目标检测基础知识与常见名词解释

任务概述

目标检测任务是找出图像和视频中人们感兴趣的物体，并且同时检测出他们的位置和大小。与图像分类不同，目标检测不仅要解决分类问题，还要解决定位问题。

有以下五大应用：

1.行人检测 2.面部检测 3.文本检测 4.交通标注与红绿灯检测 5.遥感目标检测

两个发展阶段：传统目标检测算法时期（1998-2014）和基于深度学习的目标检测算法时期（2014-至今）

传统算法

> **流程：**
>
> > 1.选择感兴趣区域，选取可能包含物体的区域；
> >
> > 2.对可能包含物体的区域进行特征提取；
> >
> > 3.对提取的特征进行检测分类。
>
> **缺点：**
>
> > 1.识别效果不够好，准确率不高；
> >
> > 2.计算量较大，运算速度慢；
> >
> > 3.可能产生多个正确的识别结果

基于CNNs的目标检测算法

> - **anchor-based 方法**
>
> > - 包括一阶段和二阶段检测算法，一阶段算法比二阶段算法速度快但精度低
> >
> > > two stage method:
> > >
> > > > step1:从图像中生成region proposals
> > > >
> > > > step2:从region proposals生成最终的物体边框

> 代表算法：RCNN、SPPNet、Fast RCNN、Faster RCNN、FPN、CascadeRCNN

one stage method:

> 不需要region proposal阶段，直接产生物体的类别概率和位置坐标值，经过一个阶段即可直接得到最终的检测结果，因此有着更快的检测速度
>
> 代表算法：YOLO系列、SSD、RetinaNet、

- 缺点：

1.Anchor的大小、数量、长宽比对于检测性能的影响很大，因此Anchor based的检测性能对于Anchor的大小、数量和长宽比都非常敏感

2.固定的Anchor极大地损害了检测器的普适性，导致对于不同任务，其Anchor都必须重新设置大小和长宽比。

3.为了去匹配真实框，需要生成大量的Anchor，但是大部分的Anchor在训练时标记为负样本，所以就造成了样本极度不均衡问题

4.在训练中，网络需要计算所有的Anchor与真实框的IOU，这样就会消耗大量内存和时间

- **anchor-free方法**
  - 摒弃anchor，通过基于边框/确定关键点的方式来完成检测，大大减少了网络超参数的数量

    > 代表算法：CornerNet、CenterNet、FSAF、FCOS、SAPD

  - 缺点：

    1. 正负样本不均衡：我们通常在特征图所有点上均匀采样 Anchor，而在大部分地方都是没有物体的背景区域，导致简单负样本数量众多，这部分样本对于我们的检测器没有任何作用。
    2. 超参难调：Anchor 需要数量、大小、宽高等多个超参数，这些超参数对检测的召回率和速度等指标影响极大。此外，人的先验知识也很难应付数据的长尾问题，这显然不是我们乐意见到的。
    3. 匹配耗时严重（训练阶段）：为了确定每个 Anchor 是正样本还是负样本，通常要将每个 Anchor 与所有的标签进行 IoU 的计算，这会占据大量的内存资源与计算时间

常用数据集

Pascal VOC

http://host.robots.ox.ac.uk/pascal/VOC/

有VOC2007和VOC2012两个数据集。

包含约10,000张带有边界框的图片用于训练和验证。含有20个类别。具体包括

> Person: person
>
> Animal: bird, cat, cow, dog, horse,sheep
>
> Vehicle: aeroplane, bicycle, boat,bus, car, motorbike, train
>
> Indoor: bottle, chair, dining table,potted plant, sofa, tv/monitor

ILSVRC

https://image-net.org/challenges/LSVRC/

ILSVRC是ImageNet Large Scale Visual Recognition Challenge的缩写，是基于ImageNet的一个图像识别大赛，每年都会举办。ILSVRC2012就是2012年举办的，比赛组织者会发布一整套数据

MS-COCO

https://cocodataset.org/#home

是微软公司建立的数据集。对于目标检测任务，COCO包含80个类别，每年大赛的训练和和验证集包含120,000张图片，超过40,000张测试图片。下面是这个数据集中的80个类别：

Person：person

Vehicle：bicycle,car,motorcycle,airplane,bus,train,truck,boat

Outdoor：traffic light, firhydrant, stop sign, parking meter, bench

Animal：bird,cat, dog,horse, sheep, cow, elephant, bear, zebra, giraffe

Accessory：backpack, umbrella,handbag, tie, suitcase

Sport：frisbee, skis,snowboard, sports ball, kite, baseball bat , baseball glove, skateboard, surfboard, tennisracket

Kitchen：bottle, wine glass,cup, fork, knife, spoon, bowl

Food: banana, apple,sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake

Furniture: chair, couch, potted plant,bed, dining table, toilet

Electronic: tv, laptop, mouse,remote, keyboard, cell phone

Appliance: microwave, oven,toaster, sink, refrigerator

Indoor: book, clock, vase,scissors, teddy bear, hair drier, toothbrus

其他任务数据集:

行人检测:

Dataset	Year	Description	#Cites
MIT Ped.[30]	2000	One of the first pedestrian detection datasets. Consists of $\sim$500 training and $\sim$200 testing images (built based on the LabelMe database). url: http://cbcl.mit.edu/software-datasets/PedestrianData.html	1515
INRIA [12]	2005	One of the most famous and important pedestrian detection datasets at early time. Introduced by the HOG paper [12]. url: http://pascal.inrialpes.fr/data/human/	24705
Caltech [59, 60]	2009	One of the most famous pedestrian detection datasets and benchmarks. Consists of $\sim$190,000 pedestrians in training set and $\sim$160,000 in testing set. The metric is Pascal-VOC @ 0.5 IoU. url: http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/	2026
KITTI [61]	2012	One of the most famous datasets for traffic scene analysis. Captured in Karlsruhe, Germany. Consists of $\sim$100,000 pedestrians ($\sim$6,000 individuals). url: http://www.cvlibs.net/datasets/kitti/index.php	2620
CityPersons [62]	2017	Built based on CityScapes dataset [63]. Consists of $\sim$19,000 pedestrians in training set and $\sim$11,000 in testing set. Same metric with CalTech. url: https://bitbucket.org/shanshanzhang/citypersons	50
EuroCity [64]	2018	The largest pedestrian detection dataset so far. Captured from 31 cities in 12 European countries. Consists of $\sim$238,000 instances in $\sim$47,000 images. Same metric with CalTech.	1

人脸检测:

Dataset	Year	Description	#Cites
FDDB [65]	2010	Consists of $\sim$2,800 images and $\sim$5,000 faces from Yahoo! With occlusions, pose changes, out-of-focus, etc. url: http://vis-www.cs.umass.edu/fddb/index.html	531
AFLW [66]	2011	Consists of $\sim$26,000 faces and 22,000 images from Flickr with rich facial landmark annotations. url: https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/aflw/	414
IJB [67]	2015	IJB-A/B/C consists of over 50,000 images and videos frames, for both recognition and detection tasks. url: https://www.nist.gov/programs-projects/face-challenges	279
WiderFace [68]	2016	One of the largest face detection dataset. Consists of $\sim$32,000 images and 394,000 faces with rich annotations i.e., scale, occlusion, pose, etc. url: http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/	193
UFDD [69]	2018	Consists of $\sim$6,000 images and $\sim$11,000 faces. Variations include weather-based degradation, motion blur, focus blur, etc. url: http://www.ufdd.info/	1
WildestFaces [70]	2018	With $\sim$68,000 video frames and $\sim$2,200 shots of 64 fighting celebrities in unconstrained scenarios. The dataset hasn't been released yet.	

文本检测:

Dataset	Year	Description	#Cites
ICDAR [71]	2003	ICDAR2003 is one of the first public datasets for text detection. ICDAR 2015 and 2017 are other popular iterations of the ICDAR challenge [72, 73]. url: http://rrc.cvc.uab.es/	530
STV [74]	2010	Consists of ~350 images and ~720 text instances taken from Google StreetView. url: http://tc11.cvc.uab.es/datasets/SVT_1	339
MSRA-TD500 [75]	2012	Consists of ~500 indoor/outdoor images with Chinese and English texts. url: http://www.iapr-tc11.org/mediawiki/index.php/MSRA_Text_Detection_500_Database_(MSRA-TD500)	413
IIIT5k [76]	2012	Consists of ~1,100 images and ~5,000 words from both streets and born-digital images. url: http://cvit.iiit.ac.in/projects/SceneTextUnderstanding/IIIT5K.html	165
Syn90k [77]	2014	A synthetic dataset with 9 million images generated from a 90,000 vocabulary of multiple fonts. url: http://www.robots.ox.ac.uk/~vgg/data/text/	246
COCOText [78]	2016	The largest text detection dataset so far. Built based on MS-COCO, Consists of ~63,000 images and ~173,000 text annotations. https://bgshih.github.[illegible]cocotext/.	69

交通信号灯检测:

Dataset	Year	Description	#Cites
TLR [79]	2009	Captured by a moving vehicle in Paris. Consists of ~11,000 video frames and ~9,200 traffic light instances. url: http://www.lara.prd.fr/benchmarks/trafficlightsrecognition	164
LISA [80]	2012	One of the first traffic sign detection dataset. Consists of ~6,600 video frames, ~7,800 instances of 47 US signs. url: http://cvrr.ucsd.edu/LISA/lisa-traffic-sign-dataset.html	325
GTSDB [81]	2013	One of the most popular traffic signs detection dataset. Consists of ~900 images with ~1,200 traffic signs capture with various weather conditions during different time of a day. url: http://benchmark.ini.rub.de/?section=gtsdb&subsection=news	259
BelgianTSD [82]	2012	Consists of ~7,300 static images, ~120,000 video frames, and ~11,000 traffic sign annotations of 269 types. The 3D location of each sign has been annotated. url: https://btsd.ethz.ch/shareddata/	224
TT100K [83]	2016	The largest traffic sign detection dataset so far, with ~100,000 images (2048 x 2048) and ~30,000 traffic sign instances of 128 classes. Each instance is annotated with class label, bounding box and pixel mask. url: http://cg.cs.tsinghua.edu.cn/traffic%2Dsign/	111
BSTL [84]	2017	The largest traffic light detection dataset. Consists of ~5000 static images, ~8[illegible] video frames, and ~24000 traffic light instances. https://hci.iwr.uni-heidelbe[illegible]de/node/6132	21

遥感目标检测:

Dataset	Year	Description	#Cites
TAS [85]	2008	Consists of 30 images of 729x636 pixels from Google Earth and ~1,300 vehicles. url: http://ai.stanford.edu/~gaheitz/Research/TAS/	419
OIRDS [86]	2009	Consists for 900 images (0.08-0.3m/pixel) captured by aircraft-mounted camera and 1,800 annotated vehicle targets. url: https://sourceforge.net/projects/oirds/	32
DLR3K [87]	2013	The most frequently used datasets for small vehicle detection. Consists of 9,300 cars and 160 trucks. url: https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-5431/9230_read-42467/	68
UCAS-AOD [88]	2015	Consists of ~900 Google Earth images, ~2,800 vehicles and ~3,200 airplanes. url: http://www.ucassdl.cn/resource.asp	19
VeDAI [89]	2016	Consists of ~1,200 images (0.1-0.25m/pixel), ~3,600 targets of 9 classes. Designed for detecting small target in remote sensing images. url: https://downloads.greyc.fr/vedai/	65
NWPU-VHR10 [90]	2016	The most frequently used remote sensing detection dataset in recent years. Consists of ~800 images (0.08-2.0m/pixel) and ~3,800 remote sensing targets of ten classes (e.g., airplanes, ships, baseball diamonds, tennis courts, etc). url: http://jiong.tea.ac.cn/people/JunweiHan/NWPUVHR10dataset.html	204
LEVIR [91]	2018	Consists of ~22,000 Google Earth images and ~10,000 independently labeled targets (airplane, ship, oil-pot). url: https://pan.baidu.com/s/1geTwAVD	15
DOTA [92]	2018	The first remote sensing detection dataset to incorporate rotated bounding boxes. Consists of ~2,800 Google Earth images and ~200,000 instances of 15 classes. url: https://captain-whu.github.io/DOTA/dataset.html	32
xView [93]	2018	The largest remote sensing detection dataset so far. Consists of ~1,000,000 remote sensing targets of 60 classes (0.3m/pixel), covering1,415km^2 of land area. url: http://xviewdataset.org	10

评估指标

1.IoU（交并比）

IoU=两个矩形交集的面积/两个矩形并集的面积

一般将IOU值设置为大于0.5的时候，则可检测到目标物体

2.准确率、精度、召回率、F1值、FPR

True positives (TP,真正): 预测为正,实际为正
True negatives (TN,真负): 预测为负,实际为负
False positives(FP,假正): 预测为正,实际为负
False negatives(FN,假负): 预测为负,实际为正

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1-Score = \frac{2\times TP}{2\times TP+FP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$

3.PR曲线-AP值

PR曲线就是Precision和Recall的曲线，我们以Precision作为纵坐标，Recall为横坐标

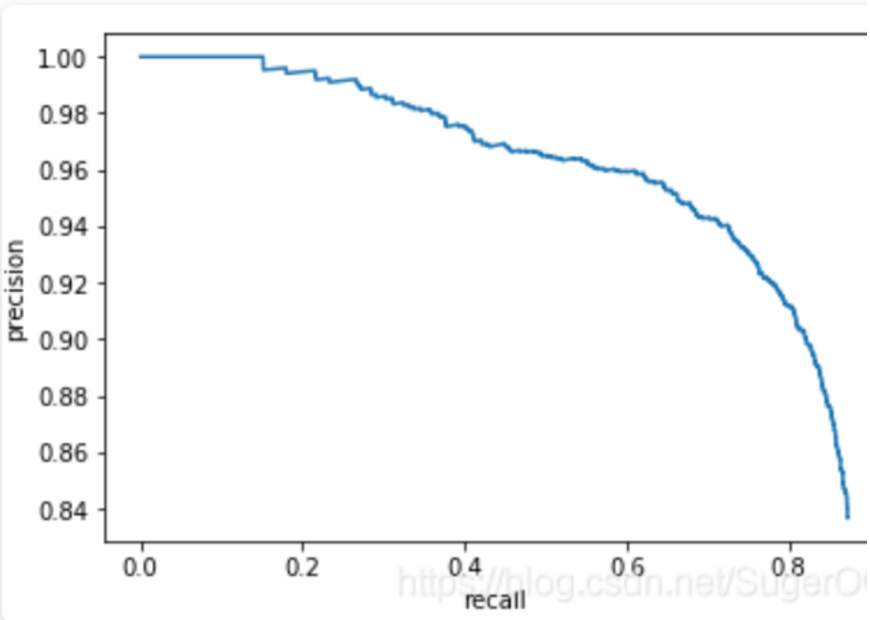

图3-3.PR曲线

如果模型的精度越高，且召回率越高，那么模型的性能自然也就越好，反映在P R曲线上就是PR曲线下面的面积越大，模型性能越好。我们将PR曲线下的面积定义为

AP(Average Precision)值，反映在AP值上就是AP值越大，说明模型的平均准确率越高。

4.ROC曲线-AUC值

ROC曲线就是RPR和TPR的曲线，我们以FPR为横坐标，TPR为纵坐标，可绘制ROC曲线

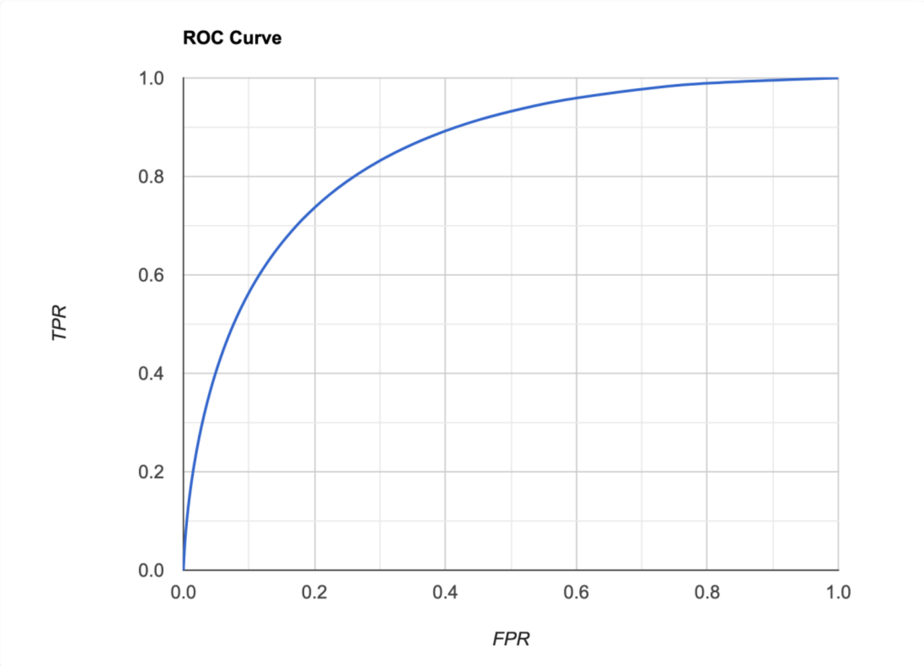

图3-4.ROC曲线

当TPR越大，FPR越小时，说明模型分类结果是越好的，反映在ROC曲线上就 是ROC曲线下面的面积越大，模型性能越好。我们将ROC曲线下的面积定义为AUC(Area Under Curve)值，反映在AUC值上就是AUC值越大，说明模型对正样本分类的结果越好。

5.MAP

Mean Average Precision(mAP)是平均精度均值，具体指的是不同召回率下的精度均值。 在目标检测中，一个模型通常会检测很多种物体，那么每一类都能绘制一个PR曲线，进而计算出一个AP值，而多个类别的AP值的平均就是mAP。

mAP衡量的是模型在所有类别上的好坏，属于目标检测中一个最为重要的指标，一般看论文或者评估一个目标检测模型，都会看这个值，这个值(0-1范围区间)越大越好。

一般来说mAP是针对整个数据集而言的，AP则针对数据集中某一个类别而言的，而percision和recall针对单张图片某一类别的。

6.FPS

Frame Per Second(FPS)指的是模型一秒钟能检测图片的数量，不同的检测模型往往会有不同的mAP和检测速度

常见术语

术语	解释
IoU	图和框的 交集/并集，判断检测是否正确的阈值，通常为 0.5。
P	每张图像中被检测出的正确目标占总目标数的多少。
AP	对于一个类别的平均精度，图像个数/总精度和。
MAP	所有类别的平均精度和/总类别数。
AP50...	AP50代表 IoU 取 0.5，AP60代表 IoU 值取 0.6。数值越高越难。
ROI	Region of Interest，有很大可能性包含检测目标的区域。
Anchor	预先设定在图像上的密集方框，用于后续检测标记。
Region Proposals	建议区域，经过 Region Proposal Network(RPN) 得到一个 region 的 p≥0.5，则这个 region 中可能具有目标，这些选出来的区域被称为 ROI（Region of Interests）。RPN 同时会在 feature map 上框定 ROI 大致位置，输出 Bounding-box。
one-stage	一步检测器，指从图片到检测结果一步到位。（*e.g.* YOLO，SSD）
two-stage	两步检测器，指分两步走，先从图片提取 ROI，再进行检测。（*e.g.* RCNN，FPN，*etc.*）
skeleton	骨骼点，常见于行为检测数据集，标记人体几个重要位置的数据。

Re-ID	行人重识别，利用计算机视觉技术判断图像或者视频序列中是否存在特定人的技术。
backbone	图像特征提取器，往往是目标检测的第一步，常用 ResNet
head	分类+定位器
neck	插在 backbone 和 detection head 之间的模块，使网络更好地融合/提取 backbone 给出的特征，提高网络性能，例如：FPN，NAS-FPN，PAN，ASFF，RFB，SPP。这部分是科研的主攻点。
NMS(Non Maximum Suppression)	非极大值抑制，是目标检测框架中的后处理模块，主要用于删除高度冗余的bbox