但在网络模型的训练中，希望**每一个目标最后由一个bounding box predictor来负责预测。因此，当前哪一个predictor预测的bounding box与ground truth box的IOU最大，这个 predictor就负责 predict object**。 这会使得每个predictor可以专门的负责特定的物体检测。随着训练的进行，每一个 predictor对特定的物体尺寸、长宽比的物体的类别的预测会越来越好。

神经网络输出后的检测流程

非极大值抑制

获取Object Detect 结果

YOLO的局限

> 1：强大的空间约束，每个grid cell预测两个box和一个类别。模型不适用于小的目标例如群体出现的鸟群。
> 2：模型从数据中学习来预测边界框，很难对新的或者特殊比例的目标进行识别，泛化能力弱。 3：输入图像经过多个向下采样层，造成数据缺失，模型使用粗糙特征预测bounding boxes导致定位错误。

优势

> 1，统一的网络架构是非常快的。基于YOLO每秒45帧实时处理图片。这个网络的一个小的版本---FAST YOLO，达到每秒155帧的惊人的实时处理图片速度 同时mAP分数是其他实时检测器的两倍。
> 和最新的检测系统相比，YOLO产生更多的定位错误，但很少在背景方面判断错误。
> YOLO学习了对目标的非常泛化的表示，他优于其他的检测方法，当从自然图像泛化到其他领域 如艺术品。

总结

YOLO联合训练，速度快，泛化能力强，健壮