中国农业大学
China Agricultural University

人工智能——图片检测和深度学习调优

胡标

目录
Contents

计算机视觉主要任务和应用场景

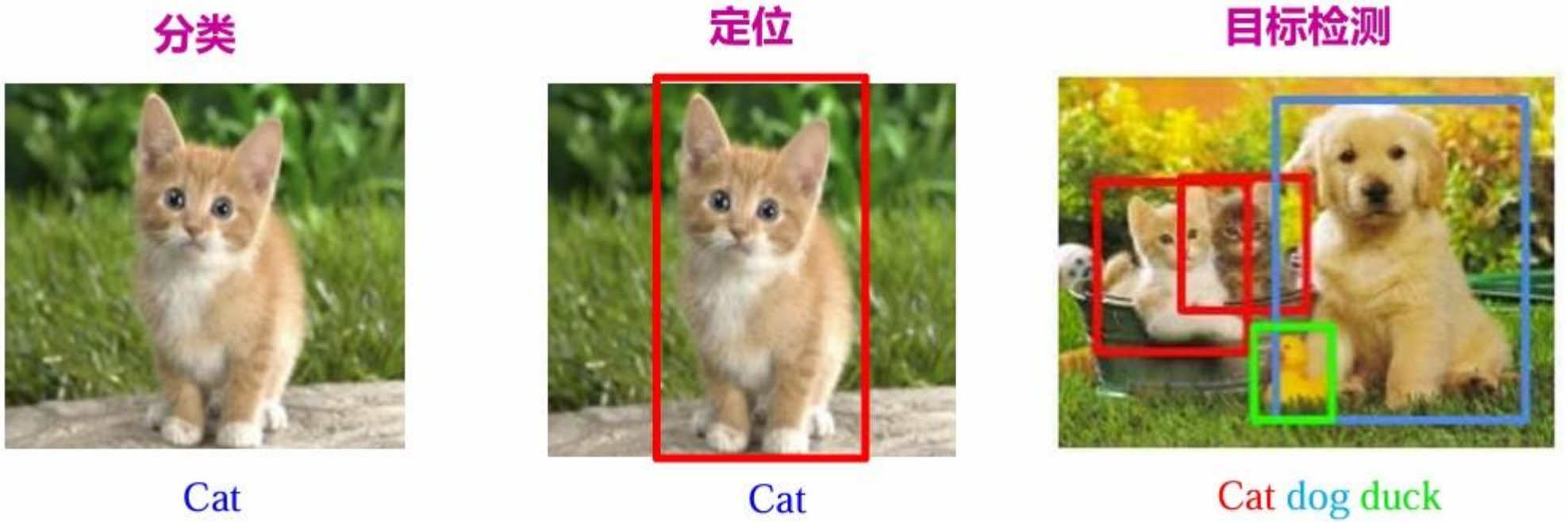
计算机视觉的任务就是让机器学会去“看”。研究者尝试着从不同的角度去解决这个问题，由此也发展出一系列的子任务：

1. 图像分类（Image Classification）

用于识别图像中物体的类别，为图像赋予一个或多个语义标签，解决 what 问题

计算机视觉的任务就是让机器学会去“看”。研究者尝试着从不同的角度去解决这个问题，由此也发展出一系列的子任务：

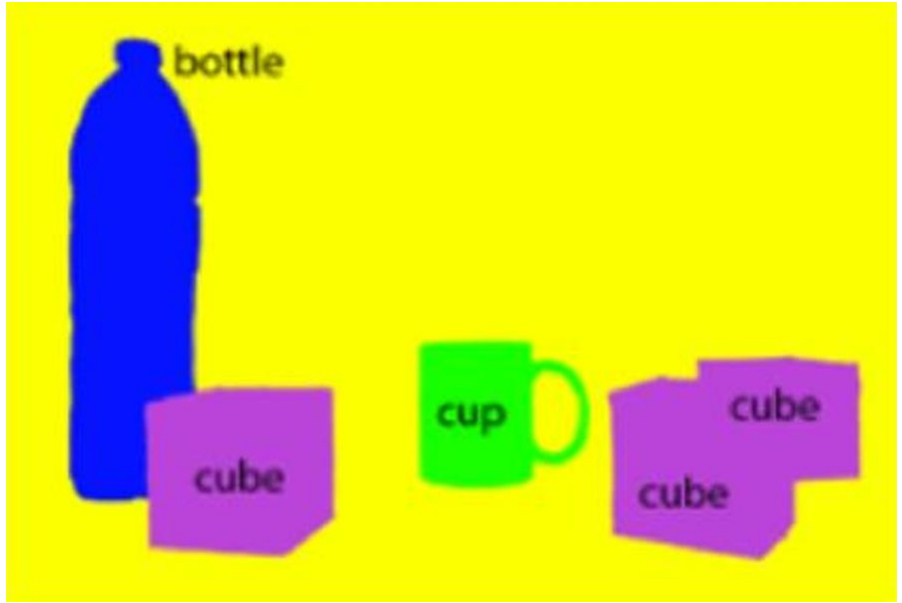
2. 目标检测（Object Detection）



引入矩形框，找到图像中多个物体的类别及所在位置，解决 what & where 问题

3. 语义分割（Semantic Segmentation）

- 用于标出图像中每个像素点所属的类别，语义分割不区分实例，只考虑像素类别，属于同一类别的像素点用一个颜色标识

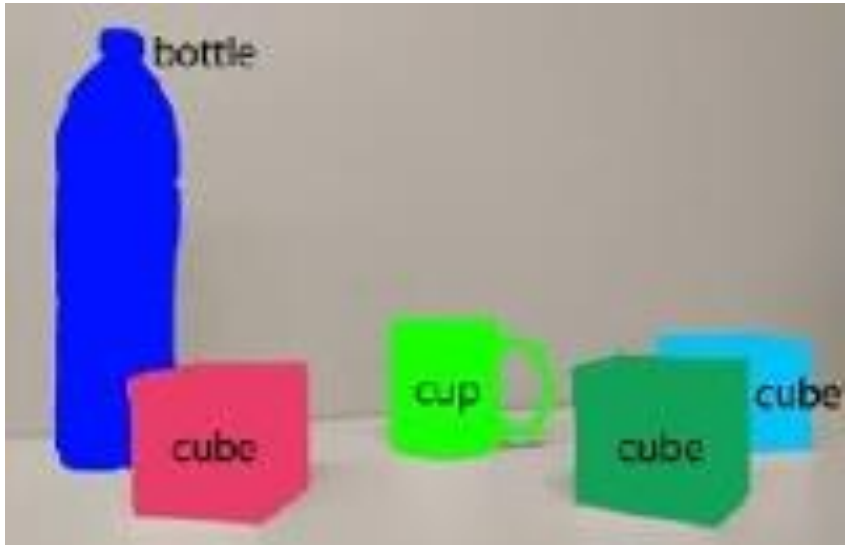


- 确定图像中物体的类别并精确勾勒出其所在位置，解决 what & where 问题
- 目标检测是标出物体的矩形框就可以了，但图像语义分割是在像素级进行识别。标出物体每个像素是属于物体还是背景等，类似于抠图

- 做分割要提供像素级的标注，而做检测只要提供矩形框标注即可。当然，有像素级标注，既可以做检测也可以做分割

4. 实例分割（Instance Segmentation）

- 实例分割任务和语义分割类似，但如果多个同类物体存在时，要将它们一一区分出来，解决 what & where 问题



- 语义分割不区分实例，只考虑像素类别
- 实例分割不但要进行像素级别的分类，还需在具体的类别基础上区别开不同的实例

- 从图像分类、目标检测，到语义分割和实例分割，是由粗粒度到精细粒度的计算机视觉任务

目标检测应用场景

消费娱乐

新零售货架商品检测

图片、视频审核

智慧交通

行人检测

车辆检测

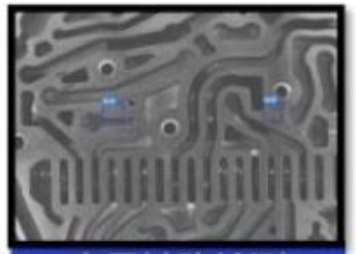
卫星遥感

地块检测

遥感目标检测

生产质检

零件计数



产品缺陷检测

设备巡检

表计巡检

设备状态监控

人脸检测

智慧医疗

眼底病变检测

肺炎检测

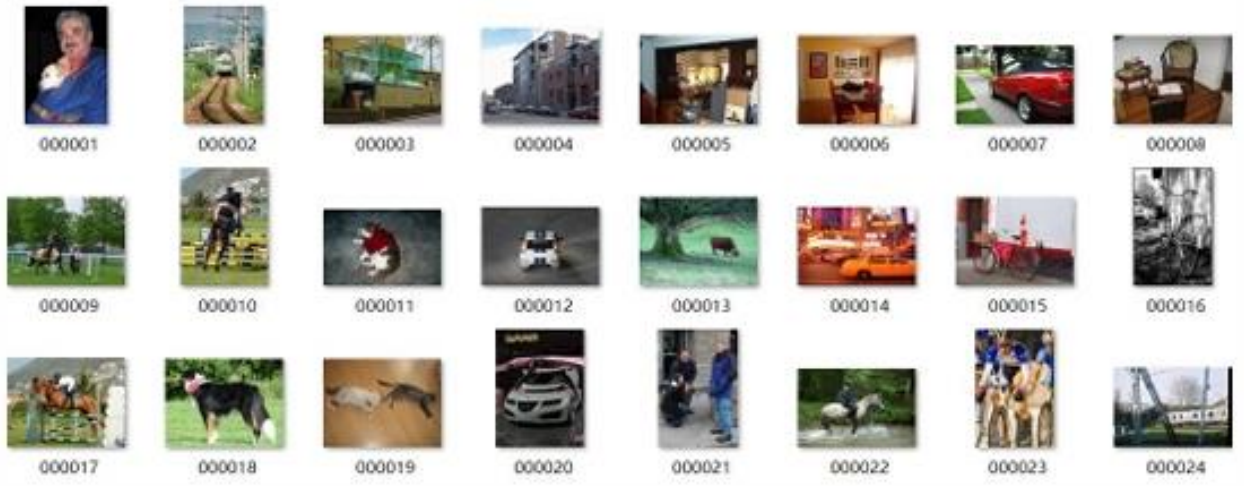
厂区安防

工服安全帽识别

目录
Contents

PASCAL VOC（The PASCAL Visual Object Classification）

- 目标检测，分类，分割等领域一个有名的数据集。从2005到2012年，共举办了8个不同的挑战赛。PASCAL VOC数据集包含20个类别，被看成目标检测问题的一个基准数据集
  - VOC2007中包含9,963张图片，共24,640个物体
  - VOC2012中包含11,540张图片，共27,450个物体




```
<annotation>
    <folder>VOC2007</folder>
    <filename>000024.jpg</filename>
    <source>
        <database>The VOC2007 Database</database>
        <annotation>PASCAL VOC2007</annotation>
        <image>flickr</image>
        <flickrid>322409915</flickrid>
    </source>
    <owner>
        <flickrid>knautia</flickrid>
        <name>Sarah</name>
    </owner>
    <size>
        <width>500</width>
        <height>335</height>
        <depth>3</depth>
    </size>
    <segmented>0</segmented>
    <object>
        <name>train</name>
        <pose>Unspecified</pose>
        <truncated>1</truncated>
        <difficult>0</difficult>
        <bndbox>
            <xmin>196</xmin>
            <ymin>165</ymin>
            <xmax>489</xmax>
            <ymax>247</ymax>
        </bndbox>
    </object>
</annotation>
```

MS COCO（Microsoft Common Objects in Context ）

- COCO数据集是Microsoft制作收集用于Detection + Segmentation + Localization + Captioning的数据集。COCO数据集共有12个大类，80个小类。
  - MSCOCO2014数据集：训练集：82783张，验证集：40504张，共计123287张
  - MSCOCO2017数据集：训练集：118287张，验证集：5000张，共计123287张


```
annotation{
    "id" : int,
    "image_id" : int,
    "category_id" : int,
    "segmentation" : RLE or [polygon],
    "area" : float,
    "bbox" : [x,y,width,height],
    "iscrowd" : 0 or 1,
}

categories[{
    "id" : int,
    "name" : str,
    "supercategory" : str,
}]
```

Object365（密集标注）

- 2019年，旷视科技发布了通用物体检测数据集Objects365，包含63万张图像，覆盖365个类别数量，边界框高达1000万个。

目标检测概述

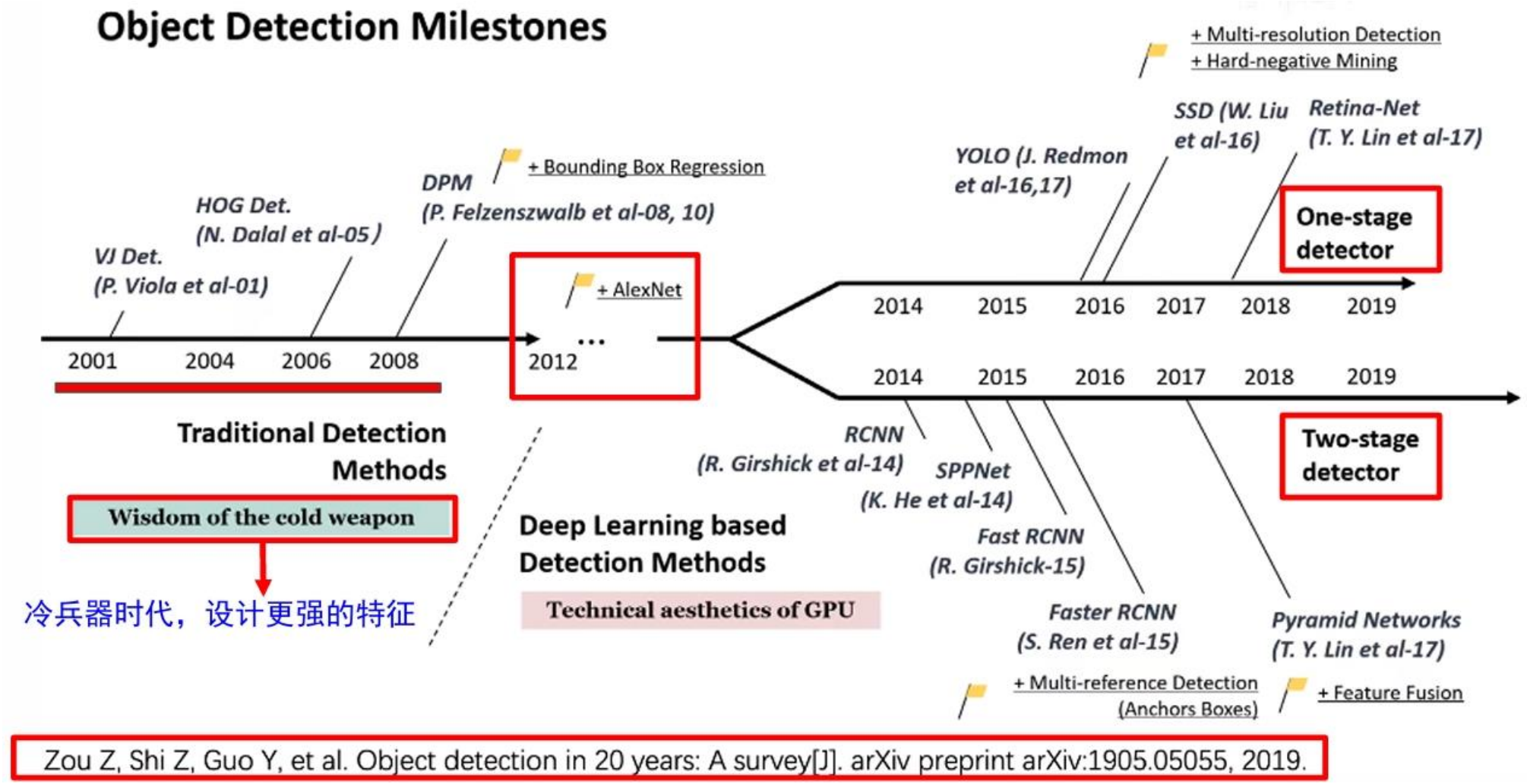
- 目标检测包括目标分类和目标定位 2 个任务。目标定位一般是用一个矩形的边界框来框出物体所在的位置

- 目标检测难点：物体的尺寸变化范围很大，摆放物体的角度，姿态不定，而且可以出现在图片的任何地方

目录
Contents

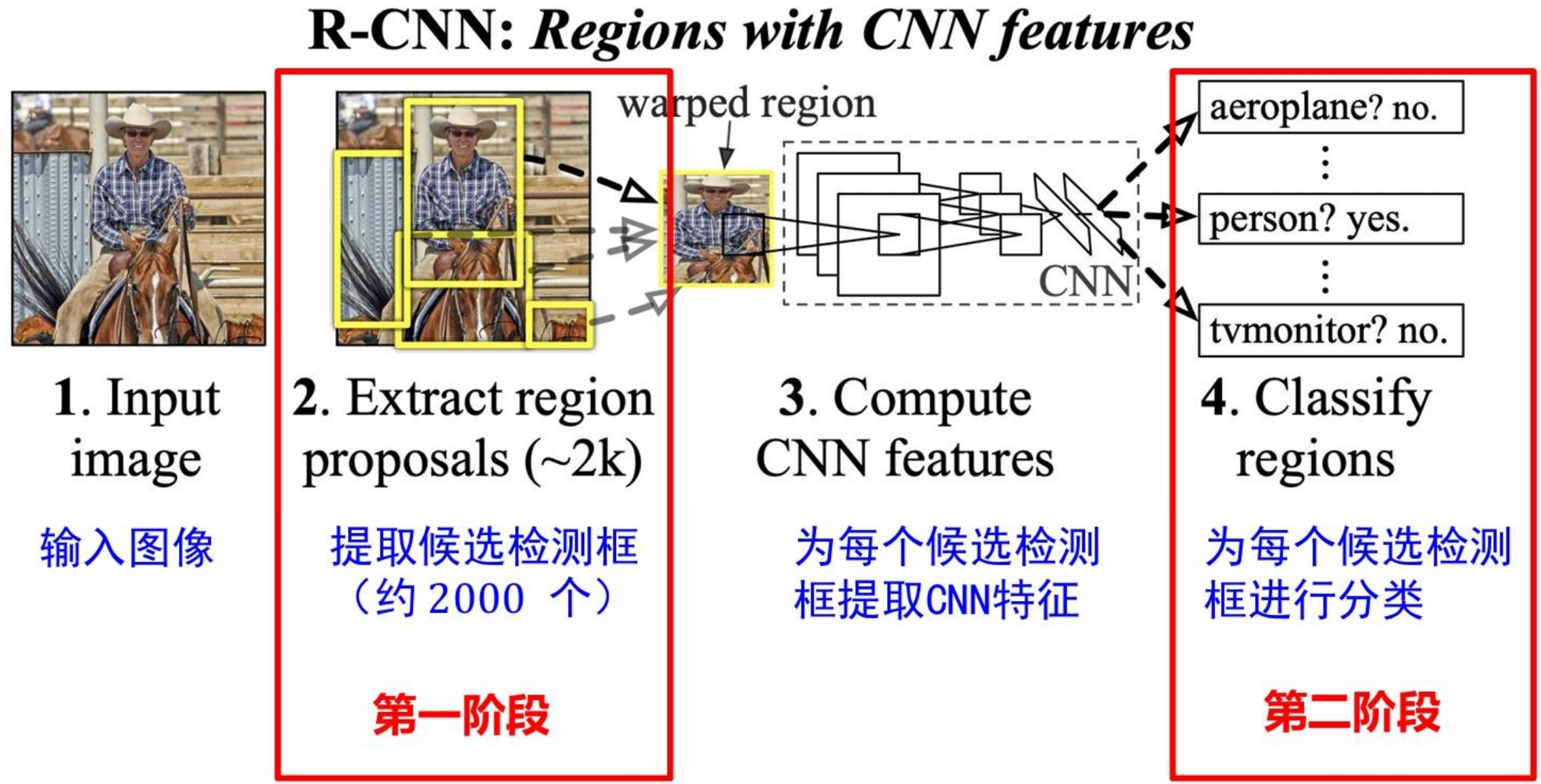
目标检测算法发展历史

- 目标检测算法分为两类：传统检测 和 基于深度学习 的目标检测算法



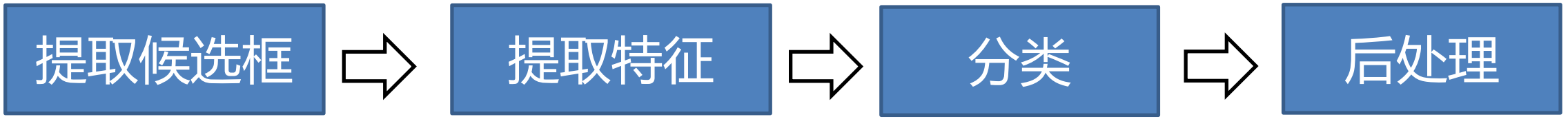
Zou Z, Shi Z, Guo Y, et al. Object detection in 20 years: A survey[J]. arXiv preprint arXiv:1905.05055, 2019.

二阶段目标检测算法：R-CNN



二阶段目标检测算法

结合图像分类的成功经验，可以将目标检测任务进行拆分，先从图像中提取若干候选框（矩形框），再逐一对候选框进行分类、甄别和调整候选框坐标



使用卷积神经网络对一幅图像进行分类不再是一件困难的事情。关键就是如何产生候选框！

一阶段目标检测算法

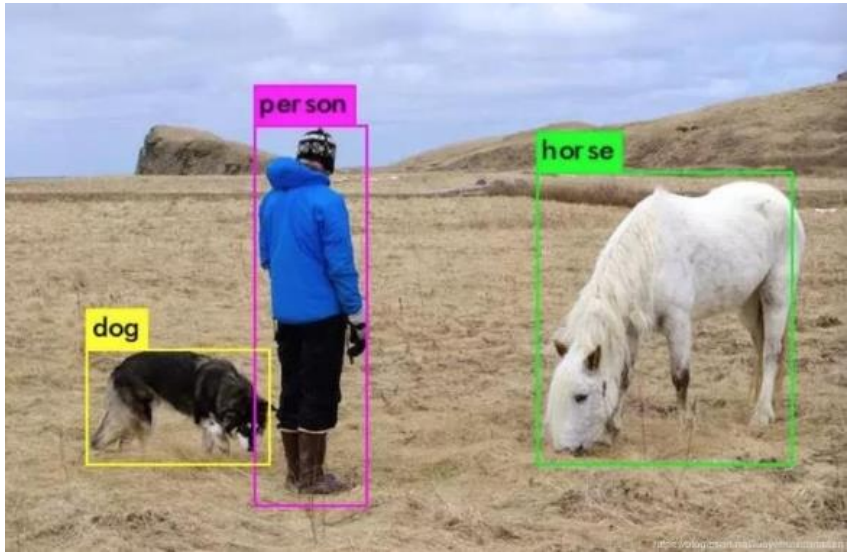
- 一阶段目标检测算法是不提取候选框，直接把全图输入到模型里，直接输出目标检测结果
- 属于端到端（输入原始数据，输出最后结果）系统，一步到位。直接在网络中提取特征来预测物体分类和位置

一阶段目标检测算法

- 两类目标检测算法各有利弊。二阶段算法因为先提取和筛选候选框，所以精度较高，但效率低；而一阶段算法实时性强，但精度可能不如二阶段算法
- 但随着算法不断改进，一阶段算法的检测精度也越来越高
- 在介绍目标检测算法之前，先介绍一些跟检测相关的基本概念，如 边界框、真实框、预测框、交并比、非极大值抑制 等

目标检测基础概念 —— 边界框

- 真实框：在检测任务中，训练数据集的标签里会给出目标物体真实边界框所对应的坐标，也就是数据集中标注的框，即真实框(*ground truth box*)。简称：*GT box*



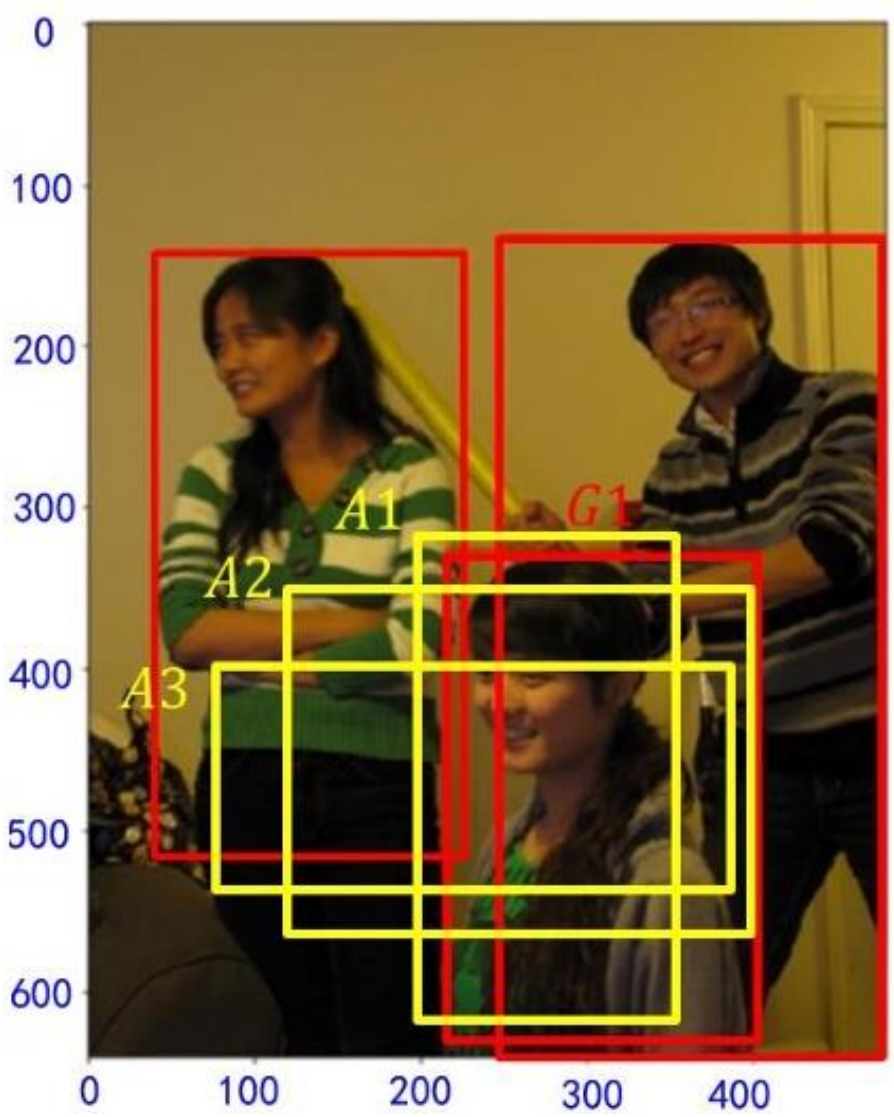
- 预测框：由模型预测输出的可能包含目标物体的边界框(*prediction box*)，也就是最终产生预测结果的边界框。简称*prediction box*
- 无论是真实框还是预测框，都属于边界框（*bounding box*）

目标检测基础概念 —— 边界框

- 要完成一项目标检测任务，输出需要哪些参数呢?
- 通常希望模型能够根据输入的图片，输出一些预测的边界框，以及边界框中所包含的物体的类别以及属于某个类别的概率值
- 因此目标检测任务的输出需要以下参数:
  1. 预测框坐标：$[x_1, y_1, x_2, y_2]$
  2. 图类别标签：L
  3. 所属类别的概率：P

目录
Contents

目标检测基础概念 —— 交并比

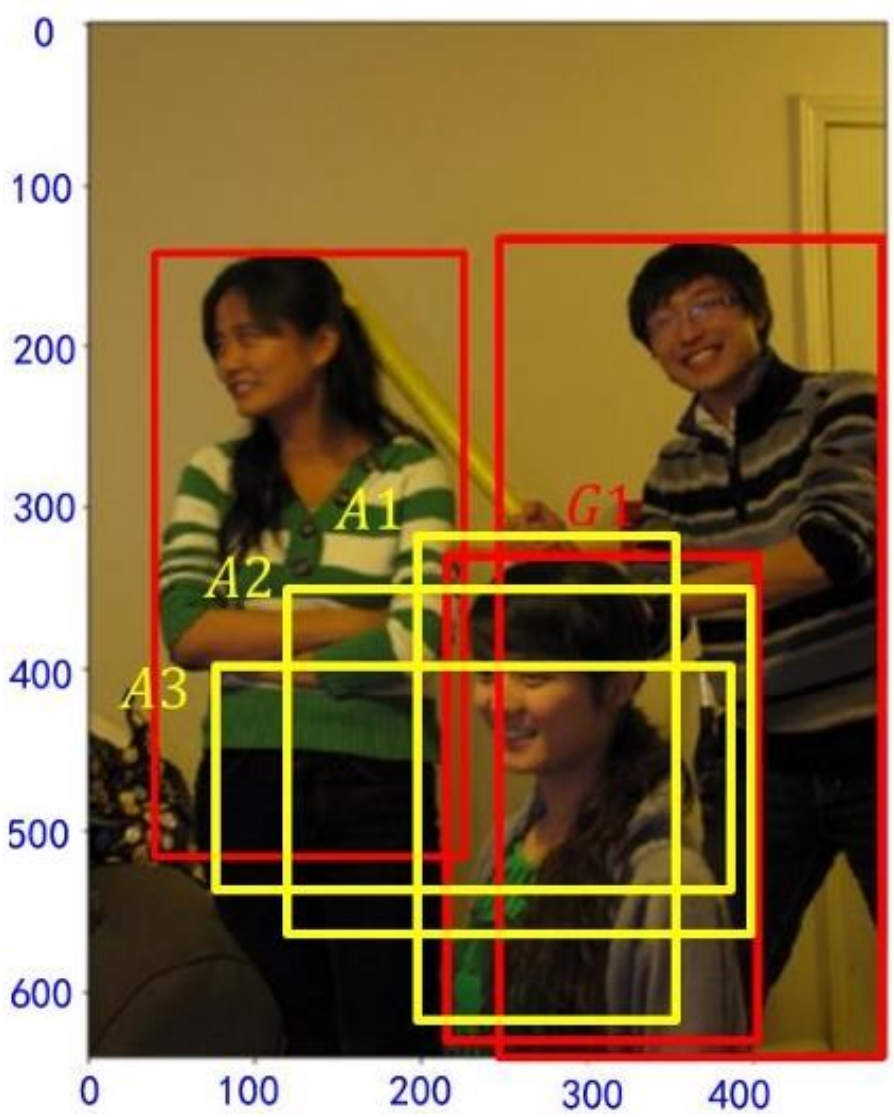


- 一张图片会对应多个预测框
- 生成的预测框与真实框重合程度如何衡量呢?

- 可以看到预测框$A1$与真实框 $G1$ 的重合度比较好。那么如何衡量这三个预测框跟真实框之间的关系呢

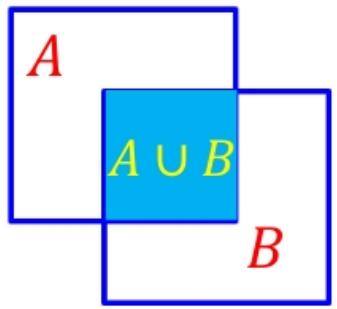
- 在检测任务中，使用 交并比（Intersection over Union，IoU）作为衡量两个框重合程度的指标

目标检测基础概念 —— 交并比

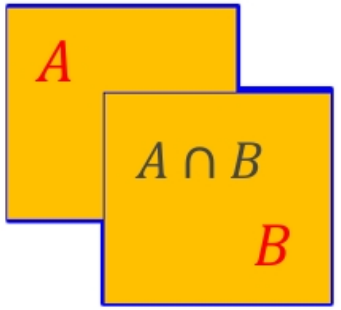


- 在检测任务中，交并比（Intersection over Union，IoU）作为衡量两个框重合程度的指标。类似数学中的集合，用来描述两个A 和B 矩形框之间的重合度关系

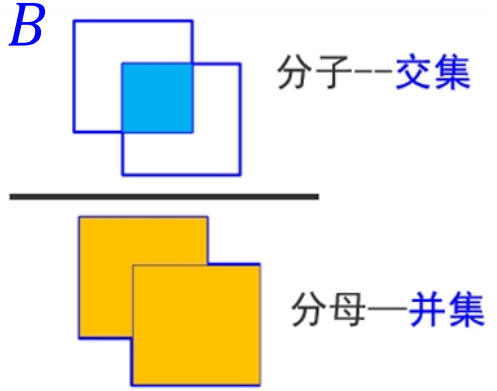
$$IoU = \frac{A \cap B}{A \cup B}$$



交集

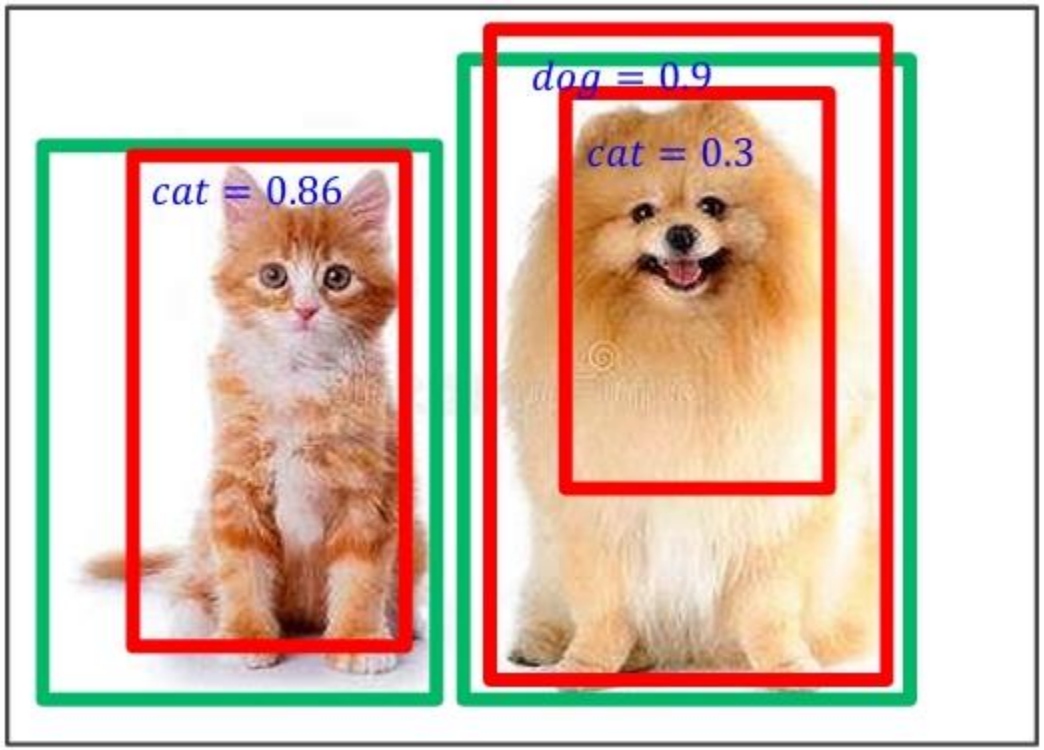


并集



交并比

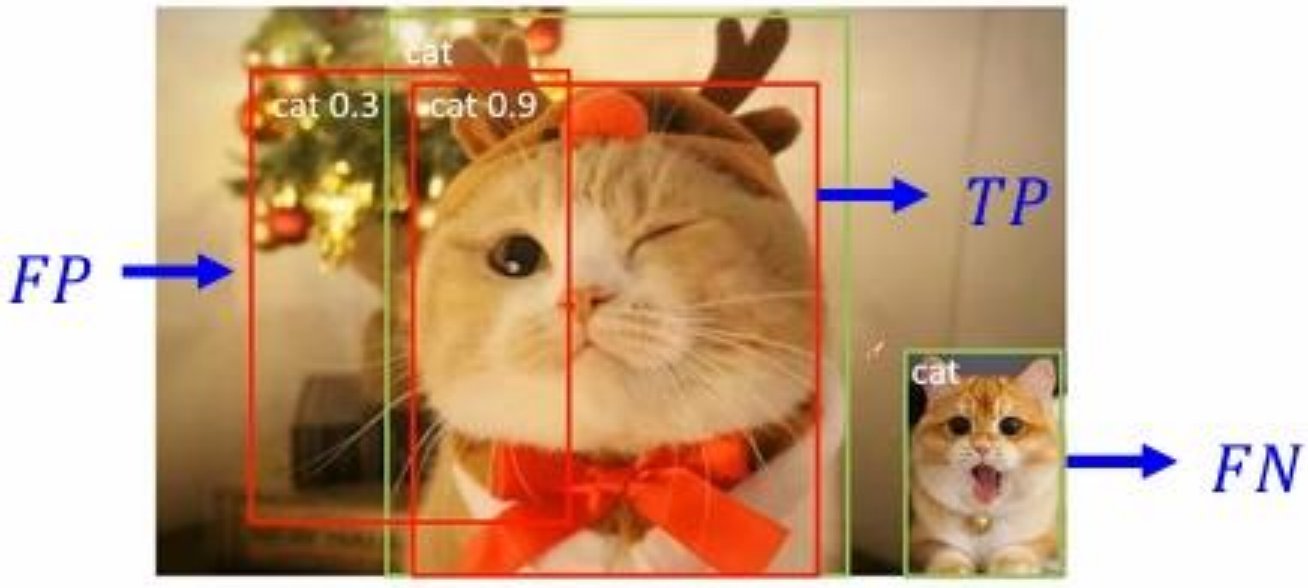
目标检测中常见指标



- 图像的目标检测和分类评判标准是有差别的。对于目标分类，通常用分类准确度来衡量好坏；但目标检测涉及分类和矩形框两个因素
- 在目标检测中，采用什么样的量化指标来表示检测的好坏呢？绿框代表人工标注的真实框，红框代表预测框

- 通过计算预测边界框和真实框的IoU，指定一个阈值来判断是否检测正确
- 还是要加上预测边界框的类别信息
- 预测目标的概率（置信度）

目标检测中常见指标



- TP（$True\ Positive$）：IoU>0.5的检测框数量
- FP（$False\ Positive$）：IoU≤0.5的检测框数量
- FN（$False\ Negative$）：没有检测到真实框的数量

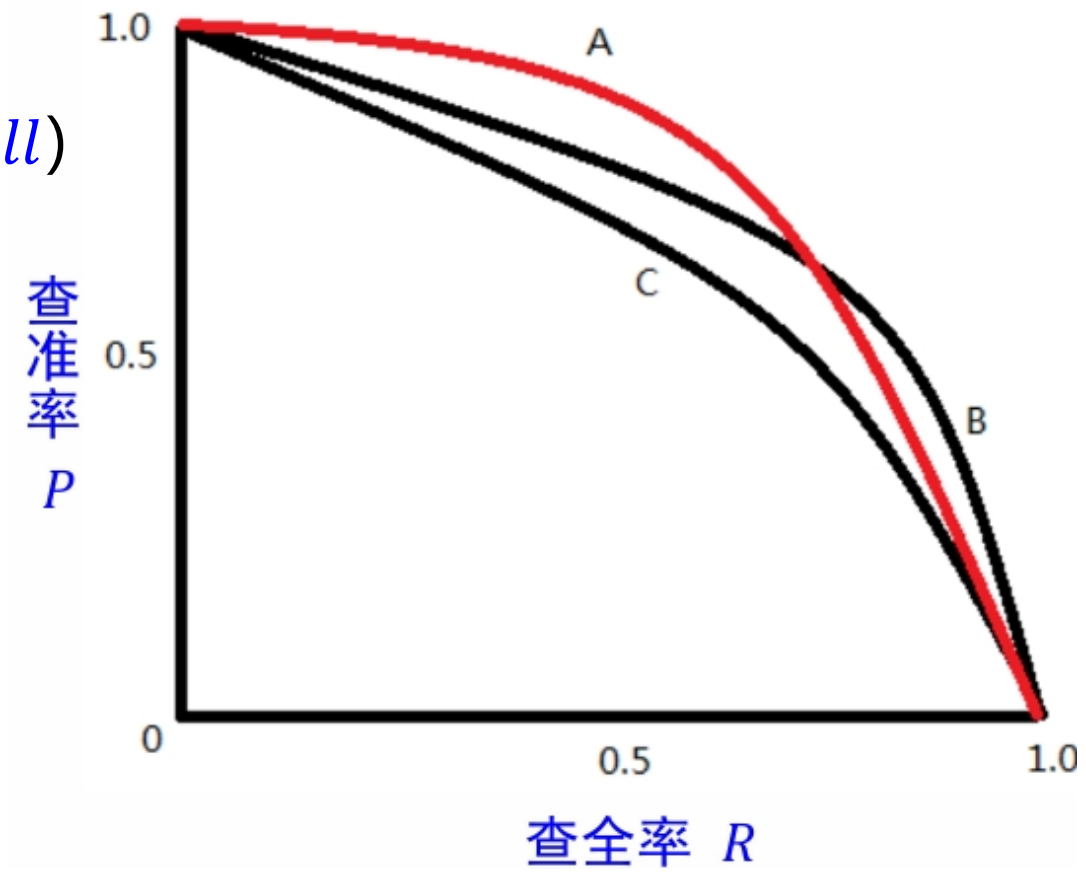
- $Precision$（查准率）：$TP/(TP+FP)$模型预测的所有目标中，预测正确的比例
- $Recall$（查全率）：$TP/(TP+FN)$所有真实目标中，模型预测正确的目标比例

mAP 的含义

- 目标检测不仅需要输出目标的类别，还要定位出目标所在的位置。那么评估分类分类问题中的精确率、召回率、准确率这些简单的指标已经不能反映出目标检测中结果的准确度
- $mAP\ (Mean\ Average\ Precision)$ 就是用来衡量目标检测算法的常用指标
- 对于目标检测任务来说，每一个类别都可以计算出 $Precision$ 和 $Recall$，于是每个类都可以画出一条 $P-R$ 曲线，而曲线下的面积就是 AP 的值

mAP 的含义

- AP(Average Precision) 的值和$P-R$曲线下的面积是成正比的，如果一个类别下的AP越大，那么也就说明$P-R$曲线下的面积越大，也可以认为该类别在$Precision$、$Recall$上表现得更加好
- 查准率（$Precision$）和查全率（$Recall$）二者绘制的曲线称为$P-R$曲线
- mAP $(mean\ Average\ Precision)$：计算出所有类别的$AP$，再求它们的平均值（一个类别对应一个 AP，目标检测往往有多个类别，所以有多个AP



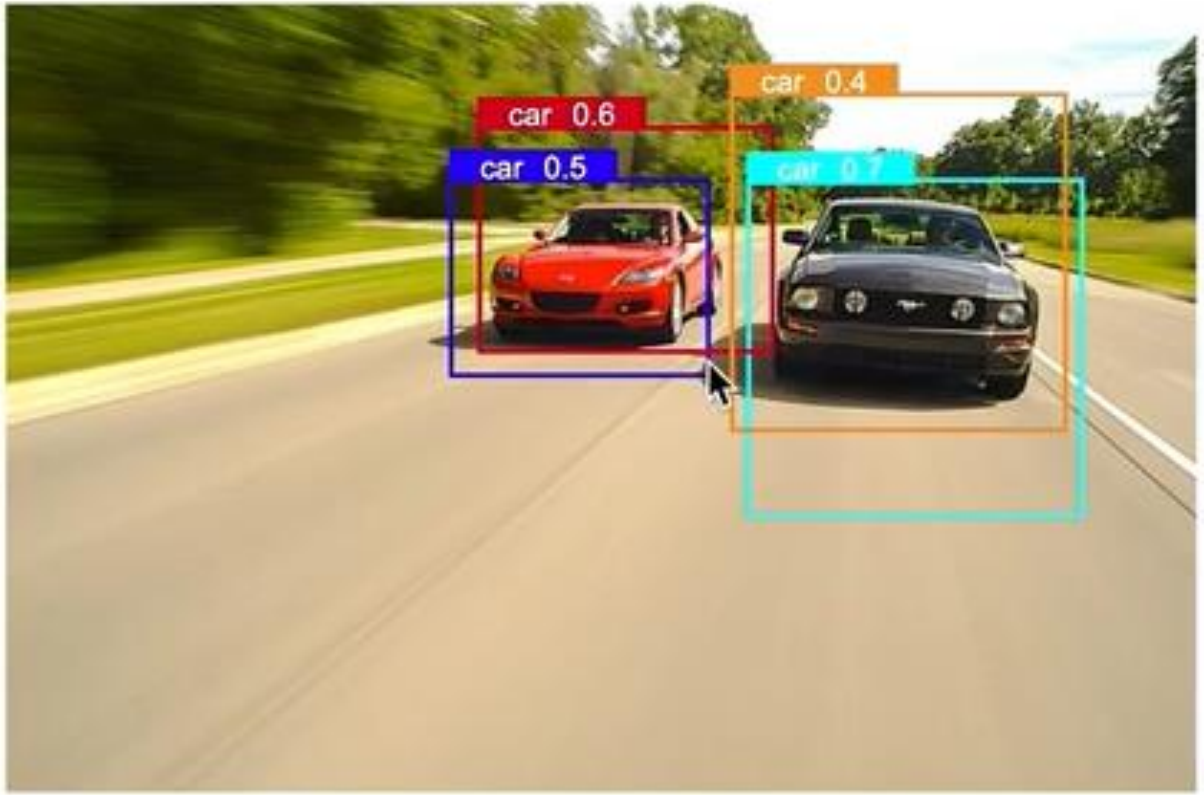
非极大值抑制

- 目标检测的过程中在同一目标的位置上会产生一些预测框，这些预测框相互之间可能会有重叠，如何消除冗余的预测框
- 为了简化输出，可以使用非极大值抑制（non-maximum suppression，NMS）合并属于同一目标的类似的预测边界框

非极大值抑制

1. 根据置信度排序： 按照预测框的置信度分数从高到低排序，优先处理高置信度的边界框。
2. 选取最高分数的框： 选出置信度最高的边界框作为当前目标框，并将其加入最终结果。
3. 计算重叠（IoU）： 对所有剩余的框，计算它们与当前选中框的交并比（Intersection over Union, IoU）。
4. 抑制重叠框： 将所有与当前选中框的 IoU 高于设定阈值的框移除，因为这些框可能是重复的检测。
5. 重复步骤 2-4： 对剩下的框重复上述过程，直到没有框可选。

非极大值抑制



预测框	坐标（xmin, ymin, xmax, ymax）	置信度
bbox1	(198, 51, 323, 147)	0.6
bbox2	(186, 73, 296, 157)	0.5
bbox3	(304, 37, 446, 180)	0.4
bbox4	(311, 73, 453, 216)	0.7

保留	候选
	bbox1（0.6） bbox2（0.5） bbox3（0.4） bbox4（0.7）

非极大值抑制

step 1： 将候选框根据置信度进行排序，保留 置信度最大检测框

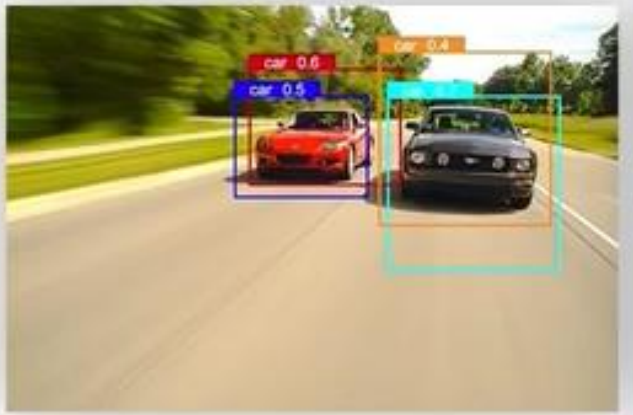
bbox4 > bbox1 > bbox2 > bbox3
0.7 > 0.6 > 0.5 > 0.4

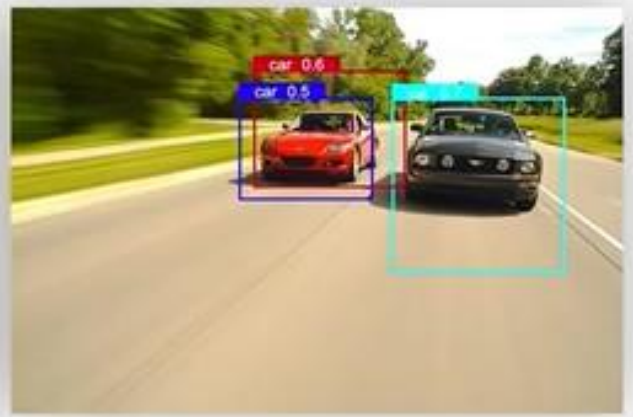
保留	候选
bbox4 （0.7）	bbox1（0.6） bbox2（0.5） bbox3（0.4）

step 2： 分别计算 保留框 与 所有候选框 的 IoU， 丢弃 IoU值大于 阈值(0.5) 的检测框

保留	候选			
bbox4 （0.7）	bbox1（0.6）	→	IoU = 0.0263	< 0.5
	bbox2（0.5）	→	IoU = 0	< 0.5
	bbox3（0.4）	→	IoU = 0.552	> 0.5 → 丢弃

保留	候选
bbox4 （0.7）	bbox1（0.6） bbox2（0.5）





非极大值抑制

step 3： 将 候选框根据置信度进行排序，保留 置信度最大检测框

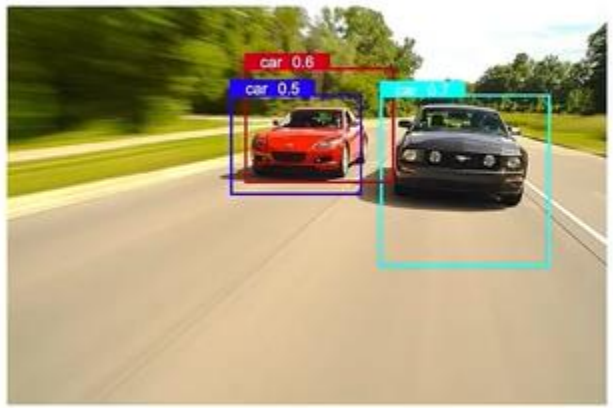
bbox1 > bbox2
0.6 > 0.5

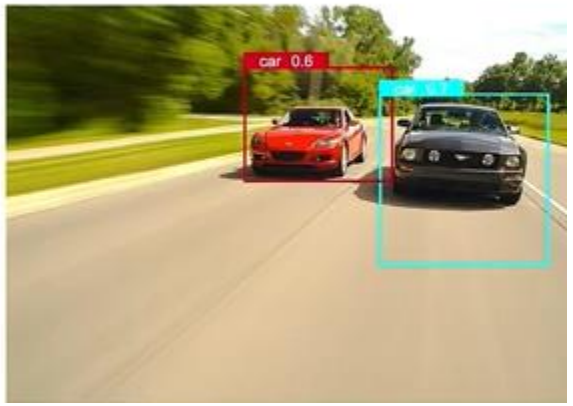
保留	候选
bbox4 （0.7）	bbox2（0.5）
bbox1（0.6）	

step 4： 分别计算 保留框 与 所有候选框 的 IoU， 丢弃 IoU值大于 阈值(0.5) 的检测框

保留	候选
bbox4 （0.7）	bbox2（0.5）
bbox1（0.6）	

bbox2（0.5） → IoU = 0.5184 > 0.5 → 丢弃





目录
Contents

调优基本原则-快速试错

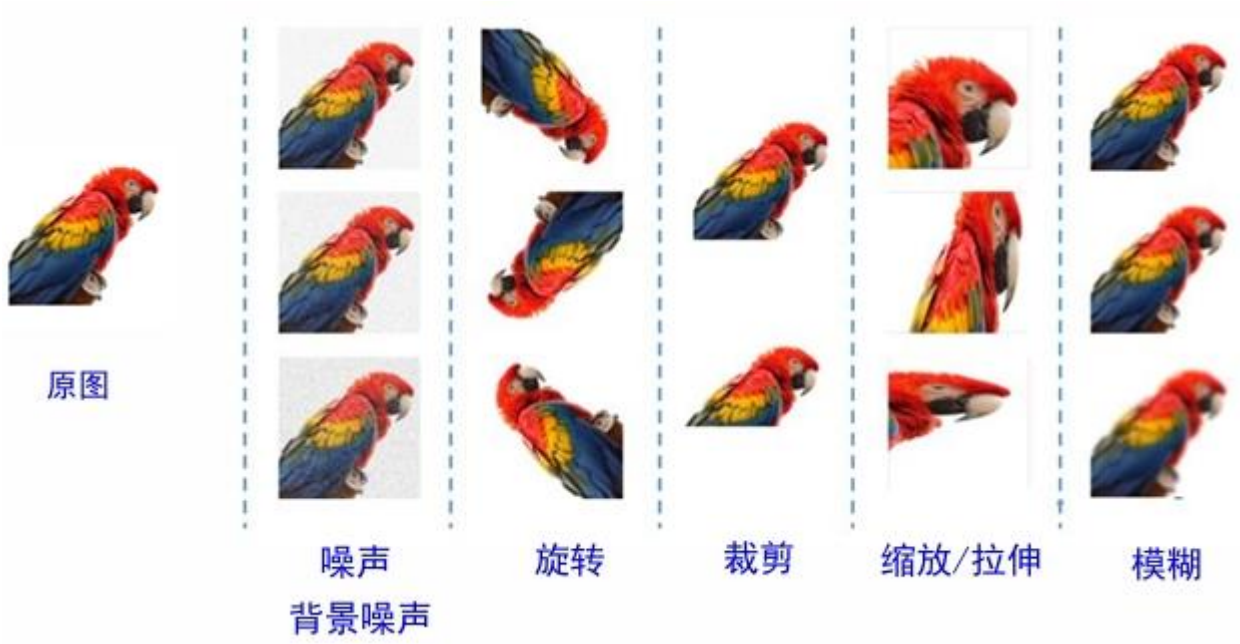
1. 由小数据到大数据
   - 因为训练数据成千上万，所以迭代训练模型需要耗时很长时间。可以先跑一个 $batch$ 的数据，快速验证训练脚本的流程是否正确，防止训练后期出错。
2. $Loss$设计要合理
   - 一般来说分类使用 $softmax$，回归使用均方误差$rMSE$
3. 观察$Loss$胜于观察准确率
   - $Loss$下降较稳定，而准确率有时是突变的（每个 $batch$ 不一样），不能反映真实情况

4. 学习率设置是否合理
   - 太大：$Loss$ 爆炸，在极值点附近来回震荡；太小：到不了极值点
5. 对比训练集和验证集的 $Loss$
   - 判断是否过拟合，训练是否足够，是否需要 early stop
   - 如果训练集和验证集的 $Loss$都在下降，说明模型还没有训练充分，应该继续训练
   - 如果验证集的$Loss$在上升，说明出现了过拟合。可采用早停止（ early stop ）策略

调优步骤

数据处理

- 数据处理最主要的方法就是做数据增强。一是现有的数据集数据量偏小，二是数据对物体的描述不全面
- 所以做数据增强的好处：一是增加数据，二是使得模型更加鲁棒。模型可以看到更多的数据变化，不至于遇到有噪声的情况模型性能下降
- 数据增强方法
  - ① 随机翻转
  - ② 随机旋转
  - ③ 随机改变亮度、对比度、饱和度



数据增强效果

1. 增加数据量
2. 采集更多的图像特征
3. 使网络可见更多的数据变化，提高模型泛化能力

调优步骤

- 网络模型主要从下面三部分来改进

1. 模型深度加深
   - 比如采用 ResNet 网络，从 18 层到 152 层
2. 模型宽度加宽
   - 比如 Inception 结构，可以导致模型宽度加宽
3. 预训练模型
   - 所谓预训练模型，是在一个大数据集上预先训练好一个模型，然后微调以适应当前自己的小数据集

预训练通常是在一个大数据集上进行

ImageNet(分类模型)

MSCOCO(检测模型)

而微调是固定前面的卷积层，而只微调全连接层。或者微调后面的 卷积层 + 全连接层

调优步骤

网络训练需要注意的问题

- 数据处理和改进这两部分完成以后，参数调优就很重要了

1. 防止梯度消失、梯度爆炸
2. 防止网络过拟合
3. 防止网络训练不稳定、不收敛

梯度消失和爆炸

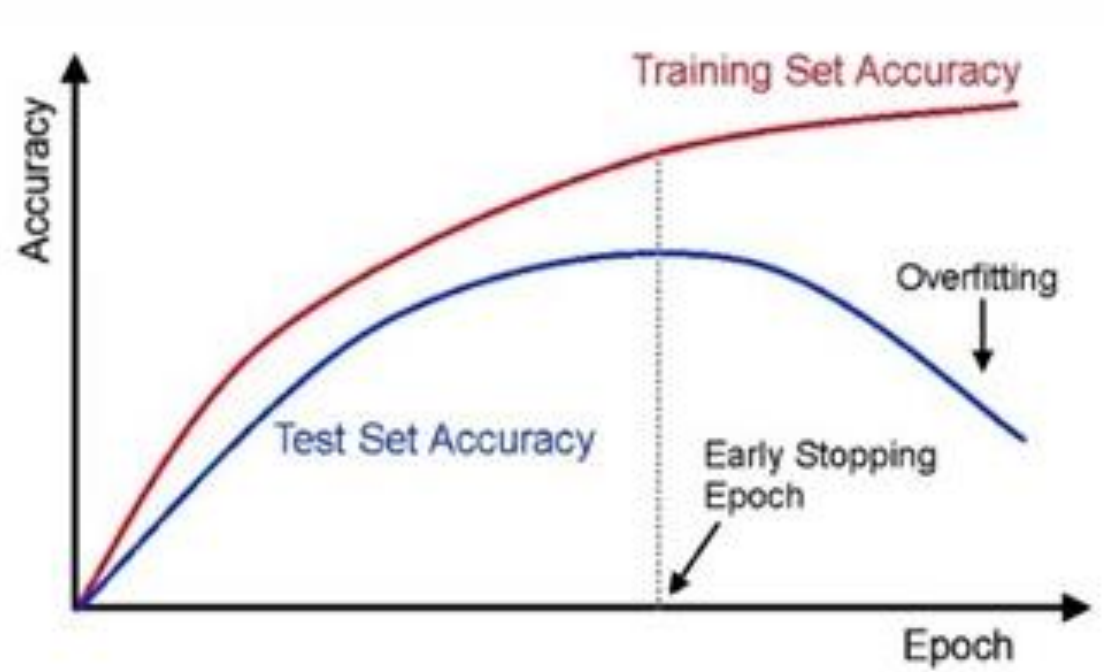
- 梯度消失和梯度爆炸都是因为网络深度而造成的。随着网络不断加深，梯度在反向传播过程当中不断累乘，如果梯度过小，随着层数的增加，梯度累乘越来越小直至消失，反之产生梯度爆炸
- 防止梯度消失和梯度爆炸通常采用的方法：
  ① 更换激活函数
  ② ResNet Block
  ③ Batch Normalization（在每次卷积之后和激活之前）
  ④ 梯度截断
  ⑤ 预训练 + 微调

梯度消失和爆炸

- 梯度截断（Gradient Clipping）
  - 梯度截断是处理梯度爆炸的方法。具体是检查误差梯度的值是否超过阈值，如果超过，则截断梯度，将梯度强行设置为阈值
  - 梯度截断可以一定程度上缓解梯度爆炸问题
- 预训练 + 微调
  - 就算发生了梯度消失（梯度反向传播不到前面几层网络），现有的参数也可以保证模型有不错的结果
  - 微调就是调后面几层卷积和全连接层，使模型适应自己的数据集

防止网络过拟合

- 造成过拟合的原因是网络模型复杂，而数据特征模式较简单，导致模型过度拟合数据集。
- 防止过拟合的常用方法：
  - ① 数据增强
  - ② 提前停止
  - ③ 权重正则化
  - ④ Batch Normalization
  - ⑤ dropout



提前停止（early stop）

调优步骤

后处理

常见后处理方法

1. 模型融合

- 可以分别训练三种模型，比如VGG、ResNet、GooglNet，这三种网络分别对测试集数据打分
- 对同一张图片，三种网络都会给出该图片属于哪一类的结果，最后采用投票的方式确定图片属于哪一类

2. 测试结果融合

- 对于分类来说，最后一层都是采用softmax。给出概率分布的结果那么 VGG、ResNet、GooglNet，这三种网络都会给出图片分类的概率分布。把这三种模型的概率分布加在一起，最后用最大值方式给出最终模型的结果

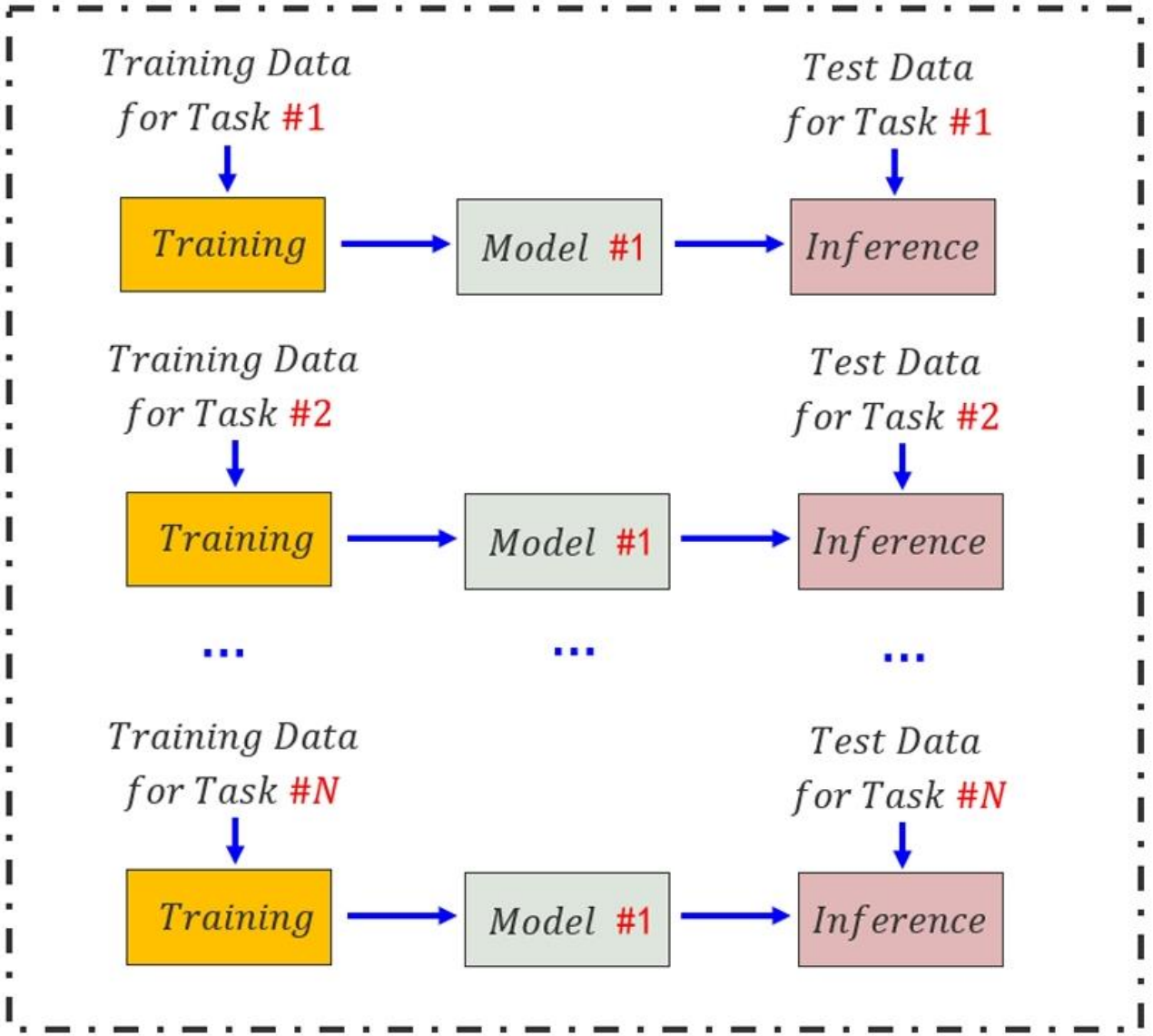
目录
Contents

深度学习：十年发展回顾

从 2012年 AlexNet 模型发布以来，深度学习经过 10 年的发展，研究范式（阶段）经历了三个境界

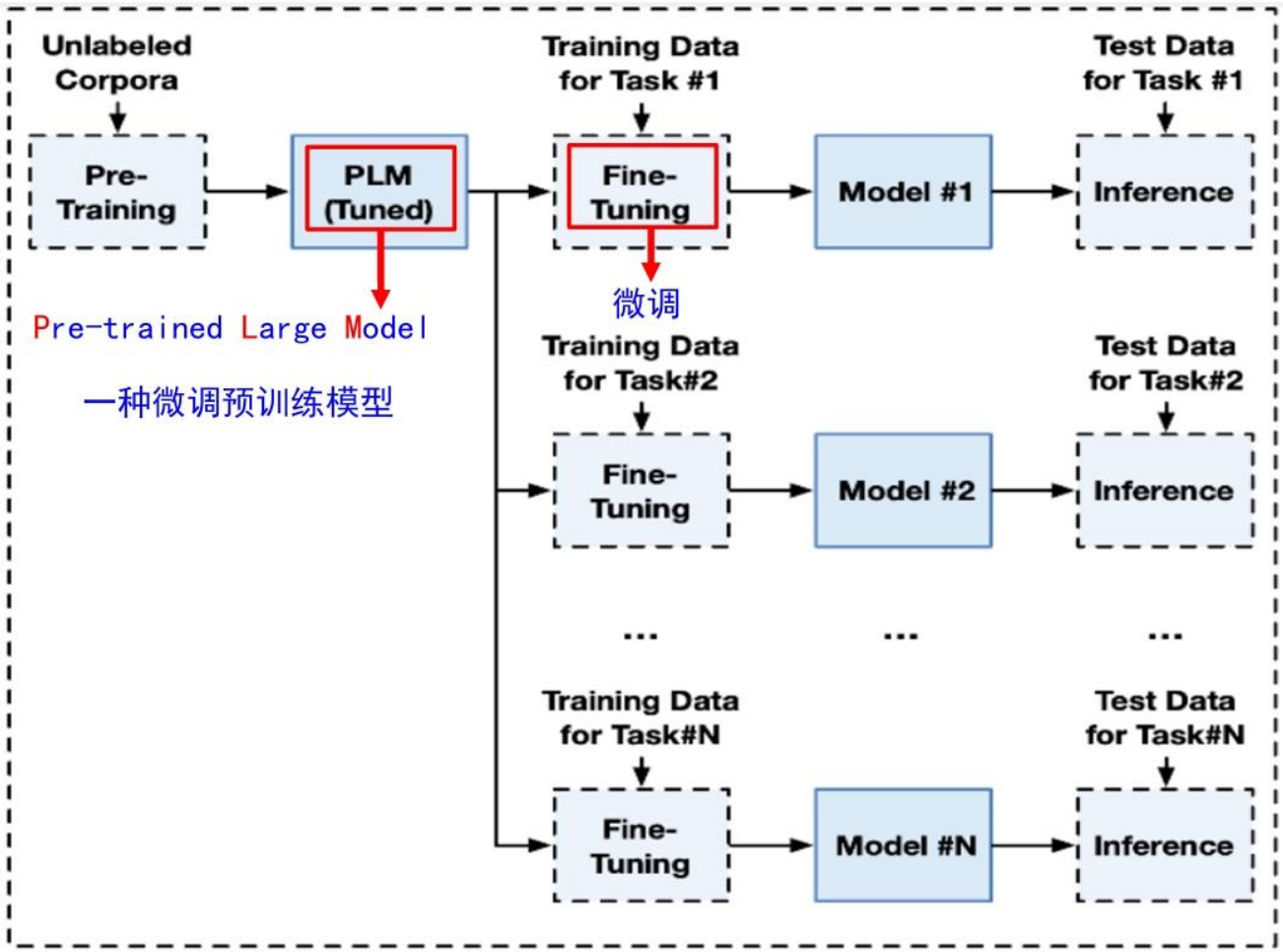
1. 深度学习第一个阶段：各自为战

- 白手起家 + 各家自扫门前雪。给定一个任务，设计一个模型，基本都是从头开
- 四种主流的深度学习网络：VGG、GoogleNet、ResNet 和 DenseNet



2. 深度学习第二个阶段：预训练大模型+大小联调（迁移学习）

- 预训练大模型采用两阶段学习
  - ① 首先在大型语料库中预先训练好一个大模型（PLM）
  - ② 然后在目标（下游）任务数据集上基于训练好的大模型进行微调（Fine-tuning），以获得适应下游任务的模型



2. 深度学习第二个阶段：预训练大模型+大小联调（迁移学习）

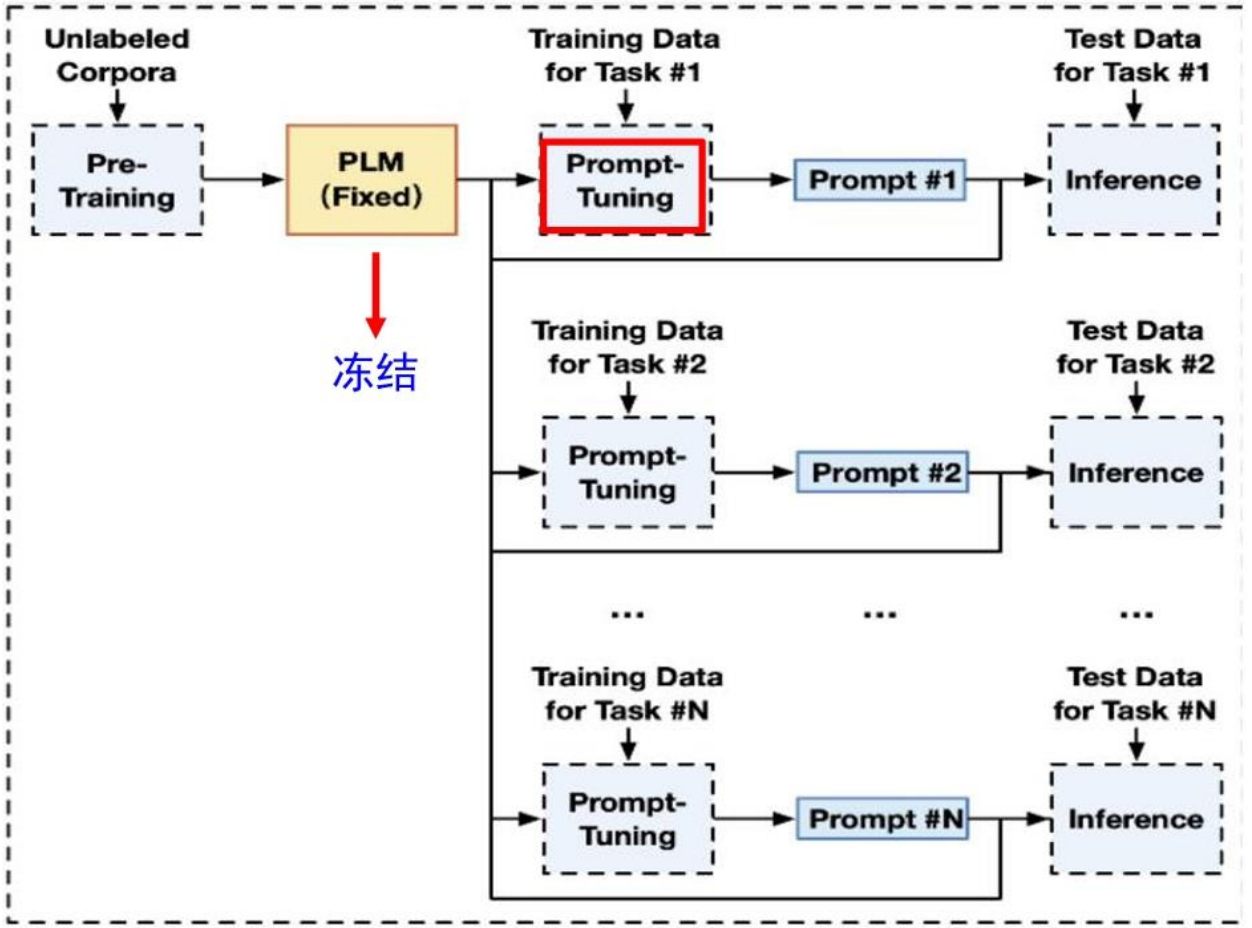
- 比如在图像视觉领域，大模型最常用的是 ResNet，语言大模型是Transformer 的各类变种，如 BERT 和 GPT 系列
- 一般来说，ResNet 和 Bert 都是 $Fine-Tune$ 最后几层，因为前几层提取了公共信息，比如 ResNet 前几层提取了颜色、纹理、形状等公共信息， BERT 前几层提取了词性，语义等公共信息。后几层往往和下游训练任务关联了
- 这种模式在诸多任务的表现上超越了传统的监督学习方法，不论在工业生产、科研创新还是竞赛中均作为新的主流方式

2. 深度学习第二个阶段：预训练大模型+大小联调（迁移学习）

- 但迁移学习模式也存在着一些问题。比如在大多数的下游任务微调时，下游任务的目标与预训练的目标差距过大导致提升效果不明显，微调过程中依赖大量的监督语料等
- GPT-3、PET 模型提出一种基于预训练语言模型的新的微调范式：Prompt-Tuning，可以让模型在小样本($few-shot$)或零样本 ($zero-shot$) 下达到理想的效果
  ① GPT-3 是 OpenAI 推出的超大规模语言生成模型，是 Generative Pre-trained Transformer 的缩写，模型训练参数量超千亿，执行各种语言处理任务，包括语言翻译、摘要和文本生成
  ② PET（Pattern-exploiting Training）是一种半监督训练，可将输入示例重新编写为填空样式的短语，在参数量少的环境下明显优于常规的监督训练

3. 深度学习第三个阶段：预训练巨模型 + 一巨托众小

- Prompt-Tuning 又称为提示学习
  - ① 提示学习就是给几个提示，给少量的训练样本就会做得很好
  - ② 主要思想通过模板将不同的下游任务转换为模型预训练时常见的形式，缩小预训练与微调时训练数据的差异性，降低了预训练模型在下游任务微调时存储和运算的资源使用，提升模型在下游任务中的表现



深度学习探图像，分类定位分割详。
调优增效方法妙，发展十年路宽广。