VOC数据集

VOC2007中包含9963张标注过的图片， 由train/val/test三部分组成， 共标注出24,640个物体。对于检测任务，VOC2012的trainval/test包含08-11年的所有对应图片。 trainval有11540张图片共27450个物体。对于分割任务， VOC2012的trainval包含07-11年的所有对应图片， test只包含08-11。trainval有 2913张图片共6929个物体。

PASCAL VOC2007数据集总共分 4 个大类：vehicle、household、animal、person，总共 20 个小类（加背景 21 类），预测的时候是只输出下图中黑色粗体的类别：

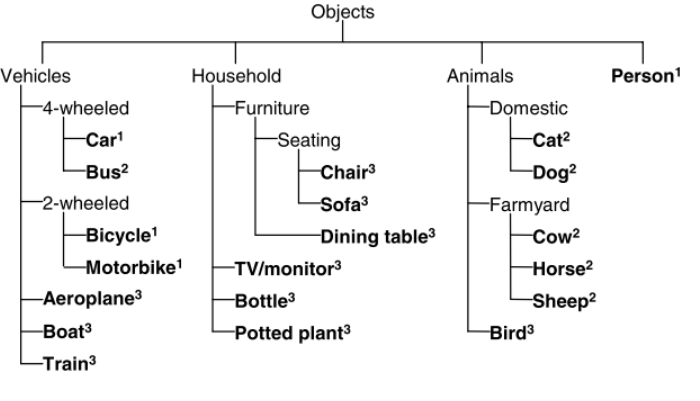


图9

2 模型结构

2.1 Faster-RCNN

2.1.1 残差网络

残差网络使用函数表示为 。残差网络结构可以避免当网络层数较深时引起的梯度消失现象。残差结构包含标准残差block以及bottle block，其中Resnet50采用bottle block作为基本block。

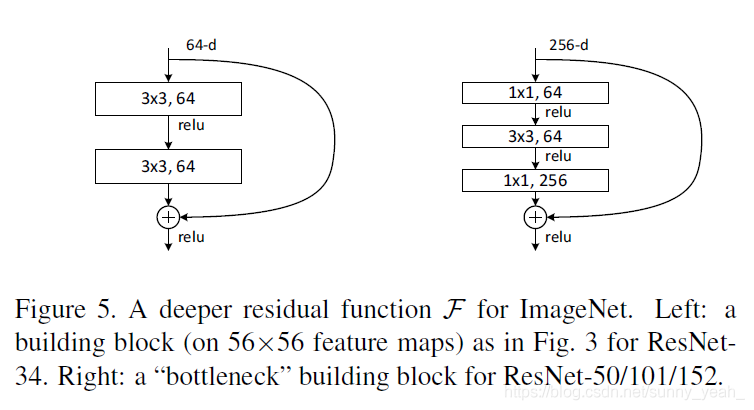


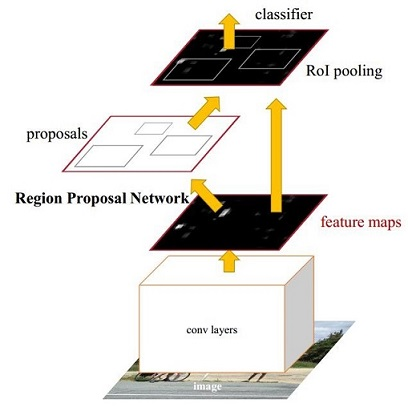
图 10 block(左) 和 bottle block(右)

2.1.2 Backbone-Resnet50

模型采用Resnet-50作为baseline，由于训练样本图片大小为32×32，因此将Resnet-50第一层kernel\_size = 7, stride = 2的卷积层删除，并移除之后的3×3 max pool层。使用kernel\_size = 3, stride = 1的卷积层代替。

经过卷积神经网络获得的特征图展平之后输入到全连接层预测最终的标签。

2.1.3 整体网络结构可视化



3 训练设置

3.1 Faster-RCNN

设备：GTX1080Ti ×1

Batchsize：4

Learning-rate：0.0002

学习率衰减：(12000, 16000)iter \* 0.1

优化器：SGD

Iter：18000

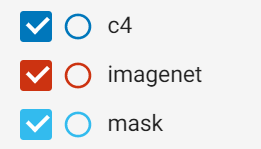
NMS Thresh:0.5

NMS TOP 6000

4 训练任务

对Faster R-CNN模型，分别进行以下训练：a) 随机初始化训练VOC；b) ImageNet预训练backbone网络，然后使用VOC进行fine tune；c)使用coco训练的Mask R-CNN的backbone网络参数，初始化Faster R-CNN的backbone网络，然后使用VOC进行fine tune；

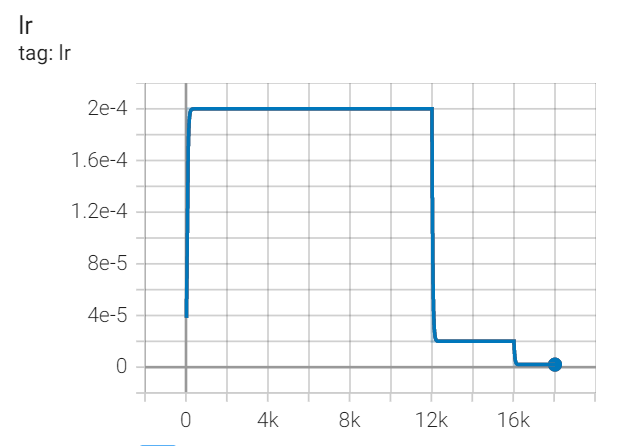
4.1图例



4.2 Training loss



4.2 Learning rate



4.3 检测结果

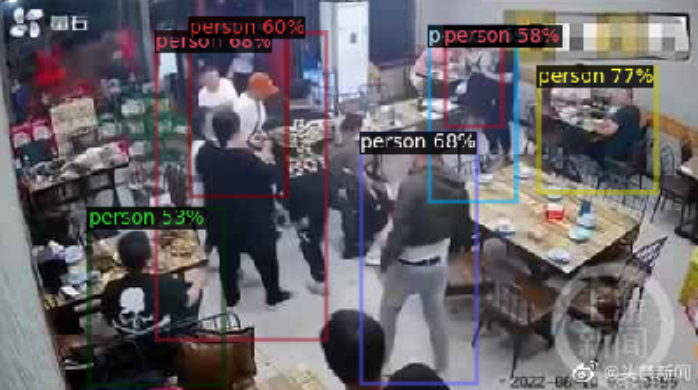
|  |  |  |  |
| --- | --- | --- | --- |
|  | AP | AP-50 | AP-75 |
| Base | 0.1794 | 0.4373 | 0.0758 |
| imageNet | 23.9888 | 53.5282 | 17.0300 |
| Mask-weight | 57.4507 | 82.2074 | 64.0020 |

4.4 可视化

随机初始化（0个物体）



Imagenet预训练（7个物体）



使用coco上mask-rcnn的权重finetune（10个物体）



结论：使用imagenet 预训练能显著提高收敛效果，提高检测效率，使用mask-rcnn上权重进行fine-tune使得效果会更好。