## 1.1算法基础

目标检测算法可以完成对于街道垃圾的实时监控，通过采集大量的街道垃圾数据集训练良好参数的识别模型，可以将其应用在街道的摄像头实时识别垃圾。

### 1.1.1目标检测算法

**目标检测是计算机视觉领域中的一个重要研究方向，也是其他复杂视觉任务的基础。**作为图像理解和计算机视觉的基石，目标检测是解决分割、场景理解、目标跟踪、图像描述和事件检测等更高层次 视觉任务的基础。

传统的目标检测一般使用滑动窗口的框架，主要包括三个步骤：利用不同尺寸的滑动窗口框住图中的某一部分作为候选区域；提取候选区域相关的视觉特征。比如人脸检测常用的Harr特征；行人检测和普通目标检测常用的HOG特征等；利用分类器进行识别，比如常用的SVM模型。

传统的目标检测中，多尺度形变部件模型DPM（Deformable Part Model）是出类拔萃的，连续获得VOC（Visual Object Class）2007到2009的检测冠军， DPM把物体看成了多个组成的部件（比如人脸的鼻子、嘴巴等），用部件间的关系来描述物体，这个特性非常符合自然界很多物体的非刚体特征。DPM可以看做是HOG+SVM的扩展，很好的继承了两者的优点，在人脸检测、行人检测等任务上取得了不错的效果，但是DPM相对复杂，检测速度也较慢，从而也出现了很多改进的方法。

基于深度学习的目标检测发展起来后，很多之前研究传统目标检测算法的研究者也开始转向深度学习。其实效果也一直难以突破。比如文献中的算法在VOC 2007测试集合上的mAP只能30%多一点，文献中的OverFeat在ILSVRC 2013测试集上的mAP只能达到24.3%。2013年R-CNN诞生了，VOC 2007测试集的mAP被提升至48%，2014年时通过修改网络结构又提升到了66%，同时ILSVRC 2013测试集的mAP也被提升至31.4%。

深度学习相关的目标检测方法也可以大致分为两派：基于区域提名的，如R-CNN、SPP-net、Fast R-CNN、Faster R-CNN、R-FCN；端到端（End-to-End），无需区域提名的，如YOLO、SSD。

R-CNN是Region-based Convolutional Neural Networks的缩写，中文翻译是基于区域的卷积神经网络，是一种结合区域提名（Region Proposal）和卷积神经网络（CNN）的目标检测方法。Ross Girshick在2013年的开山之作《Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation》奠定了这个子领域的基础，这篇论文后续版本发表在CVPR 2014，期刊版本发表在PAMI 2015。

R-CNN是第一个真正可以工业级应用的解决方案， 在此之后又先后出现了SPP-net、Fast R-CNN、Faster R-CNN、R-FCN、YOLO、SSD等研究。这些创新的工作其实很多时候是把一些传统视觉领域的方法和深度学习结合起来了，比如选择性搜索（Selective Search)和图像金字塔（Pyramid）等。

本解决方案选择Faster R-CNN作为Baseline，在此基础上进行修改，形成最终的解决方案。

### 1.1.2算法可行性分析

在目标检测任务的分支中，行人识别任务是从大量的摄像头角度的复杂街道背景下将其中的行人识别并标注，具有极其广泛的应用：智能辅助驾驶，智能监控，行人分析以及智能机器人等领域，这个领域目前的研究非常活跃。

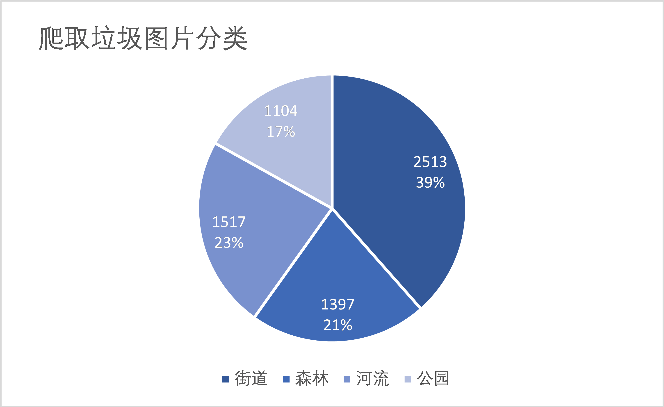
在现实应用中，可以认为多场景下的垃圾识别任务其背景、检测方式、任务复杂度与常规的目标检测任务（如自动驾驶、行人识别等）有较高的相似度，与此同时，由于街道场景相关的目标检测任务是大量存在的，在数据集方面可以通过爬虫大量获取与接到场景相关的目标检测任务数据集，如PASCAL VOC、MSCOCO、ILSVRC等，而目标检测算法在多个与本任务相关背景下，也有较高的识别准确率（在90%以上），因此认为对摄像头下的多场景垃圾目标检测任务也具有可行性。

## 1.2数据集

本项目的特点之一是采用弱监督学习来对数据进行预先标注，其中数据集的主要来源为自采集的街道垃圾数据集以及搜索引擎关键字爬虫获得，主要的场景有城市街道、河流、森林、公园等，预计在未来还可以将图像数据库扩充至更多场景如公共场所，基于多背景的图像可以保证在不同场景下保证采集到的图片的背景变化大，而垃圾的特征变化较小，可以有效地防止过拟合。将大量自采集图像和爬虫所获图像输入至基于弱监督学习模型的数据标注工具——WLM(Weak Label Master)，此工具为项目组自研发的高效数据自动标注工具，用于获取初代的垃圾检测数据库。在初代的机器标注数据库基础上，通过人工二次标注，建立可以用于模型训练的次代人工标注数据库，此数据库在用于训练弱监督学习模型的同时，也可用于训练垃圾目标检测的模型。采取这种方式可以获取大量的用于目标检测的垃圾图像数据，为后续扩充数据库，提供有力的技术支持。目前的数据库已有自采集图像1000张，网络爬取图像7000张，后续预计将扩充至20000张，强有力的数据支撑将大大增强模型的泛化能力。

### 1.2.1数据集采集

本项目的数据集主要有街道、河流、森林、公园、高山等背景下的垃圾图像。图像来源主要有两个部分：自采集，爬虫搜索引擎爬取。其中自采集的部分，由项目成员采集了403张城市街道背景下的垃圾图像；爬虫采集部分，通过Google, Baidu等搜索引擎爬取多背景下的垃圾图像6531张。图片的格式大多为.jpg或.png。其中图片中目标框的，同时采集数据的质量由于是爬虫采集，会出现过于模糊、遮挡严重或者目标大小不均匀，图像之间跨度较大，采集目标的大小无法根据使用的场景，接近应用场景的尺寸最佳，需要后续对图像进行处理，减小尺度不均匀对检测的影响。数据集的优点在于多场景下采集数据保证了图像变化的多样性，同时采集场景均为自然状态下的各种各样的照片，而不是人为地制造变化不大的图像数据，有利于防止模型过拟合。



图x：爬取垃圾图片类别

### 1.2.2数据集标注

首先通过探索性数据分析(Exploratory Data Analysis)将数据大致归类并对异常或难样本采取数据清洗，去除缺失值和噪声数据,得到训练可用的数据集。

1. **基于弱监督学习的数据自动标注**

监督学习通过学习大量训练样本来构建预测模型，其中每个训练样本都有一个标签标明其真值输出。尽管当前的技术已经取得了巨大的成功，但是值得注意的是，由于数据标注过程的高成本，很多任务很难获得如全部真值标签这样的强监督信息。因此，能够使用弱监督的机器学习技术是可取的。机器学习在各种任务中取得了巨大成功，特别是在分类和回归等监督学习任务中。预测模型是从包含大量训练样本的训练数据集中学习，每个训练样本对应一个事件或对象。训练样本由两部分组成：一个描述事件/对象的特征向量（或示例），以及一个表示真值输出的标签。在分类任务中，标签表示训练样本所属的类别；在回归任务中，标签是一个与样本对应的实数值。大多数成功的技术，如深度学习 ，都需要含有真值标签的大规模训练数据集，然而，在许多任务中，由于数据标注过程的成本极高，很难获得强监督信息，在图像分类任务中，训练数据的Groud-Truth标签由人类注释者给出; 虽然很容易从互联网上获取大量图像，而由于人工成本，只能注释一小部分图像。在面对大量数据集的情况下，传统的人工标注方式成本非常高，而且很耗时间，在某些数据集的标注中，需要领域专家才能够进行数据的精准、正确标注，这一问题则会更加明显。

为了解决这个问题，项目组基于弱监督学习，利用已经训练好的预训练模型，搭建一个可以用来预测垃圾标签信息的目标检测神经网络。则检测任务的类型类似于若监督学习中的不完整监督：部分样本label缺失。将获取的数据集均看作样本label的缺失，不完全监督考虑那些我们只拥有少量有标注数据的情况。

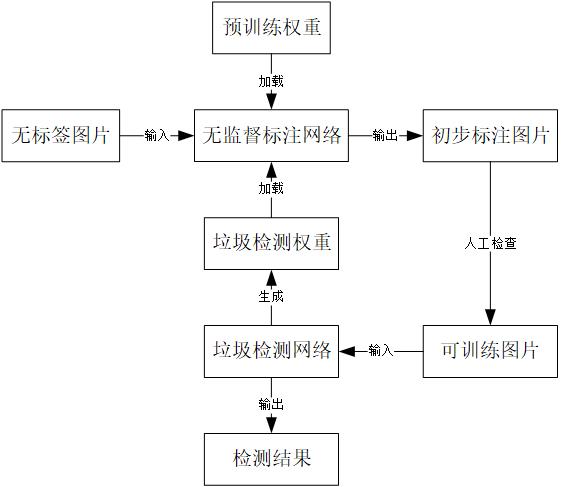
这些有标注数据并不足以训练出好的模型，但是我们拥有大量未标注数据可供使用。形式化表达为，模型的任务是从训练数据集中学习，其中训练集中有 n 个有标注训练样本（即给出的样本）和 u = m - n个未标注样本；其他条件与具有强监督的监督学习相同。将 n个有标注示例称为标注数据，将 u 个未标注示例称为未标注数据。

则由标注实例训练出的模型称为标注网络，其loss为对于多目标检测任务采用focal loss:

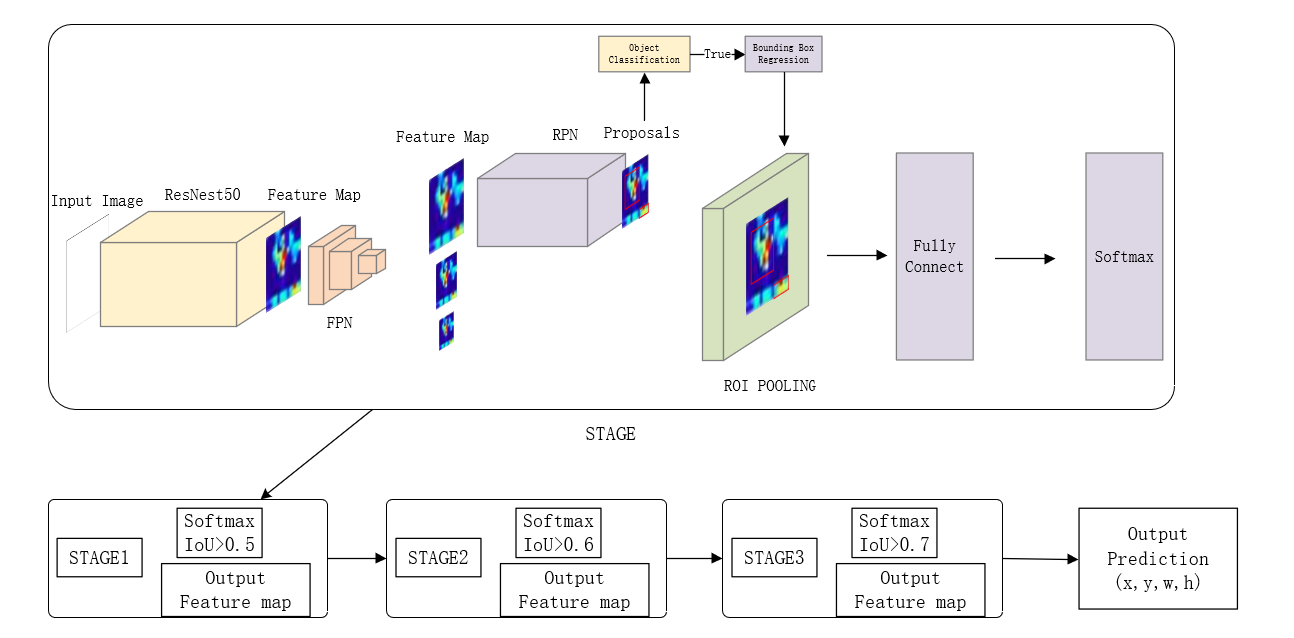
在本任务中为单目标检测任务，采用但目标检测常用的Triplet loss：

其中ρ是边距参数，是一个超参，d\_ij d\_ik :分别表示anchor sample 分别与positive sample、negative sample的欧式距离。

未标注示例u输入至标注网络，获得初代标签。



标注网络以ResNet50作为模型的骨干神经网络，设置三阈值Stage，逐步提升，不同阶段采用的IoU阈值不同，如果使用较低的IOU阈值，那么会学习到大量的背景框，产生大量的噪声预测。但是如果采用较高的阈值，这个检测器的表现往往会变得很差，两个主要的原因，第一就是随着IOU阈值的增加，正样本的数量会呈指数级的减小，因此产生过拟合。第二就是推理过程中出现于IOU的误匹配，也就是在训练优化感知器的过程中的最优IOU与输入proposal的IOU不相同，出现误匹配，这样很大程度上降低了检测精度。对于本任务，本项目经过实验，得到最优IoU阈值。



在初次使用时，加载第三方的预训练权重（如ImageNet），在训练垃圾目标检测网络后可以得到垃圾检测的权重，之后可以在下次使用标注网络标注时加载该权重，从而提高模型标注的准确率。

这一方法的三大优点是：

1、可以大规模标注，每个标注函数都可以用于成百上千个数据样本的标注。

2、可以利用海量的未标注数据，来构建大量虽然不完美但是足够好的大型训练数据集

3、这些标注可以用于训练一个具有大特征集的强大的判别分类器。即使只使用少量标注图像，每个数据样本依然可以有上千个特征。只经过测试，只需标注原数据集中30%的图片，就能让标注网络启动数据集预标注，能够一键标注剩余数据集。对比手工标注训练出的模型效果，其标注的准确率达到了95%以上。

（二）人工监督二次标注

通过弱监督学习的标注网络获得的初代标注数据并不能直接用于训练，但是已经完成了大部分的标注任务，大大减少了工作量。在这个基础上，项目组将对垃圾图像进行二次标注，确保图像是标注正确的，得到图像中异物标注框的四维向量信息(x, y, w, h)，其类别为garbage，将数据集按照 PASCAL VOC 格式进行4:1的比例划分为训练集和测试集。

## 1.3数据增广与增强

### 1.3.1 总论

垃圾分类作为典型的大规模多尺度图像分类问题，其数据有较高的复杂性。

首先，少数种类的垃圾数据量较小。数据量小就要求我们对少量已有的数据进行精细地调整，使图片的信噪比达到最优，另外我们可以使用不同的图像增广方法，获得不同特征的图像以扩充数据集。

其次，对于有效信息离散分布的图像，我们使用可以使用弱监督模型对垃圾区域进行目标检测，使用目标检测的结果生成新图像，以达到扩充数据集的目的。

最后，对于大部分数据，存在明显的多分辨率、多尺度、多粒度、特征复杂的特点。

于是，针对部分分类数据量小，大部分分类数据特征复杂的特点，兹设计如下的数据增广与增强解决方案。

### 1.3.2解决方案概述

解决方案分为两个子方案：数据增广框架、数据增强框架。数据增广用于扩充数据集、数据增强用于加强有效特征。并且，该解决方案有增强分类特征、优化不平衡数据集、推理时间短的特点。

需要指明的是，在传统上Data Augment可以被译作数据增广或数据增强，这是一个翻译问题。在本文中为了方便表达，使用数据增广一词描述时间成本较高的数据预处理操作，反之使用数据增强一词描述时间成本低的操作。

其中，数据增广的主要目的是扩充数据集的数据量，解决分类不平衡问题；数据增强的目的是增强数据特征，解决图像的光照差异大等问题，使用最简单、最根本的方法有效增强模型的鲁棒性。

**（一）数据增广框架**

扩充数据集的经典方法是对图像进行随机裁剪、翻转、色彩变换等操作，将生成的新图片作为新样本，以达到扩充数据集的目的。

扩充数据集的挑战性方案是基于对抗神经网络的样本生成，通过对抗神经网络训练的方式得到优秀的样本解码器，在解码器中输入与垃圾分类图像相似性高的特征，完成样本生成任务。

经典方法的特点是高效、易实现，GAN方案的特点使训练难度大，成本较高，但样本生成的效果较好。

在本方案中，我们经过综合考量，提出新的样本生成方法。

这种方法基于发展日渐成熟的目标检测模型和传统的图像处理技术，具体流程如下：

* 使用少量bounding box数据与mask数据训练目标检测模型，获得弱检测模型1
* 使用弱检测模型1对新获得大量图像进行目标检测，获得推理结果
* 使用推理得到的bounding box及mask目标结果，在目标结果区域进行适应性色彩修正、适应性噪声补充等操作，并添加到其它背景图片中

**（二）数据增强框架**

**（三）解决方案特点**

特点一：增强分类特征

使用AutoML算法，对限定条件下的数据增强或增广方法进行参数搜索，获得有效的数据增强手段，使得样本特征明显、鲁棒性高。

特点二：优化不平衡数据集

通过样本生成方法优化不平衡数据集。传统样本生成基于GAN的方法，生成器训练难度大、训练成本高。本数据增强解决方案根据垃圾分类问题特点，设计有针对性的解决方案。方案概述如下：

通过弱目标检测模型提取不平衡数据分类的mask或bounding box，再将提取结果与其它图像做拼接，拼接过程中对mask或bounding box做适应性色彩修正、适应性噪声补充、翻转、旋转、移位、缩放、裁剪等操作。通过这种方式，对不平衡分类数据进行扩充，扩充的数据特征自然，适用于模型训练。

特点三：推理时间短

本解决方案基于高效的数据增强第三方库Imgaug，推理时间短，落地方法多。使用于有着海量数据的垃圾分类问题。

### 1.3.3相关工作

谷歌ImageNet搜索得到的Auto-augment策略

第三方库Imgaug

## 1.4模型设计

### 1.4.1 Faster RCNN模型步骤

(1)首先在网络中输入自定义大小图片;

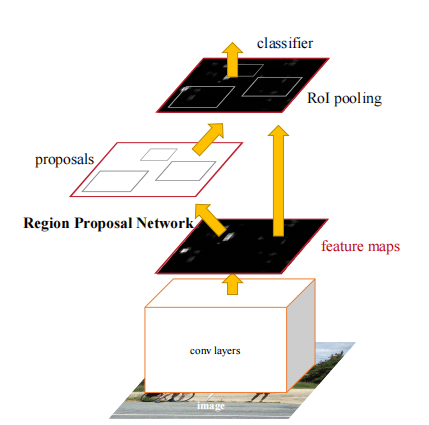
(2)在主干网络中时通过使用一组卷积、Relu层、池化层来对图像特征进行提取，从而得到图像特征信息，获取特征图(Feature map)。之后将得到的Feature map传输到RPN网络层，以便于生成候选区域和全连接层进行目标类别的分类。

(3)使用RPN网络主要用于生成区域推荐(ROIS)，在RPN网络层中通过使用Softmax函数来将Anchors进行二分类，判断是目标物还是不是目标物。接下来运用Bounding box regression对Anchors进行校正，这样可以得到更准确的Proposals。

(4)实现目标物的分类，把在第2个步骤得到的高维特征图和第3个步骤得到的输出区域--起运送到RoI池化层中，在神经网络模型的末尾运用全连接层，把目标物进行类别区分，实现目标物体的多分类。

(5)实现边框的回归利用Proposal feature maps计算每个Proposal所属地的不同类别概率信息，利用Boundingboxregression来得到物体检测框的准确位置，从而得到目标对象的检测结果。

Fastern RCNN模型的整体结构如图所示：

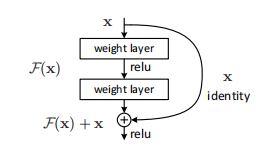


图Fastern RCNN的模型结构

1.2.3.2改进后的Faster RCNN模型介绍

(1)主干网络修改

首先针对于本次垃圾分类数据集的特点选用ResNet101网络作为主干网络，由于分类的类别数较多，我们需要提取较多的图像特征，因此也就需要较深的网络结构，然而当网络的深度过深时，传统的卷积网络或者全连接网络会出现信息丢失、梯度消失、梯度爆炸等问题，导致很深的网络无法训练。而 ResNet 即残差网络，能够很好地解决这一问题。其网络块结构如图 2所示



图ResNet的块结构

输入信息 X 经过两层权重映射后得到 ，也可以直接跳过权重层，残差块将

和 两者逐元素相加并用Relu 函数激活后输出得，则残差函数为：

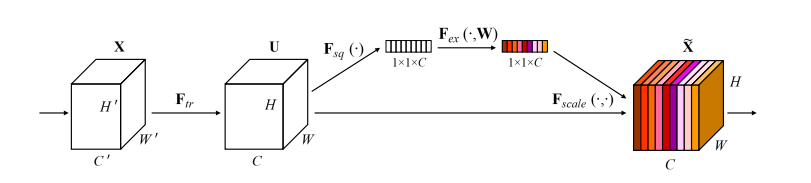


可以看到，输出中包含了输入原本的信息和输入经过权重层映射后的信息这样一来不仅能够防止信息丢失，更有效地利用信息，并且模型拟合残差函数要比直接拟合更简单，简化了学习目标，缩短了模型训练时间。

ResNet有不同的网络层数，比较常用的是 50-layer，101-layer。他们都是由图 2 的 ResNet块结构在一起实现的，101层的残差网络深度较深，能充分提取图像特征且相较于VGGNet其更网络结构更简单，因此我们以 ResNet101为网络结构搭建了图像分类神经网络模型。

(2)添加SE模块

为了合理利用有限的视觉信息处理资源，人类需要选择视觉区域中的特定部分，然后重点关注它。在神经网络中，注意力机制通常是一个额外的神经网络，能够硬性选择输入的某些部分，或者给输入的不同部分分配不同的权重。在神经网络中引入注意力机制有很多方法，以卷积神经网络为例，可以在空间维度增加引入Attention机制（如inception网络的多尺度，让并联的卷积层有不同的权重），也可以在通道维度（Channel）增加Attention机制，SE模块就是在通道维度（Channel-wise）上增加注意力机制。如下图所示：



图SE模块的结构图

首先进行压缩(Squeeze)，通过全局池化（global pooling），将每个通道的二维特征（H×W）压缩为1个实数，论文是通过平均值池化的方式实现。这属于空间维度的一种特征压缩，因为这个实数是根据二维特征所有值算出来的，所以在某种程度上具有全局的感受野，通道数保持不变，所以通过Squeeze操作后变为1×1×C，Squeeze的公式为：

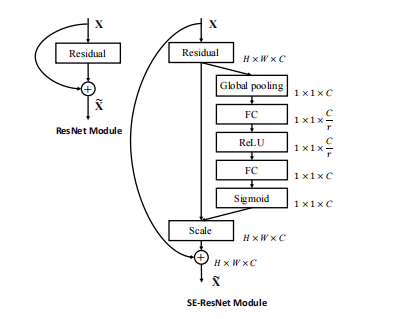


其次进行Excitation（Fex），通过参数来为每个特征通道生成一个权重值，这个权重值是如何生成就很关键了，论文是通过两个全连接层组成一个Bottleneck结构去建模通道间的相关性，并输出和输入特征同样数目的权重值。

最后进行Scale（Fscale），将前面得到的归一化权重加权到每个通道的特征上。论文中的方法是用乘法，逐通道乘以权重系数，完成再通道维度上引入attention机制，Scale的公式为：



在本网络结构中，选择把SE模块嵌入到ResNet当中，如下图所示，Global pooling就是Squeeze操作，FC + ReLu + FC + Sigmoid就是Excitation操作，具体过程是首先通过一个全连接层（FC）将特征维度降低到原来的1/r，然后经过ReLu函数激活后再通过一个全连接层（FC）生回到原来的特征维度C，然后通过Sigmoid函数转化为一个0~1的归一化权重。



图SE模块嵌入到ResNet结构图

(3)修改负样本挖掘策略(Hard Negative Mining Method)

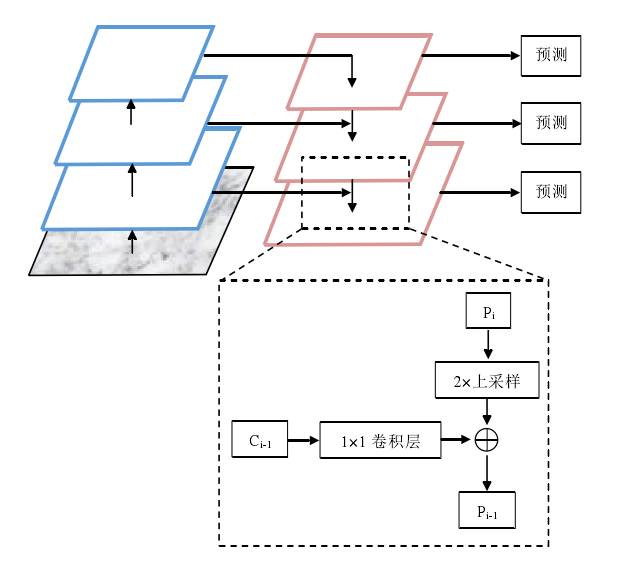
在 two-stage 的目标检测方法中，经过区域生成算法或者网络生成的 Region proposals 通常会经过正负样本的筛选和比例平衡后，才送入之后的检测网络进行训练。但是正负样本的定义和训练的比例是需要人为定义的。如果要筛选hard mining。通常使用的都是 hard negative mining的方法。但是此方法不适用于end-to-end的模型。因为会大大降低模型的训练速度，所以本模型使用OHEM方法来自动地选择 had negative 来进行训练，不仅效率高而且性能好。

对于一次 SGD 迭代，计算过程如下:先计算出特征图，可读 RoI 网络对所有 RoI 执行前向计算并计算每个 RoI 的损失，然后选择 hard RoIs。把这些 hard RoIs 输入到可读可写的 RoI 网络中执行前向前向计算和反向传播更新网络，并把可读可写的 RoI 网络的参数赋值给只可读的网络，一次迭代就完成了，速度和效率提升明显。

(4)添加特征金字塔网络(FPN)

在以往的Fastern RCNN进行目标检测时，无论是RPN还是Fastern RCNN，ROI都作用在最后一层，对于大目标的检测没有问题，但是对于小目标的检测就有些问题。因为对于小目标来说，当进行卷积池化到最后一层，实际上语义信息已经没有了，对于一个ROI映射到某个Feature map的方法就是将底层坐标直接除以Stride,显然随着网络的加深，映射过去后就越小，甚至可能就没有。所以为了解决多尺度检测的问题，引入了特征金字塔网络。

FPN的大致思路可以概成的深层特征，同时也有效地利用了网络中的浅层特征。FPN的大致思路可以概括为自底向上、自顶向下以及横向连接，示意图如所示。以ResNet模型为例，先根据模型的前向过程，自底向上地生成一系列尺寸不同的特征图{C2、C3、C4、C5}，这些特征图分别为ResNet 模型中conv2、 conv3\_ x、conv4\_ \_x以及conv5\_ x等残差模块的输出，相对于输入图像的步长分别为{4、8、16、 32},C1层由于尺寸过大且局部感受野过小，不纳入计算过程中。随后，在C5.上通过1x1的卷积操作进行维度变换得到P5，然后在特征图P5上通过最近邻上采样，将长宽各扩大1倍，向下获得尺寸与C4相同的特征图，将该特征图横向与维度变换后的C4相连，其中各元素相加融合得到P4，随后依次向下得到P3、P2。最后在P2~P5后分别加入3x3的卷积层，消除上采样过程中的混叠效应，输出特征均调整为256个通道。



图特征金字塔结构

(5)损失函数的修改

二分类模型的期望效果是模型总是给正样本输出1，负样本输出0，但模型的拟合程度有限，并不能完美做到这一点。在实际预测中，一般认为大于0.5的属于正样本，而小于0.5的属于负样本。

二分类模型表现不佳的原因通常在于正负样本比例不平衡，在检测过程中，锚框（Anchor Box）近似于滑动窗口（Sliding Window）的方式会使正负样本比例接近于1000：1，而且大多数负样本都是易于分类的样本，即easy example，这些easy example虽然loss很低，但是数量众多，占据了loss中的大部分，还主导了梯度，因此我们可以通过对损失函数的修改来平衡正负样本的重要性并易于区分hard example和easy example，降低简单样本的权重，让损失函数能够更关注于hard negative example，二分类问题的标准loss交叉熵损失的公式为：

对于二分类模型用Sigmoid函数激活，所以公式为

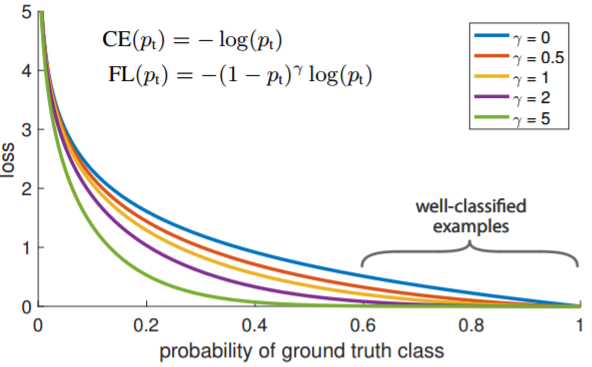


通过设定两个阈值，例如0.4和0.6，当模型对某个负样本的输出小于0.4，或者模型对某个正样本的输出大于0.6时，不需要再根据这个样本来更新模型，而只有在0.4~0.6之间的样本才会更新模型，这样就达到了让模型更关注于hard negative example的目的，从而使的训练后的分类效果更好

因此，交叉熵损失修改为[Kaiming He](https://link.zhihu.com/?target=https%3A//scholar.google.com/citations%3Fuser%3DDhtAFkwAAAAJ%26hl%3Den)提出的Focal loss函数

γ的作用是调节权重曲线的陡度，例如在训练中负样本远比正样本多的时候，模型会倾向于数目多的负类，这时候负类的 就会比较小，而正类的 会较大，这时模型就会关注更多的正样本，如图所示。





图Focal loss改变参数伽马的损失曲线

在上图中，X轴表示预测真实值的概率，Y轴是给定预测值下的损失值。当γ=0时，Focal loss为交叉熵损失。从图像中可以看出，当模型以0.6的概率预测真实值时，交叉熵损失在0.5左右，因此为了减少损失，模型会以更高的概率预测真实值，这样会降低模型的鲁棒性，使得模型可能出现过拟合的情况。但是在类别不平衡的情况下，Focal loss会将模型的注意力转向稀有类别，这在一定程度上减小了过拟合情况出现的概率。

(6)标签平滑正则化

在深度学习样本训练的过程中，当采用one-hot标签去进行计算交叉熵损失时，可能会出现只考虑到训练样本中正确的标签位置（one-hot标签为1的位置）的损失，而忽略了错误标签位置（one-hot标签为0的位置）的损失。因此，模型虽然可以在训练集上拟合的很好，但是由于其他错误标签位置的损失没有计算，导致预测的时候，预测错误的概率增大。为了解决这一问题，标签平滑的正则化方法便应运而生。

为了达到这个目标，改进的方法是：在每次迭代时，并不直接将(xi,yi)放入训练集，而是设置一个错误率ε，以1-ε的概率将(xi,yi)代入训练，以ε的概率将(xi,1-yi)代入训练。 这样模型在训练时，既有正确标签输入，又有错误标签输入，训练出来的模型不会“全力匹配”每一个标签，而只是在一定程度上匹配。这样，如果出现错误标签，模型受到的影响会更小。

当没有标签平滑计算的损失只考虑正确标签位置的损失，而不考虑其他标签位置的损失，这就会出现一个问题，即不考虑其他错误标签位置的损失，这会使得模型过于关注增大预测正确标签的概率，而不关注减少预测错误标签的概率，最后导致的结果是模型在训练集上拟合效果非常良好，而在其他的测试集结果表现不好，模型泛化能力较差。

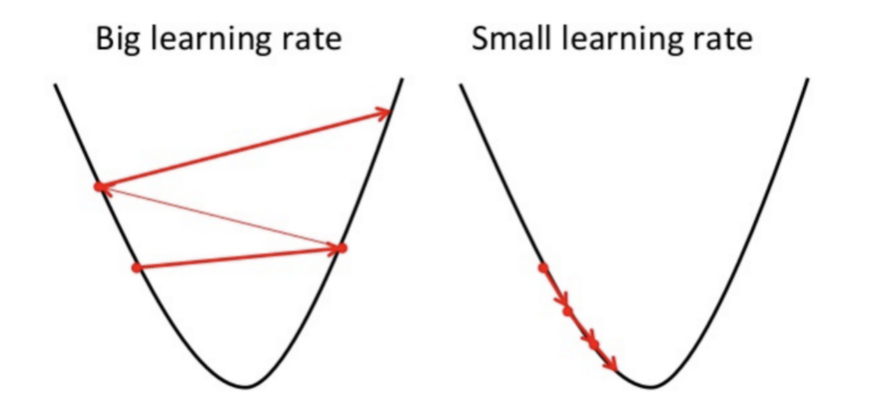
在进行平滑后的样本交叉熵损失不仅包含训练样本中正确的标签位置（one-hot标签为1的位置）的损失，也包含其他错误标签位置（one-hot标签为0的位置）的损失，使模型的学习能力提高，让模型往增大正确分类概率并且同时减小错误分类概率的方向前进。

标签平滑的实质是促使神经网络中进行Softmax激活函数激活之后的分类概率结果向正确分类靠近，即正确的分类概率输出较大（对应的one-hot标签为1位置的Softmax概率较大），并且同样尽可能的远离错误分类（对应的one-hot标签为0位置的Softmax概率较小），即错误的分类概率输出较小。

(7)优化方法修改

传统的深度学习优化方法是学习率调度器和自适应优化器，通常来说优化器都有一个学习率超参数，这是影响模型性能的最重要的超参数之一，在简单背景下，学习率是固定的，后来经实践发现，采用“模拟退火”的策略有助于模型更快的收敛，从而达到全局最优状态。

如果设置单一学习率，学习率的取值是难以确定的，在选择较大的学习率时，模型在早期会很快的逼近全局最优状态，但是在全局最优状态附近，大学习率可能会导致模型超过最优点，在最优点附近无法收敛；而较小的学习率会增加模型的迭代次数，延长训练时间，有可能会使模型陷入局部最优状态，从而不能收敛。因此应该保持学习率不断减小使其到达最优点，但是这种方法无法避免模型陷入局部最优状态，如图所示。



图模型陷入局部最优解示意图

而“模拟退火”的策略是指在模型训练的早期阶段，模型选择较大学习率，使得模型更快的逼近全局最优状态；而在训练后期，换成小学习率以逐渐收敛至全局最优状态；但是并不始终保持学习率一直减小，而是周期性的提高学习率，以便在模型陷入局部最优解的时候，学习率会向较高的学习率振荡以跳出鞍点。传统学习率调度器正是参考了模拟退火策略，选取base\_lr和max\_lr，让学习率lr在base\_lr和max\_lr之间周期性增加和减小，一次迭代（cycle）包含两个步长（step\_size），经过每次步长学习率lr会通过衰减函数scale\_fn更新一次，以此有效的加快收敛并提高准确率。

在传统学习率调度器的基础上，可以采用OneCycleLR（单周期学习率调度器）优化，设置整个训练过程只有一个cycle，学习率首先从初始值上升至max\_lr，之后从max\_lr下降至低于初始值的大小，一般设置为max\_lr的1/5或1/10，单周期的长度略小于要训练的周期总数，在最后的迭代中，将学习率设置为远低于max\_lr的值（如max\_lr的1/10）。以使得在学习率较大时，学习率可以作为正则化方法发挥作用，防止模型过拟合。

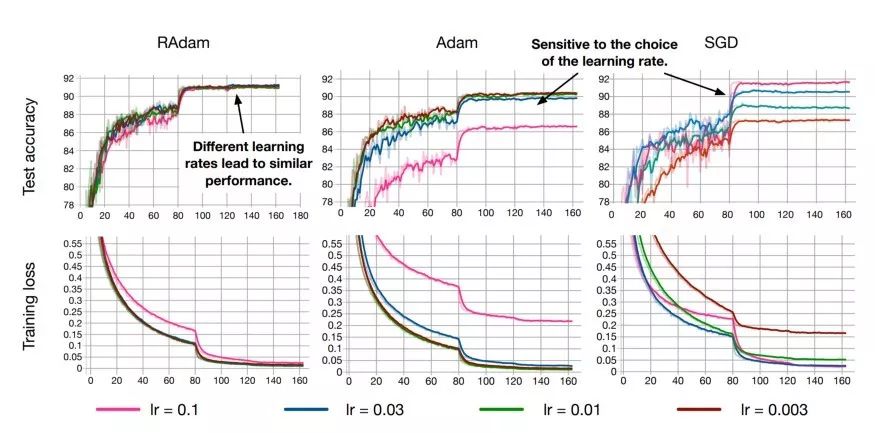
![MR$DCA]T`V1E_NKNUPUX](R](data:image/png;base64,)

图Onecyclelr的学习率变化曲线

自适应优化器并不使用单独的学习率调度器，而是将学习率优化直接嵌入到优化器本身，传统的自适应优化器如应用最广泛的Adam通过权重来管理学习率。Adam使用动量和自适应学习率来加快收敛速度，它的超参数非常少，改变它并不会对性能产生太大影响。

但是Adam具有鲁棒性不好的缺点，它常常会收敛到不太好的局部最优解，只能通过预热（warmup）策略来解决，即最初几次迭代，都用很小的学习率，以此来缓解收敛问题。这一问题出现的根源是模型早期缺乏足够的数据样本，因此学习率方差较大，自适应率的方差也较大。

因此改进模型采用优化算法RAdam来代替Adam优化模型。Radam和Adam相比降低了自适应学习率的方差，输入步长，衰减率，通过迭代计算梯度和动量的二阶矩，移动偏差的修正和方差修正范围，使用非自适应动量更新参数，RAdam的和Adam，SGD对比如图所示。



图RAdam的和Adam，SGD对比效果

Radam算法对于初始学习率具有鲁棒性，在一个很宽的范围内表现出了一致的性能，训练曲线末端高度重合。RAdam提供了一个动态启发式方法来提供自动化的方差衰减，从而消除了在训练期间预热所涉及手动调优的需要。此外，RAdam对学习速率变化具有更强的鲁棒性，并在各种数据集和各种AI体系结构中提供更好的训练精度和更强的模型泛化能力。