机器学习期中报告

研究方向：行人搜索

学院：电子与信息工程学院

专业：计算机科学与技术

学号：2230771

姓名：包广垠

完成日期：2022 年 11 月 19 日

目 录

装 订 线

1 背景

行人搜索旨在从一系列未经裁剪的图像中对行人进行定位与识别，融合了行人检测和行人重识别两个子任务。该任务最早于 2013 年在 ACM 多媒体大会上提出，首次将行人检测与行人重识别任务整合为一个任务。

行人检测属于目标检测的子任务，旨在从大量的照片或者视频数据中找到行人的位置和大小，通常来说需要使用矩形的框将其框出来，而这一任务不需要识别框选出来的行人是谁；行人重识别则是行人匹配的任务，通常是从一系列已经裁剪好的行人图像中匹配一个和待匹配目标最像的人。将行人检测与行人重识别结合的行人搜索任务更具有实用价值，在可以应用在刑侦中，从而节省大量的人力资源开销。

由于行人搜索的任务结合了两个独立性较强的子任务，因此其解决方案通常有两种：两阶段模型和端到端模型。两阶段模型是分步骤解决行人搜索任务：首先使用深度学习方法将图片中的行人检测出来，对于检测出来的模型，使用行人重识别的模型进一步完成匹配任务。两阶段的模型需要分开进行两个子任务，使得原始任务的难度降低，但也会使得行人搜索的任务效率下降，因此端到端的行人搜索算法的研究受到了更多的重视和研究。端到端的行人搜索网络将行人检测和行人重识别集中到一个网络中处理，使得网络的训练和推理可以针对一个网络来完成，这样的模型会更具有实用性。

2 研究现状

以行人搜索领域的常用数据集（CUHK-SYSU、PRW）为例，在行人搜索上表现较好的模型和方法如下（数据截至到 2022 年 11 月 7 日）：

1）CUHK-SYSU：

Rank	Model	MAP↑	Top-1	Paper	Code	Result	Year	Tags
1	SeqNeXt+GFN	96.4	97.0	Gallery Filter Network for Person Search			2022	
2	SeqNeXt	96.1	96.5	Gallery Filter Network for Person Search			2022	
3	GLCNet+CBGM	95.8	96.2	Global-Local Context Network for Person Search			2021	
4	GLCNet	95.5	96.1	Global-Local Context Network for Person Search			2021	
5	ROI-AlignPS	95.4	96.0	Efficient Person Search: An Anchor-Free Approach			2021	
6	NAE+SeqNet+CBGM	94.8	95.7	Sequential End-to-end Network for Efficient Person Search			2021	
7	OIM+SeqNet+CBGM	94.3	95.0	Sequential End-to-end Network for Efficient Person Search			2021	
8	AlignPS+	94	94.5	Anchor-Free Person Search			2021	
9	NAE+SeqNet	93.8	94.6	Sequential End-to-end Network for Efficient Person Search			2021	
10	OIM+SeqNet	93.4	94.1	Sequential End-to-end Network for Efficient Person Search			2021	
11	AlignPS	93.1	93.4	Anchor-Free Person Search			2021	

2）PRW

Rank	Model	mAP↑	Top-1	Paper	Code	Result	Year
1	SeqNeXt+GFN	58.3	92.4	Gallery Filter Network for Person Search			2022
2	SeqNeXt	57.6	89.5	Gallery Filter Network for Person Search			2022
3	ROI-AlignPS	51.6	84.4	Efficient Person Search: An Anchor-Free Approach			2021
4	GLCNet+CBGM	47.8	87.8	Global-Local Context Network for Person Search			2021
5	NAE+SeqNet+CBGM	47.6	87.6	Sequential End-to-end Network for Efficient Person Search			2021
6	GLCNet	46.7	84.9	Global-Local Context Network for Person Search			2021
7	NAE+SeqNet	46.7	83.4	Sequential End-to-end Network for Efficient Person Search			2021
8	OIM+SeqNet+CBGM	46.6	84.9	Sequential End-to-end Network for Efficient Person Search			2021
9	AlignPS+	46.1	82.1	Anchor-Free Person Search			2021
10	AlignPS	45.9	81.9	Anchor-Free Person Search			2021
11	OIM+SeqNet	45.8	81.7	Sequential End-to-end Network for Efficient Person Search			2021

从两个数据集的 benchmark 中可以看到：除了基于 Anchor-free 的模型/方法均是基于 SeqNet 的（包括在排行榜前几位的 SeqNeXt 和 GLCNet）。SeqNet 这一网络结构来源于 2021AAAI 的论文《Sequential End-to-end Network for Efficient Person Search》,该论文的第一作者单位是我们同济大学电子与信息工程学院计算机科学与技术系，由苗夺谦教授团队发表。

基于以上两个理由，我打算复现的即为 SeqNet 这一工作。

装 订 线

3 算法原理

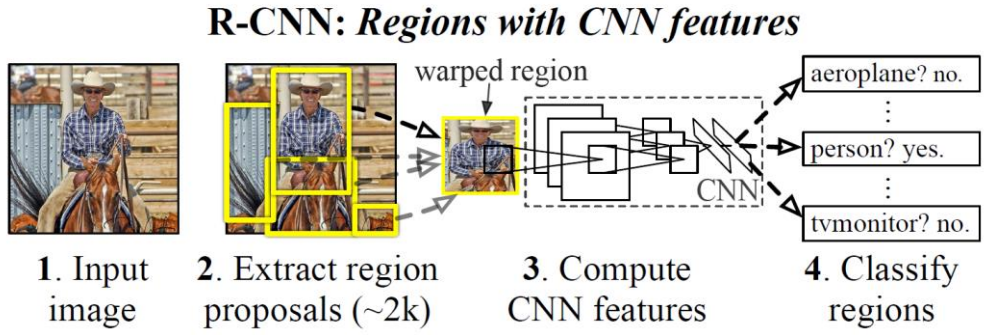
3.1 RCNN

RCNN 是目标检测领域的经典方法，出自于 2014 年的论文《Rich feature hierarchies for accurate object detection and semantic segmentation》。

RCNN 将目标检测分为分类与回归两个并行进行的子任务，分类任务是指将 bounding box 中内容进行分类，回归任务是指通过回归的思想寻找 bounding box 的准确位置。其推理过程包含以下四个步骤：

1. 使用 Selective Search 算法生成区域提案（Proposals）；
2. 将面积大小不相同的 proposals 缩放到统一的大小；
3. 将得到的结果输入特征提取网络得到特征向量；
4. 使用分类器得到区域的类别和 bboxes 的回归值。

算法的示意图如下：



在实际使用中，对于一张图片，RCNN 的方法要独立的将提取的两千多个区域独立地输入特征提取网络，并且需要训练复杂的 SVM 分类器，因此该方法的效率很低。

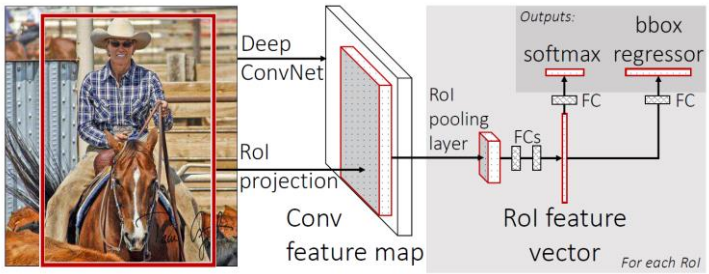
3.2 Fast RCNN

Fast RCNN 是对 RCNN 的改进，出自于 2015 年的论文《Fast R-CNN》。

Fast RCNN 提出了 ROI pooling 的方法，使得不需要单独地将 Selective Search 得到的 Proposals 输入特征提取网络，二是可以并行处理这些特征。该方法的推理过程包含以下几个步骤：

1. 使用 Selective Search 算法生成区域提案（Proposals）；
2. 将原始图像输入特征提取网络得到整个图像的特征图（feature map），将 proposals 与特征图上的区域对应起来，并将其 pooling 到指定大小；
3. 使用两层全连接网络将得到的结果提取为 ROI 特征向量；
4. 使用 softmax 和 bbox regressor 得到结果。

算法的示意图如下：

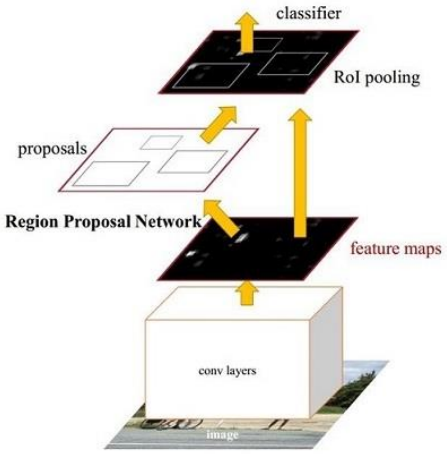


该方法由于使用了 ROI pooling 的方法和使用简单分类器，使得目标检测的效率大幅度提高，但仍然存在需要使用 Selective Search 算法生成大量冗余且不精确的区域提案（Proposals）的不足之处。

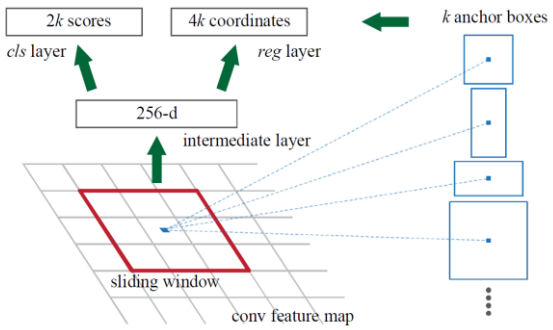
3.3 Faster RCNN

Faster RCNN 由是对 Fast-RCNN 的进一步改进，出自于 2016 年的论文《Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks》。

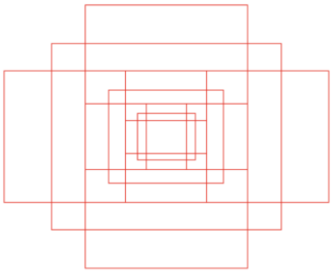
该方法的突出贡献在于提出了 RPN（region proposal network）网络，使得不再需要使用 Selective Search 算法生成区域提案。Faster RCNN 的网络结构如下，是由 Fast RCNN 和 RPN 组成：



RPN 网络的结构图如下：



其推理过程如下：首先在 feature map 上，使用 3×3 的滑动窗口提取 256 维特征向量，使用该特征向量得到一组 anchor boxes 的分类和回归结果。一组 anchor boxes 包含 9 个 anchor，对应原图上{1:1 1:2 2:1}×{128^2 256^2 512^2}的九个大小不同的区域，即中心点对应的如下区域：



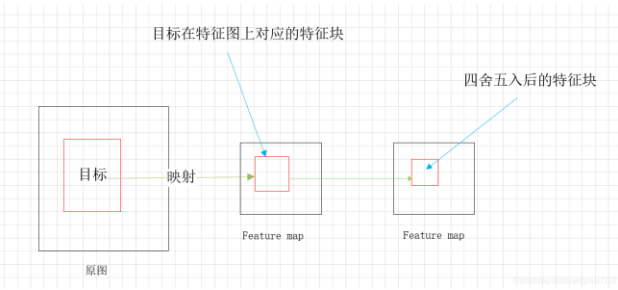
然后排除超过原始图像边界的区域，将这些区域计算其前景/背景的类别和回归参数，从而得到区域提案 proposals。然后使用非极大抑制算法（NMS）排除高度相似的 proposals，将最终剩下的 proposals 输入原来的 Fast RCNN 中。

Faster RCNN 的 RPN 结构使得模型可以自行生成较为精确的区域提案，同时 RPN 结构可以和 Fast RCNN 联合训练，这使得目标检测真正具有一个端到端的算法。同时，在 Faster RCNN 中，特征提取网络使用了性能更好的 VGG 网络，使得目标检测的精确度得到提升。

3.4 ROI-Align

ROI-Align 是对 ROI pooling 的一个改进，该方法来源于 2017 年的论文《Mask R-CNN》中。

由于 ROI pooling 的输入是 proposals 的坐标，这些坐标由 RPN 计算得到，由于经过对 anchor boxes 的一次回归，所以 proposal 映射到 feature map 上是一个浮点数的坐标，往往是无法对应特定像素的，因此需要进行一次取整，让 proposals 与 feature map 对映。而计算 ROI pooling 时，还要经过一次取整，才能将其映射到指定大小。这两次取整操作其实让 ROI pooling 得到的特征对应的原始区域与 proposal 相差很大，导致了目标检测的性能不佳。

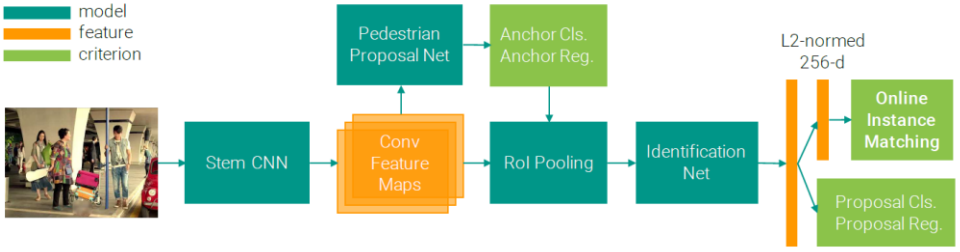


而 ROI-Align 使用双线性插值的方法计算每一个浮点坐标在 feature map 上对应的值，从而进行类似于 ROI pooling 时不会产生很大误差。

3.5 OIM

OIM 是行人搜索网络的一个损失，其出自 19 年的论文《Joint Detection and Identification Feature Learning for Person Search》。

其对应的网络结构如下图所示：



相较于 Faster RCNN，其不同之处在于：该网络原来提取的 ROI 特征向量从另一个分支使用全连接层降维到 256 维，并进行 L2 规范化，使用这个结果进行在线实例匹配（Online Instance Matching, OIM）。

在行人搜索中，需要将 query 中的人像与 gallery 中的检测到的人进行匹配，而 query 中的人像通常类别很多，因此使用 softmax 并不很合适。

在模型推理时，该模型首先将 query 中的人像通过网络的 ReID 阶段提取的特征向量保存到 LUT（lookup table）中，然后将 gallery 中的图像通过全部的网络，计算 proposal 区域得到的人的特征向量，与 LUT 中的一一比较余弦相似度，从而得到最匹配的人，同时还要在线更新 LUT 中最匹配记录的特征向量：

$$v_t \leftarrow \gamma v_t + (1-\gamma)x$$

并进行 L2 规范化。

在模型训练时，使用 OIM 的损失来自两部分，其思想是拉近同一行人的特征向量的距离、拉远不同行人间特征向量的距离，首先计算被框中的行人与被标注前景的余弦相似度，再计算被框中的行人与未被标注前景的余弦相似度，构建 softmax 概率：

$$p_i = \frac{\exp(v_i^T x/\tau)}{\sum_{j=1}^{L}\exp(v_j^T x/\tau)+\sum_{k=1}^{Q}\exp(u_k^T x/\tau)}$$

$$q_i = \frac{\exp(u_i^T x/\tau)}{\sum_{j=1}^{L}\exp(v_j^T x/\tau)+\sum_{k=1}^{Q}\exp(u_k^T x/\tau)}$$

其中，LUT 仍需在线更新，而未被标注前景是指最近的 batch 中的。

然后构建 OIM 部分的损失函数：

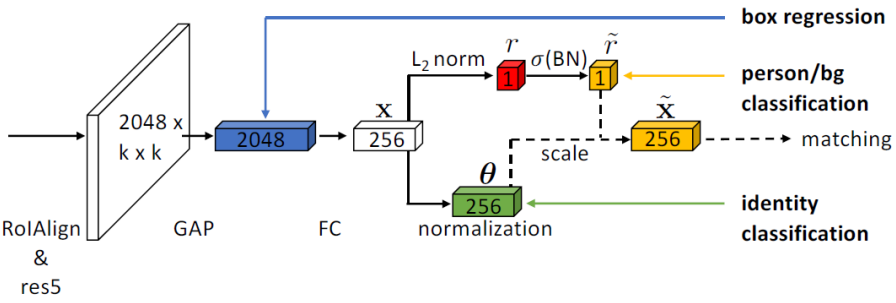
$$\mathcal{L} = \mathrm{E}_x\left[\log p_t\right]$$

整个网络在四个损失的监督下联合训练。

3.6 NAE

NAE 也是对应一种特定行人搜索网络的损失，其出自 20 年的论文《Norm-Aware Embedding for Efficient Person Search》。

该方法对应的网络结构图如下：



NAE 与 OIM 网络不同之处在于前景/背景分类所使用的特征，OIM 使用的是 ROI 特征向量，该特征向量无法很好地反映前景/背景信息，因此在 OIM 的基础上，NAE 网络将提取出的 L2 规范化后的特征向量的长度作为前景/背景分类依据，这个指标可以很好地反映前景/背景信息。

将特征向量 x 进行 L2 规范化的公式如下：

$$\mathbf{x} = r \cdot \boldsymbol{\theta}$$

θ 为 L2 规范化后的单位长度向量，与 OIM 一致，该向量用于训练过程匹配目标人像。而 r 作为 x 的长度，其经过 BN 操作后可以很好的代表前景/背景信息，由于匹配相似度计算为：

$$\text{sim}(\tilde{\mathbf{x}}_q, \tilde{\mathbf{x}}_g) = \tilde{\mathbf{x}}_q^T \tilde{\mathbf{x}}_g = \tilde{r}_g \cdot \boldsymbol{\theta}_q^T \boldsymbol{\theta}_g$$

r 接近 1，代表前景，此时匹配相似度越高；r 越接近 0，代表背景，此时匹配相似度越低。对于 r 的训练采用交叉熵作为监督：

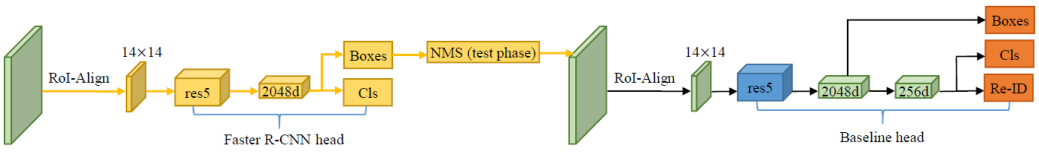
$$\mathcal{L}_{\text{det}} = -y \log(\tilde{r}) - (1-y) \log(1-\tilde{r})$$

NAE 网络在 OIM 基础上进一步精心设计前景/背景分类监督，使得特征向量 x 在在匹配过程中的余弦相似度更具有意义。从而进一步提高行人搜索的准确度。

3.7 SeqNet

SeqNet 是序列化的行人搜索网络，出自 21 年的论文《Sequential End-to-end Network for Efficient Person Search》，在 CUHK-SYSU 数据集上以 94.8 的 mAP 成为当时的 SOTA。

SeqNet 在 NAE 的基础上增加了两个改进，一个是增长了网络结构，使得模型可以使用更精确的 proposals 来进行后续的 ReID，如下图所示：



可以看到，ReID 阶段使用的 proposals 是经过完整的 Faster RCNN 回归输出的更为准确的

proposals，这无疑为行人匹配提供了更高质量的区域提案，这使得端到端的模型准确度可以赶上两阶段模型的准确度。

另一个改进是 SeqNet 提出了一种上下文二分图匹配（CBGM）的方法，该方法充分利用了图像中的上下文信息，即图像中可能包含的其他行人的信息，使得匹配过程不是单一地用框出的人与所有待匹配人进行，而是考虑整个图像中所有人与待匹配人的最佳匹配。以论文中的例子进行说明，如下图：

若仅考虑(a)与(c)、(d)的匹配，那么(a)将与更高概率的(d)错误匹配；但如果考虑 Query 的整张图像的匹配，即考虑(a)、(b)与(c)、(d)的匹配，则(b)与(d)将先匹配，此时(a)就能正确地和(c)匹配。这个问题可以抽象为二分图的最佳匹配问题，使用经典的 Kuhn-Munkres 算法求解。

装 订 线

4 实验过程

4.1 实验环境

4.1.1 操作系统

Ubuntu 18.04

4.1.2 环境版本

python	3.7.13
black	20.8b1
flake8	3.9.0
isort	5.8.0
numpy	1.16.4
pillow	6.1.0
scikit-learn	0.23.1
scipy	1.5.1
tabulate	0.8.7
torch	1.2.0
torchvision	0.4.0
tqdm	4.48.2
yacs	0.1.8
future	0.18.2
tensorboard	2.4.1
protobuf	3.19.0

4.1.3 显卡

NVIDIA TESLA V100 32G

4.2 数据集

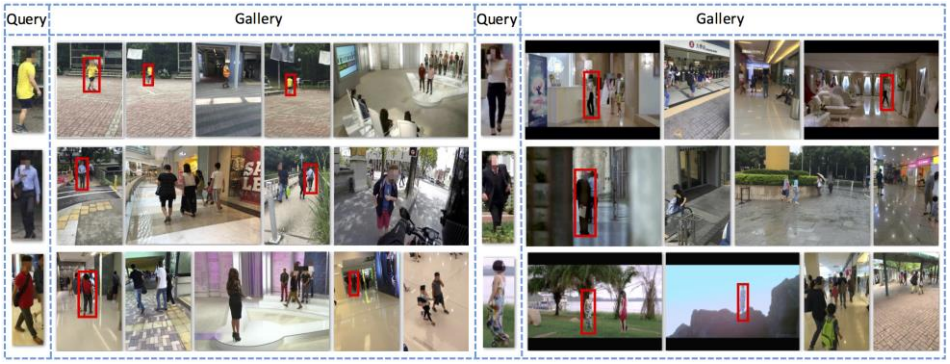
4.2.1 CUHK-SYSU

CUHK-SYSU 是行人搜索领域最常用的数据集。

该数据集是一个大规模的人员搜索基准，包含 18184 张图像和 8432 个身份。

根据图像来源，数据集可以分为两部分：街道捕捉和电影。在街拍中，图像通过手持摄像机收集，跨越数百个场景，并尝试包括视点、照明、分辨率、遮挡的变化；此外，数据集还选择影视剧作为另一种图像采集来源，因为它们提供了更多样化的场景和更具挑战性的视角。

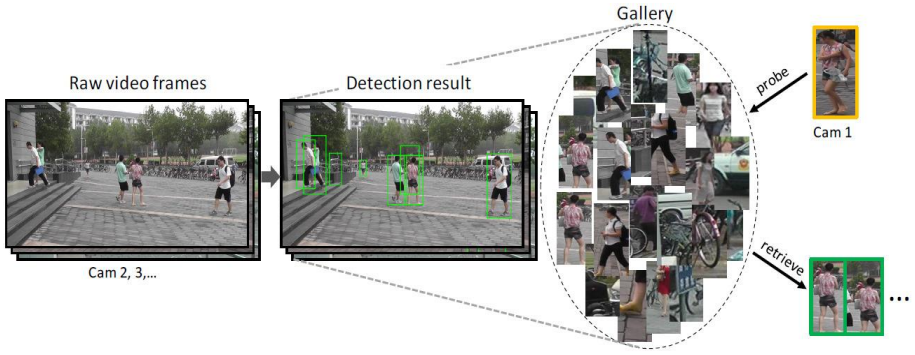
该数据集为人的重新识别和行人检测提供注释。每个查询人出现在至少两个图像中，并且每个图像可以包含多个查询人和多个背景人。数据被划分为训练集和测试集。训练集包含 11206 幅图像和 5532 个查询人，测试集包含 6978 幅图像和 2900 个查询人。数据集的概况如下图所示：



4.2.2 PRW

PRW 是行人搜索的另外一个常用数据集。

PRW (Person Re-identification in the Wild) 是一个人员重识别数据集。该数据集采集于清华大学，通过六个摄像机，采集共 10 小时的视频。数据集被分为训练、验证和测试集。训练集包含 5134 帧和 482 个 ID，验证集共 570 帧和 482 个 ID，测试集则包含 6112 帧和 450 个 ID。每帧中出现的所有行人都会被标注边界框，同时分配一个 ID。如下图所示：



4.3 训练配置

训练参数和数值如下：

参数	含义	数值
模型输入		
C.INPUT.DATASET	数据集	"CUHK-SYSU"
C.INPUT.DATA_ROOT	数据集路径	"data/CUHK-SYSU"
C.INPUT.MIN_SIZE	输入最小尺寸	900
C.INPUT.MAX_SIZE	输入最大尺寸	1500
C.INPUT.BATCH_SIZE_TRAIN	训练时批大小	5
C.INPUT.BATCH_SIZE_TEST	测试时批大小	1
C.INPUT.NUM_WORKERS_TRAIN	训练时数据转移线程	5
C.INPUT.NUM_WORKERS_TEST	测试时数据转移线程	1
学习参数		
C.SOLVER.MAX_EPOCHS	最大训练轮次	20
C.SOLVER.BASE_LR	基础学习率	0.003
C.SOLVER.WARMUP_FACTOR	线性预热因子	1.0/1000
C.SOLVER.WEIGHT_DECAY	权重延迟率	0.0005
C.SOLVER.SGD_MOMENTUM	随机梯度下降的动量	0.9
损失函数参数		
C.SOLVER.LW_RPN_REG	RPN 回归损失的权重	1
C.SOLVER.LW_RPN_CLS	RPN 分类损失的权重	1
C.SOLVER.LW_PROPOSAL_REG	候选框回归损失的权重	10
C.SOLVER.LW_PROPOSAL_CLS	候选框分类损失的权重	1
C.SOLVER.LW_BOX_REG	BBox 回归损失的权重	1
C.SOLVER.LW_BOX_CLS	BBox 分类损失的权重	1
C.SOLVER.LW_BOX_REID	OIM 损失的权重	1
RPN 参数		
C.MODEL.RPN.NMS_THRESH	非极大抑制算法的阈值	0.7
C.MODEL.RPN.BATCH_SIZE_TRAIN	anchors 个数	256
C.MODEL.RPN.POS_FRAC_TRAIN	训练时正负样本比例	0.5
C.MODEL.RPN.POS_THRESH_TRAIN	正样本的 IOU 置信度	0.7
C.MODEL.RPN.NEG_THRESH_TRAIN	负样本的 IOU 置信度	0.3
C.MODEL.RPN.PRE_NMS_TOPN_TRAIN	训练时 NMS 前候选框数	12000
C.MODEL.RPN.PRE_NMS_TOPN_TEST	测试时 NMS 前候选框数	6000
C.MODEL.RPN.POST_NMS_TOPN_TRAIN	训练时 NMS 后候选框数	2000

C.MODEL.RPN.POST_NMS_TOPN_TEST	测试时 NMS 后候选框数	300
ROI 参数		
C.MODEL.ROI_HEAD.BN_NECK	是否使用 BN	True
C.MODEL.ROI_HEAD.BATCH_SIZE_TRAIN	ROI 训练集大小	128
C.MODEL.ROI_HEAD.POS_FRAC_TRAIN	ROI 正例比率	0.5
C.MODEL.ROI_HEAD.POS_THRESH_TRAIN	ROI 正样本阈值	0.5
C.MODEL.ROI_HEAD.NEG_THRESH_TRAIN	ROI 负样本阈值	0.5
OIM 参数		
C.MODEL.LOSS.LUT_SIZE	LUT 表大小	5532
C.MODEL.LOSS.CQ_SIZE	CQ 表大小	5000
C.MODEL.LOSS.OIM_MOMENTUM	LUT 更新动量	0.5

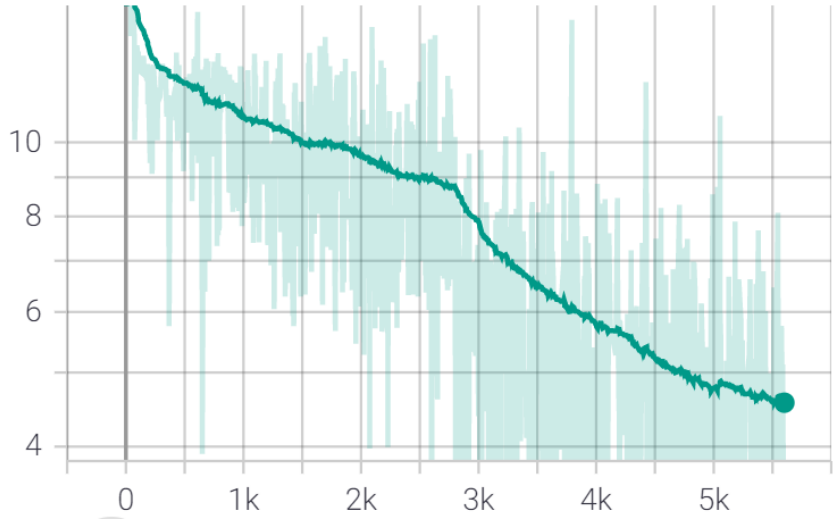
装 订 线

4.4 训练过程可视化

使用 Tensorboard 可视化训练损失的变化过程：

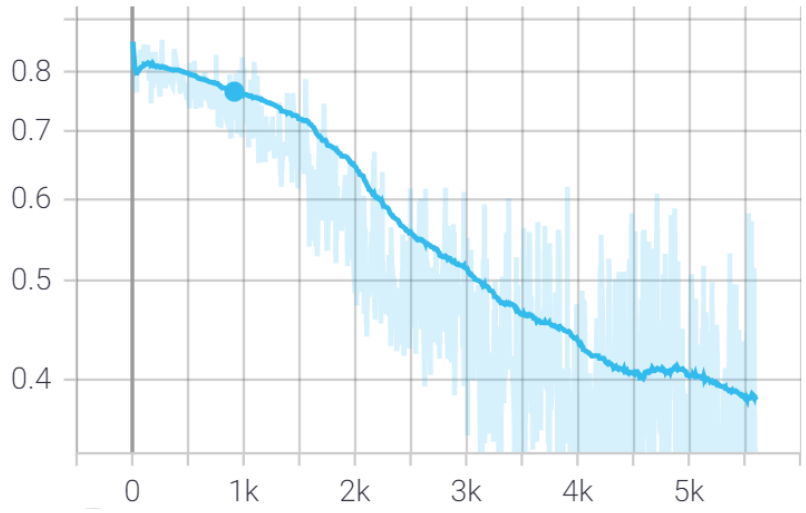
4.4.1 ReID 损失

ReID 损失是行人搜索任务中与搜索精度最直接的损失，它反映了匹配的准确率。



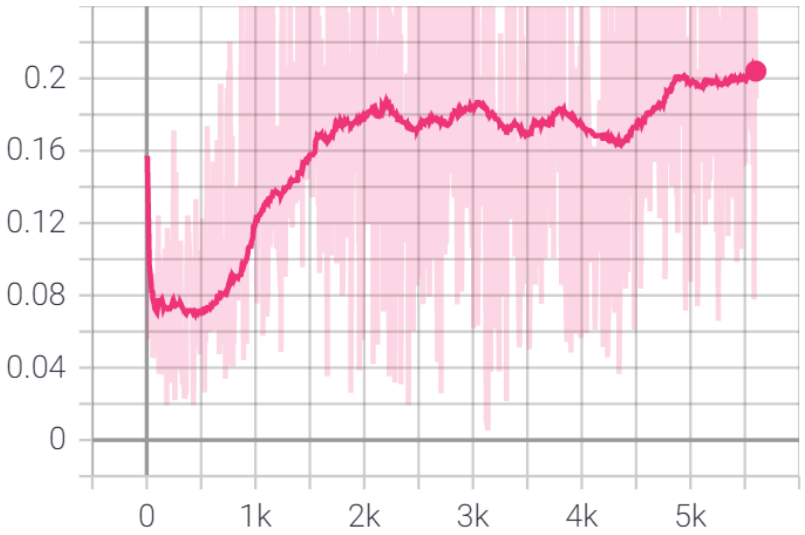
可以看到随着训练过程的进行，ReID 损失不断下降，这表明模型的匹配精度在逐步提升。

4.4.2 Box 分类损失



可以看到随着训练过程的进行，Bounding Box 分类损失不断下降，这表明模型框出来的物体的分类精度在逐步提升。

4.4.3 Box 回归损失



可以看到随着训练过程的进行，Bounding Box 回归损失趋于平稳，这表明模型框出来的物体的位置精度在趋于平稳。

5 实验结果

5.1 评价指标

5.1.1 mAP

mAP 是目标检测领域最常用的评价指标。

对于每张图片中的每个预选框，将其与 IOU 值最大的 GT 框相对应起来，若 IOU 值大于 0.5 且没有 IOU 更大的框与之对应，则该预选框是一个正样本；若 IOU 值小于 0.5 或有其他 IOU 值更大的预选框与之对应，则该预选框是一个负样本。根据候选框分类置信度与给定阈值的关系进行预测：

TP——置信度大于阈值的正样本

TF——置信度大于阈值的负样本

FN——置信度小于阈值的样本或者没有候选框对应的 GT

然后，计算准确率与召回率：

P=TP/(TP+TF)

R=TP/(TP+FN)

在不同的阈值下得到不同的准确率和召回率，从而绘制 P-R 曲线，曲线下的面积对应的类别的 AP 值。

计算所有类别的 AP，取平均数就得到 mAP 值。

5.1.2 top-k 准确率

top-k 准确率是 ReID 领域的常用指标。

top-k 指的是将待匹配图片与 k 个最有把握的图片相匹配，只要 k 个中有一个匹配成功，则认为匹配成功，一般而言 top-1 使用较多。

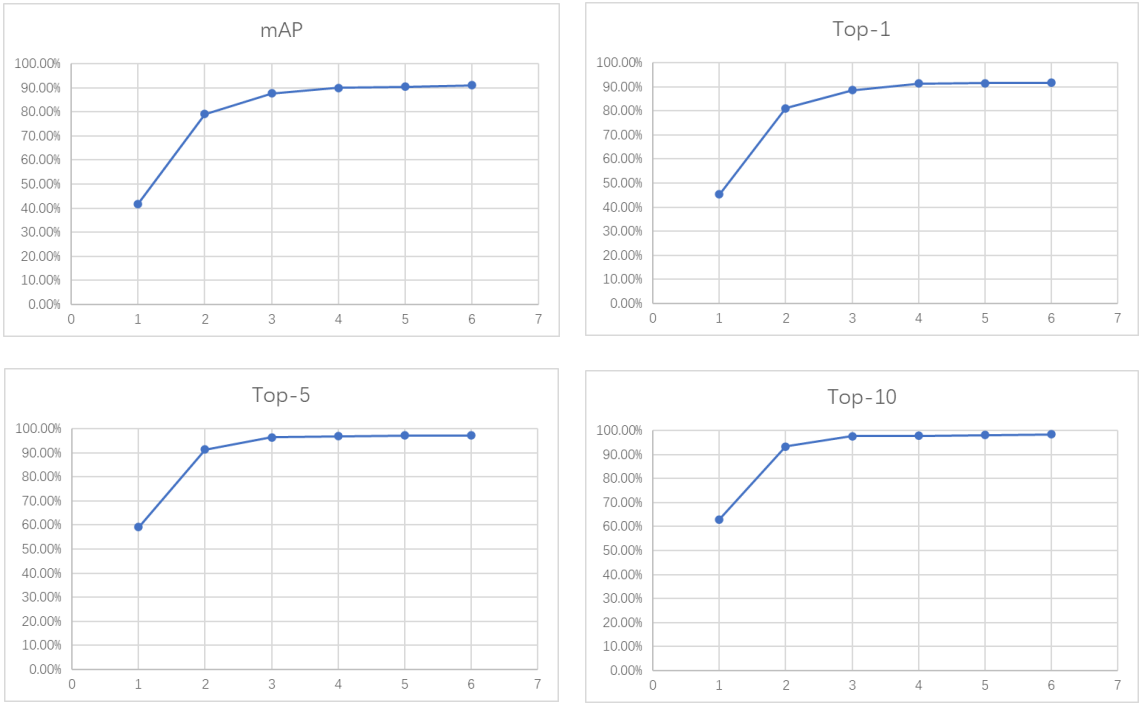
5.2 实验结果

5.2.1 SeqNet

每训练完一个 epoch 评估一次模型的精度，得到的结果（召回率，AP，mAP，top-k）如下表所示：

epoch	recall	AP	mAP	Top-1	Top-5	Top-10
1	34.65%	26.6%	41.74%	45.31%	59.10%	62.93%
2	63.73%	60.36%	79.05%	81.10%	91.38%	93.31%
3	84.19%	79.23%	87.69%	88.59%	96.52%	97.69%
4	87.37%	82.37%	90.02%	91.45%	96.93%	97.86%
5	88.66%	83.79%	90.49%	91.55%	97.28%	98.21%
6	89.73%	84.59%	91.10%	91.76%	97.21%	98.38%

绘制折线图如下：



可以看到，随着训练轮次的增加，模型的准确率（mAP，top-1，top-5，top-10）都在上升，当训练轮次为 6 时，模型达到了 91.10%的 mAP 和 91.76%的 top-1 准确率。

5.2.2 CBGM

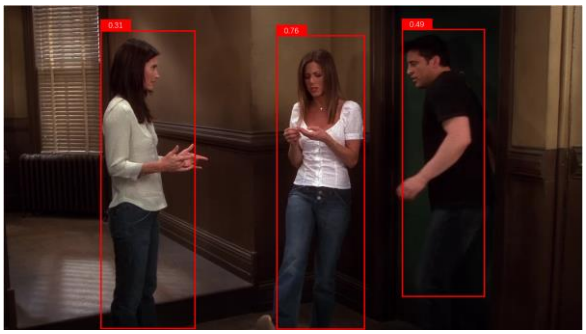
方法	mAP	Top-1
SeqNet	91.10%	91.76%
SeqNet+CBGM	91.68%	92.34%

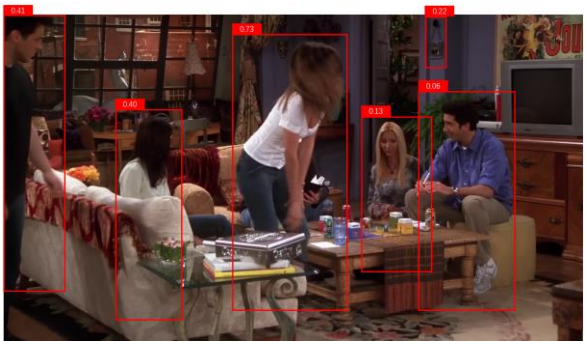
可以看到加入 CBGM 算法后匹配准确率有所提升。

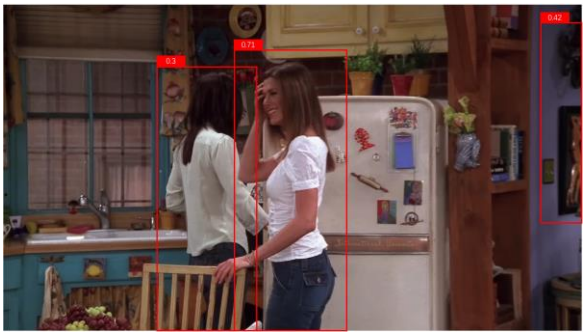
5.3 实例演示

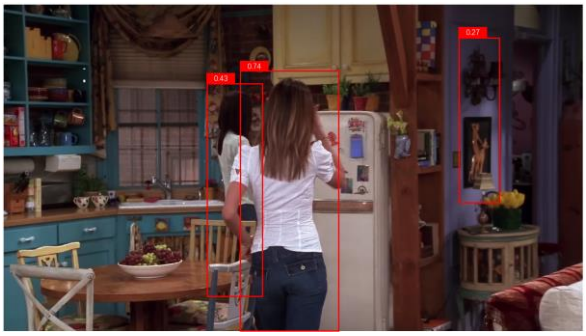
Query:

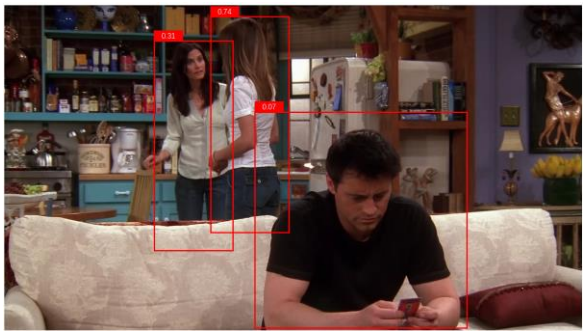
Gallery1-5:











装 订 线

6 总结与思考

6.1 实验总结

在本次《机器学习理论与应用》课程中，我了解到目标检测、行人搜索领域的一些常用算法，也接触到一些比较新的算法。

在确定了以“行人搜索”问题的为我的期中报告的复现内容之后，我从 paper with codes 上搜索到这一领域的前沿研究内容，将 SeqNet 作为我最终需要复现的内容。为了学习 SeqNet 这一模型的结构，我从目标检测的基础方法——RCNN 开始学习，依次学习了 RCNN、Fast RCNN、Faster RCNN 和 mask RCNN，了解到目标检测任务的通用做法。在此基础上，我学习行人搜索任务内容，了解相关数据集信息，并进一步学习了行人搜索的一些方法，包括 OIM 和 NAE 两个方法。最后在此基础上学习 SeqNet，并进行实验的复现，由于实验条件和时间有限，只得到了比原论文略差的结果。

6.2 思考

本次所复现的行人检测方法是一种两阶段的行人检测策略，所谓两阶段指的是候选框的生成和行人的匹配是分开的，这是基于 Faster RCNN 的结果。Faster RCNN 通过 RPN 网络生成候选区域提案，再将区域提案输入到后续的 ReID 网络进行目标检测和识别，加入了 OIM 和 NAE 两个方法的改进后，网络可以用于进行行人搜索。

SeqNet 的优势在于该论文注意到了 proposal 不够准确的问题，因而将经过回归参数调整的区域候选框输入第二阶段的 ReID。由于使用了更为准确的区域提案，目标检测的精度得到了提高。再加之论文提出了上下文二分图匹配算法，使用 CBGM 算法，在推理时考虑了上下文信息，使得模型能够进一步提到精度。

行人搜索或者说目标检测发展至今已经有很多的模型，除了基于 Faster RCNN 的两阶段的方法，还有基于 YOLO 的一阶段方法，也有基于 FCOS 的无锚框的方法，还有基于 Transformer 的方法。目前，行人搜索领域的 SOTA 方法主要是基于 Faster RCNN 的和基于 FCOS 的，因而可以考虑将 YOLO 架构用于行人搜索领域，或者将视觉 Transformer 用于行人搜索领域。另一方面，在匹配过程中，合理利用上下文信息对提高匹配精度有很大帮助，如何合理使用上下文信息也是一个值得研究的内容。

7 参考文献

[1] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[2] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[3] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[5] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[6] Xiao T, Li S, Wang B, et al. Joint detection and identification feature learning for person search[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3415-3424.

[7] Chen D, Zhang S, Yang J, et al. Norm-aware embedding for efficient person search[C].Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 12615-12624.

[8] Li Z, Miao D. Sequential end-to-end network for efficient person search[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(3): 2011-2019.