一、 基于 DeeplabV3 Plus 的行车视频检测

蔡畅威（21210980025）、汪加西（21210980069）

- 问题介绍

使用以 Mobile-Net 为主干网络的 DeeplabV3+，对一段网络下载的行车视频作语义分割。检测后视频由约六百张测试图片合成，时长 1 分 40 秒。视频及项目链接请见文末。

图 1：部分图片的测试结果

二、 使用三种 Faster-RCNN 的目标检测

- 问题介绍

在 VOC 上训练主干网络为 resnet-50 的 Faster-RCNN。考虑以下三种方式

1. 随机初始化 resnet-50，即在调用 torchvision.models 时将 pretrained_backbone 设置为 False；
2. 使用在 ImageNet 上训练好的 resnet-50 作为主干，将 pretrained_backbone 设为 True；
3. 在 COCO 上预训练 Mask-RCNN，使用 Mask-RCNN 的 resnet-50 初始化 Faster-RCNN 的主干。须首先从 torchvision 下载 Mask-RCNN-resnet50 模型，将其 backbone 的权重赋值给 Faster-RCNN，即 fasterr.backbone = nn.Sequential(maskr.backbone)。

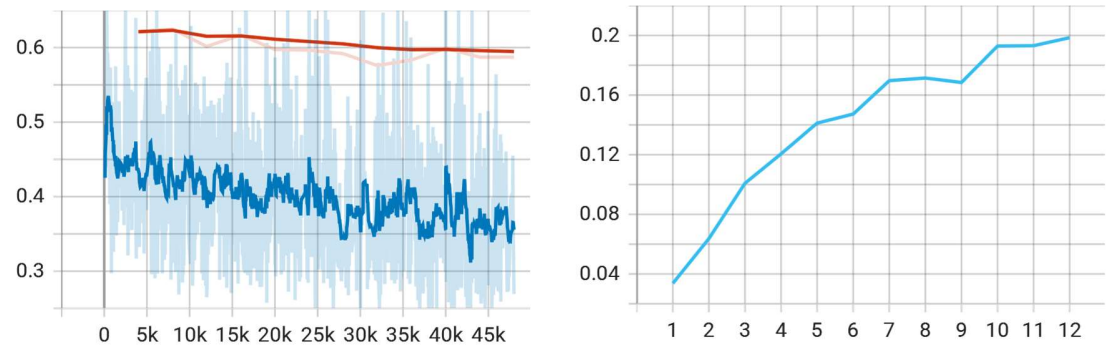

图 2(a)：损失（左）和 mAP（右）—随机初始化

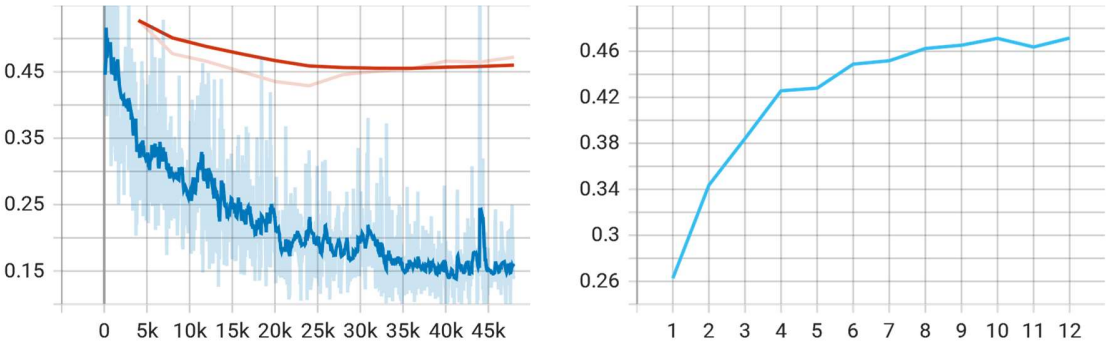

图 2(b)：Imagenet 预训练

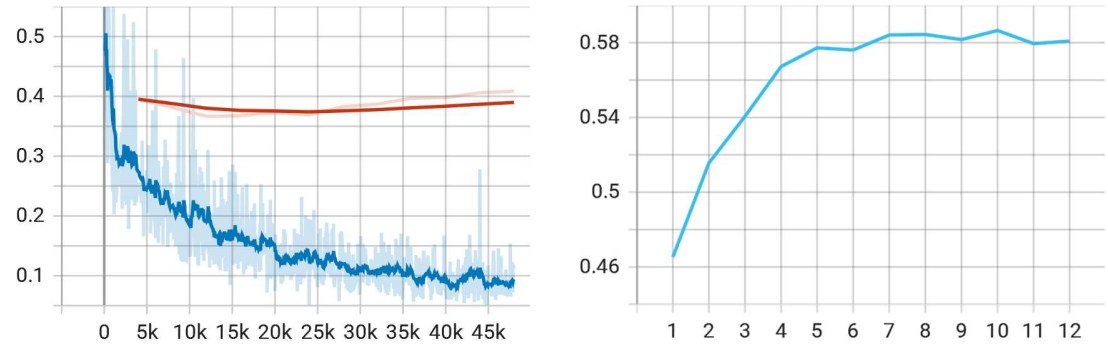

图 2(c)：COCO 预训练

将数据集分为训练、测试，样本量分别为 4000 和 1000，使用 Adam 优化器，初始学习率均设为 1×10^{-5} 并在每个 epoch 后将学习率下调 20%。训练 12 个 epoch，每 20 个 iteration 在 train 上输出总损失，每个 epoch 后在 test 上输出总损失和 mAP。

如图 2 为三种训练方式下损失和 mAP 迭代，其中 mAP 是在 IOU=0.5 时、对单张图片取平均计算的。可见在学习率等参数相同的情况下，COCO 预训练模型的效果最好：其在第一个 epoch 后 mAP 达到 0.465，高于 Imagenet 预训练 (0.263) 和随机初始化 (0.034)。在 12 个 epoch

内，三个模型 mAP 的最高值分别为 0.199、0.472 和 0.587，具体见表 1 (其中 *epoch* 表示达到最高 mAP 和最低损失所需的训练轮数)。为进一步比较三个模型，另取 3000 张测试图片，计算模型在各个检测类别上的 mAP，如表 2 (a)-(c)。

表 1：三种训练方式下 Faster-RCNN 的 mAP 和损失

	$\max mAP$	*epoch*	$\min loss$	*epoch*
随机初始化	0.199	12	0.576	8
Imagenet 预训练	0.472	12	0.429	6
COCO 预训练	0.587	10	0.367	3

表 2(a)：随机初始化 Faster-RCNN 检测精度 (mAp=0.060)

plane	0.143	bus	0.072	table	0.050	plant	0.013
bike	0.083	car	0.129	dog	0.032	sheep	0.040
bird	0.017	cat	0.035	horse	0.117	sofa	0.021
boat	0.016	chair	0.013	motor	0.125	train	0.071
bottle	0.006	cow	0.048	person	0.073	Tv	0.094

表 3(b)：Imagenet 预训练 (mAp=0.439)

plane	0.477	bus	0.520	table	0.290	plant	0.376
bike	0.485	car	0.546	dog	0.509	sheep	0.431
bird	0.407	cat	0.567	horse	0.500	sofa	0.390
boat	0.278	chair	0.296	motor	0.482	train	0.515
bottle	0.354	cow	0.450	person	0.450	tv	0.455

表 4(c)：COCO 预训练 (mAp=0.587)

plane	0.652	bus	0.678	table	0.429	plant	0.446
bike	0.559	car	0.648	dog	0.667	sheep	0.520
bird	0.602	cat	0.652	horse	0.653	sofa	0.558
boat	0.457	chair	0.518	motor	0.641	train	0.651
bottle	0.530	cow	0.627	person	0.614	tv	0.645

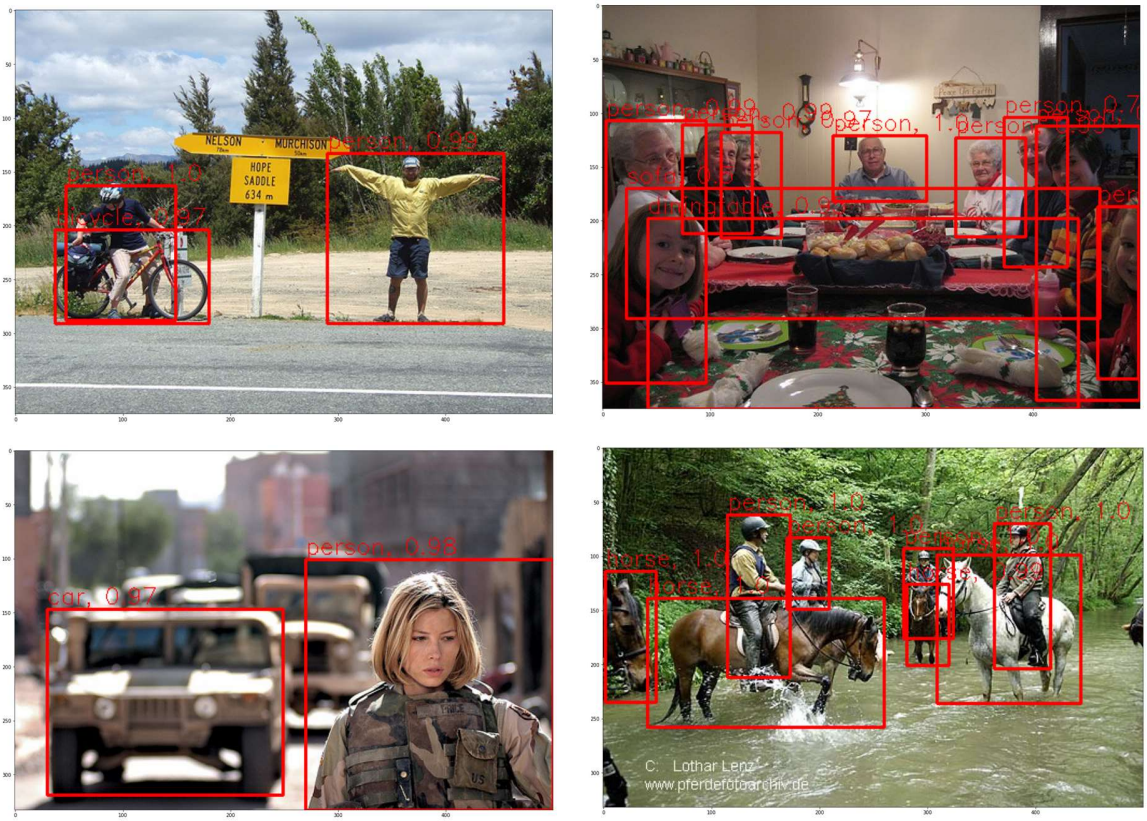

图 3(a)：使用 COCO 预训练 Faster-RCNN 的目标检测

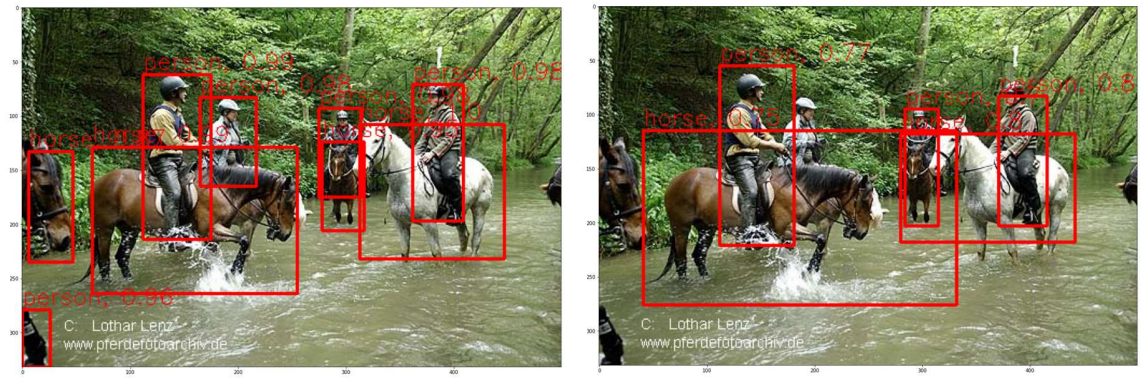

图 3(b)：Imagenet 预训练（左）、随机初始化（右）

使用三种模型对挑选出的图片作目标检测，如图 3 (a)、(b)。可见 COCO 预训练 Faster-RCNN 的检测效果较好，其对复杂图片 (如右下子图包含多个人、马且框有重叠) 也能准确识别。另外两个模型的效果要稍差些，如图 3 (b) 左下角的物体被错误地识别为人，右子图未能检出图片左侧的马头且框的置信度比较低。

三、 使用 VIT 的 CIFAR-100 图像分类

- 问题介绍

所使用的 Transformer 网络模型为 VIT，来自论文 An Image is Worth 16×16 words: Transformers for Image Recognition at Scale (Alexey, 2020) 中的 Vision Transformer。该模型首先在大型数据集上预训练 VIT，再在规模较小的下游任务微调 (添加一个初始化全零，维度为 $D \times K$ 的前馈预测层，其中 K 为下游类的个数)。虽然 VIT 可处理任意长度的输入序列，但预训练得到的位置嵌入可能是混淆的。作者为此对预训练位置嵌入在原始图像中的坐标作二维插值，改善了位置嵌入的表示。我们使用官方 Github 提供的预训练模型，再在 CIFAR-100 任务上作微调。该模型版本为 VIT-B/16，与 resnet-101 参数量相近 (均为 8.5×10^8 左右)。

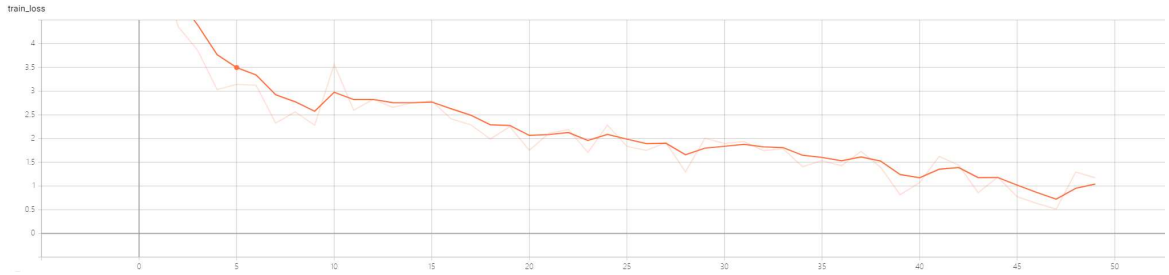

图 4(a)：训练集上损失

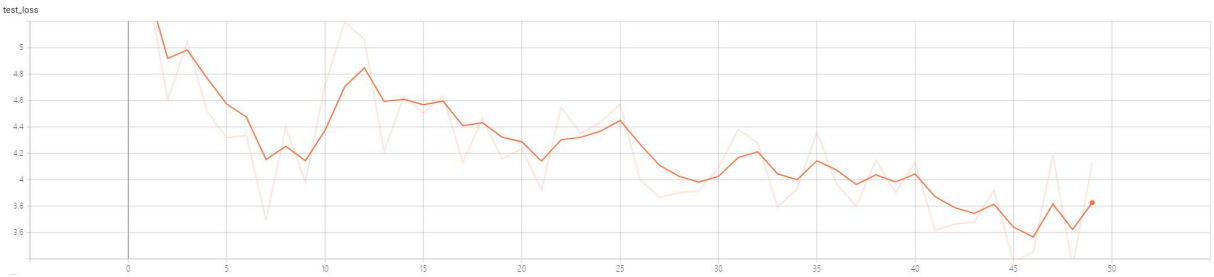

图 4(b)：测试集上损失

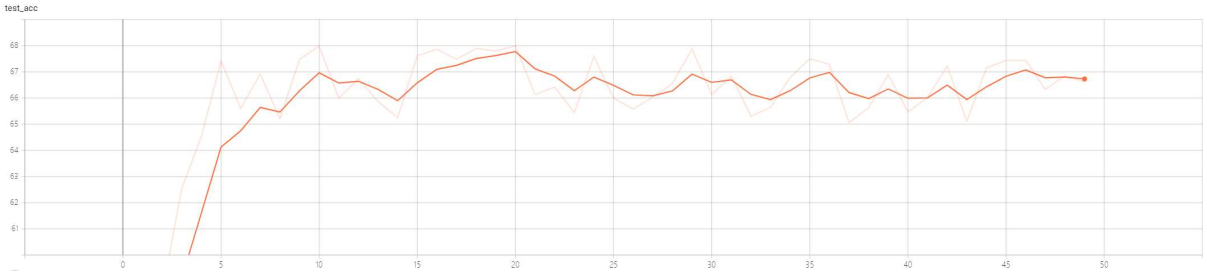

图 4(c)：测试集上准确率

- 模型训练

设置 epoch 为 50，学习率 0.001，batch 大小为 8，使用 SGD 优化器。数据增强方法为：对每张训练图片，以各 1/3 的概率作 cutmix, cutout 和 mixup。图 4 (a)-(c) 为损失及准确率迭代情况。可见损失收敛速度较快，且由于预训练，模型准确率在前 3 个 epoch 内就已达到 0.6 以上。模型最终在测试集上准确率为 0.6662，高于 resnet：期中作业使用 resnet-18 的准确率仅为 0.5866，resnet-101 为 0.62 左右 (参考网络上另一项目)。

- 期末项目 (包含 1-3) 请见 Github repo：

https://github.com/pitter-patterz/final_pj

- 视频 / 模型下载链接为：

https://pan.baidu.com/s/1ry7nWFbDlZJsGRP0xhusiw?pwd=sjwl (pwd:sjwl)

https://pan.baidu.com/s/1-OF-MkzxYYBQTXHPpctmdw?pwd=sjwl (pwd:sjwl)

https://pan.baidu.com/s/1Jl7Afbr-GDblsmgrd5dAMw (pwd:rjjb)