YOLO:You Only Look Once:Unified,Real-Time Object Dection

摘要

实时的物体检测器1231241fffff 特别快，有两个版本：YOLO:45fps,Fast YOLO:155fps
把物体检测当成一个回归问题来对待
使用一个单个的神经网络直接从图片中预测bounding-box和物体所属类别的概率
YOLO的主要错误来源于：定位错误，很少有false positives(把背景当做物体)

1.介绍

RCNN的缺点：优化慢并且困难，因为每个组件都要被分别训练
YOLO检测的过程：把输入图片的大小调整为：448×448,跑一个卷积神经网络，用模型的检测信度用一个阈值来筛选结果
YOLO的特定：

1. 特别快，在Tian X GPU上，没有批处理的版本：45fps,快速版本150fps ,可以用来实时处理视频
2. 在做预测时，是对图片进行整体处理，相当于编码了类和外观的上下文信息，这样可以减少把背景当作物体的错误发生
3. YOLO学习了物体的正则化的表示，在自然图像上训练，在艺术图片上测试

2.统一检测

我们的系统把输入图片分成$S \times S$个格子，物体的中心在那个格子，就那个格子对检测这个物体负责
每个格子预测B个bounding box和信度得分，信度得分反应了box包含物体的信度，和box的精确度，信度定义为$Pr(Object) * IOU_{pred}^{truth}$
每个bounding box有5个预测：x,y,w,,h,信度
每个格子预测C个条件类概率$Pr(Class_i|Object)$，为每个格子，至于猜测类概率，而不是为每个box预测概率
在测试阶段，计算 $Pr(Class_i|Object) * Pr(Object) * IOU_{pred}^{truth}$，表示特定类的信度得分，表示格子中出现这个类物体的概率和预测的Box适合这个物体的程度
预测最后被编码成一个$S \times S \times (B * 5 + C)$的一个张量
在PASCAL VOC中B=2,S=7,C=20

网络设计

网络框架是GoogLeNet分类网络，有24个卷积层和2个全连接层，使用1×1的层后面跟一个3×3的卷积层来约减网络
Fast YOLO是使用了9个卷积层，并且每层的滤波器数量也更少，其他的跟YOLO的是一样的
网络最后的输出是$7 \times 7 \times 30$的张量

训练

在ImageNet 1000类的竞赛数据集上预训练网络，预训练的网络结构是图三的前20个卷积层后面加一个平均池化层和一个全连接层，训练时间为：一周
在预训练的网络上加了随机初始化的4个卷积层和2两个全连接层用于做检测任务
输入的图片的分辨率是448×448，更多细粒度的视觉信息
最后一层预测类概率和bounding-box的坐标，这些数据都正则化到0到1之间
优化的是最小二乘误差 ， 缺点：

1. 分类误差和定位误差一样，有许多不包含物体的格子，对结果影响太大，导致模型不稳定
   解决方法：增加定位误差的权重，减少不包含物体的预测的信度误差的权重，$\lambda_{coord} = 5, \lambda_{noobj} = 0.5$
2. 大box中的错误跟小box中的错误的权重相同，小的偏差在大box中没有在小box中重要，，解决方法：预测bounding-box的宽和高的平方根，可以证明

YOLO会为每个格子预测多个候选框，在训练阶段，每个物体只需要一个候选框，选择跟真实物体的IOU最高的那个候选框 在VOC2007和VOC2012上的训练分为135个阶段，当在VOC2012上测试时，VOC2007的测试数据也用于训练，batch size是64，冲量是0.9,衰减是0.0005
学习率：第一阶段，学习率缓慢的从10^{-3}到10^{-2}，因为刚开始梯度不可靠
10^{-2}训练75个阶段，10^{-3}训练30个阶段10^{-4}训练30个阶段
为了防止overfiting,使用dropout和增广数据

推论

在VOC上，网络为每张图片预测了98个bounding box和为每个Box预测了类概率
网格的设计增加了空间多样性，一般的物体落入那个网格就预测那个网格，但打得物体可能会靠经多个网格的边缘，这样就预测多个网格，然后使用非极大值抑制

YOLO的局限性

网格的空间约束，限制了模型能够预测的相邻物体的数目，比如很难预测出一群鸟
对待小的错误在大box和小box中相同，对于小物体，小的IOU误差也会对网络优化过程造成很大的影响，从而降低了物体检测的定位准确性

3.跟其他检测系统的比较

DPM分为很多步，YOLO同时做特征提取，bounding box预测，非极大值抑制，上下文相关 ，YOLO比DPM更快更精确
RCNN:跟RCNN相比，使用了网格，减少了同一个物体多次检测的情况，比RCNN更少的候选bounding box,RCNN2000个，YOLO 98个，YOLO是把每个组件整合到一个网络，联合优化
Fast,Faster RCNN:比YOLO慢

4.实验

YOLO可以用来给Fast RCNN重新打分，减少把背景识别成物体的错误
YOLO推广到新的领域的数据库的效果比别的方法要好

跟其他实时监测系统的比较

用VGG_16训练的YOLO比YOLO慢
DPM,fastest DMP是实时的物体检测，30Hz,100Hz,但是检测的精确率很低
R-CNN不能实时检测，而且因为没有好的候选域 bounding box，精确性大大降低
fast R-CNN加速了分类阶段，但是依然使用selective search 产生候选域 bounding box，而且速度不能用于实时物体检测，0.5fps
Faster R-CNN使用神经网络产生候选域bounding box,最精确的版本是7fps,更小，不太精确的版本是18fps,VGG_16 Faster R-CNN的mAP比YOLO高10个点，但是速度是YOLO的6倍，Z-F Faster RCNN的速度是YOLO的2.5倍，但精确度比YOLO低

VOC 2007上的错误分析

YOLO的定位错误比其他的错误的总和还要多
Fast RCNN 有很多background错误，是YOLO的3倍

Fast R-CNN和YOLO的结合

通过使用YOLO消除Fast R-CNN的背景检测的错误，提高性能
每个R-CNN预测的bounding box用YOLO去检查，看是否预测到一个相似的bounding box ,如果是，就基于YOLO预测到概率和两个Box的重叠部分给这个预测一个推动
结合的模型，把Fast R-CNN的mAP提高了3.2% 结合的模型，对速度来说不好

VOC 2012上的结果

在VOC2012上mAP是57.9% 主要是小的物体检测的不好，比如，瓶子，羊，电视机

在艺术作品中实现人的检测

艺术作品的图片在像素级别跟自然图片相差很大，但在物体的大小和形状上是相似的
YOLO在Picasso数据集上也表现的比较好