基于深度学习的局部实例搜索

朱周华,高 凡
(西安科技大学 通信与信息工程学院,陕西 西安 710054)

摘 要:针对传统实例搜索方法准确率和视觉相似度低下的问题,提出一种利用卷积神经网络提取图像全局特征和区域特征的实例搜索方法。该方法经过初步搜索、重排和查询扩展三个阶段实现实例搜索任务,通过微调策略和在重排阶段对特征匹配方法的改进进一步提高检索性能,并将其应用到局部实例搜索任务,即利用残缺图像检索得到整幅图像,在此基础之上,加入在线检索功能。在 Oxford 5k 和 Paris 6k 两个公开数据集上进行实验验证,结果表明,整幅图像的检索 mAP 值和视觉相似度都得到了很大提升,局部实例检索的 mAP 值均高于其他文献中整幅图像的检索,仅比文中整幅图像的检索低 0.032。因此,提出的实例搜索方法不仅提高了实例搜索的准确率,也增强了目标定位的准确性,同时很好地解决了局部实例搜索问题。

关键词:深度学习;局部实例搜索;区域特征;微调;特征匹配
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2020)09-0036-07
doi:10.3969/j.issn.1673-629X.2020.09.007

Local Instance Search Based on Deep Learning

ZHU Zhou-hua, GAO Fan
(School of Communication and Information Engineering, Xi'an University of Science and Technology, Xi'an 710054, China)

Abstract: In order to solve the problem of low accuracy and visual similarity of traditional instance search methods, an instance search method for extracting global and regional features of images using convolutional neural networks is proposed. The method realizes instance search task through three stages: filtering stage, spatial re-ranking and query expansion. The retrieval performance is further improved by fine-tuning strategy and the improvement of the feature matching method in the re-ranking stage. It is applied to the local instance search task, that is, using the incomplete image retrieval to obtain the whole image. On this basis, the online search function is added. Experiments are carried out on two public datasets of Oxford building 5k and Paris building 6k that the mAP value and visual similarity of the whole image are greatly improved. The mAP value of the local instance retrieval is higher than that of other literatures. The retrieval of images is only 0.032 lower than the retrieval of the entire image. Therefore, the proposed method not only improves the accuracy of instance search, but also enhances the accuracy of target location, and solves the problem of local instance search.

Key words: deep learning; local instance search; regional features; fine-tuning; feature matching

0 引 言

信息时代,数字图像和视频数据日益增多,人们对于图像检索的需求也随之增大。传统的基于内容的图像检索(CBIR)都是定义在图像级别的检索,查询图片背景都比较单一,没有干扰信息,因此可以提取整个图片的特征进行检索。但是,现实生活中的查询图片都是带有场景的,查询目标仅占了整幅图的一部分,直接将查询图与数据库中的整幅图像进行匹配,准确率必然会很低。因此,考虑使用局部特征进行实例搜索。

实例搜索是指给定一个样例,在视频或图像库中找到包含这个样例的视频片段或者图片,即找到任意场景下的目标对象。早期,实例搜索大多采用词袋模型(bag-of-words, BoW)对图像的特征进行编码,其中大部分都采用尺度不变特征变换(SIFT)[1]来描述局部特征。Zhu 等人[2]首先使用 SIFT 提取查询图片和视频关键帧的视觉特征,接着采用词袋算法对特征进行编码得到一维向量,最后根据向量之间的相似性,返回一个排好序的视频列表,文中借鉴了传统视觉检索的一些方法,但是没有很好地结合实例搜索的特点。2014 年有学者[3]提出采用比 BoW 性能更好的空间

收稿日期:2019-11-03 **修回日期**:2020-03-06
基金项目:国家自然科学基金(61671352)
作者简介:朱周华(1976-),女,副教授,研究方向为信号处理及应用;高 凡(1994-),女,硕士研究生,研究方向为图像处理。

Fisher 向量[4]和局部特征聚合描述符(VLAD)[5]来描述 SIFT 特征的空间关系,从而进行实例检索。

随着深度学习的发展,深度卷积神经网络特征被广泛应用于计算机视觉的各个领域,如图像分类[6-7]、语音识别[8]等,均取得了不错的效果,因此有学者也将其引入到图像检索领域。起初,研究者们利用神经网络的全连接层特征进行图像检索[9],后来很多研究者开始转向卷积层特征的研究[10],并且证明卷积层特征的性能更好。

Eva 等人[11]采用词袋模型对卷积神经网络(CNNs)提取的特征进行编码,然后分别进行初次搜索,局部重排,扩展查询,从而实现实例检索。

实例检索需采用局部特征实现,因此许多生成区域信息的方法相继出现,最简单的是滑动窗口,之后有学者提出使用 Selective Search 生成物体候选框[12-13],但是这些方法将生成候选区域和特征提取分开进行。Faster R-CNN[14]是一个端到端的网络,它可以同时提取卷积层特征和生成候选区域。文献[15]提出将微调之后的目标检测网络 Faster R-CNN 应用到实例检索中,使用区域提议网络(region proposal network, RPN)生成候选区域,从而得到查询图区域特征与数据库图像区域特征,特征匹配之后排序得到检索结果,在两个建筑物数据集上取得了不错的效果。何涛在其论文中[16]针对 Faster R-CNN 网络效率较低的问题提出了端到端的深度区域哈希网络(DRH),使用 VGG16 作为特征提取器,滑动窗口和 RPN 网络得到候选区域,并将两种方法进行对比,整个网络最后阶段对特征进行哈希编码并计算汉明距离进行排序,从而得到检索结果,文中为了排除不同场景、不同光照和拍照角度产生的干扰,使用局部信息进行检索。以上两篇文献尽管均使用局部信息进行检索,但都是为了排除干扰信息将查询图中的目标标记出来,查询图依然是整幅图像。

实际搜索图像时某些图片会有残缺,此时就无法通过标记目标进行检索,因此实例检索除了常见的利用局部特征进行整幅图像的检索之外,局部图像的检索亦有着非常重要的现实意义。现有的检索方法的检索效果不是很理想,因此文中针对以上两个问题首先改进整幅图像的检索并提高其检索性能,之后利用图像的部分信息(例如建筑物的顶部、嫌疑人的部分特征等)检索得到整幅图像,实现局部图像的检索。同时,考虑到实际检索时,输入均为一幅图像,输出为一组图像,因此,文中在局部实例检索的基础之上加入在线检索功能,可以实现局部图像的实时搜索。因此,主要有以下两大方面的贡献和创新:

(1)基于深度学习的实例搜索。一方面,通过微调策略提高了实例搜索的精确度;另一方面,针对候选框得分(scores)和余弦距(cosine)两种相似性度量方法存在的不足,提出将两种方法相结合,以获得更好的检索效果。

(2)基于深度学习的局部实例搜索。由于局部图像的检索具有重大的现实意义,将全局实例搜索算法应用在局部实例检索任务中,即利用残缺图片信息搜索得到整幅图像,并加入在线检索功能,输入局部查询图,便可以得到查询结果和所属建筑物的名字。

1　相关理论

1.1　基于 Faster R-CNN 的区域特征提取

如图 1 所示,Faster R-CNN 由卷积层(Conv layers),RPN 网络,RoI pooling,分类和回归四部分构成。

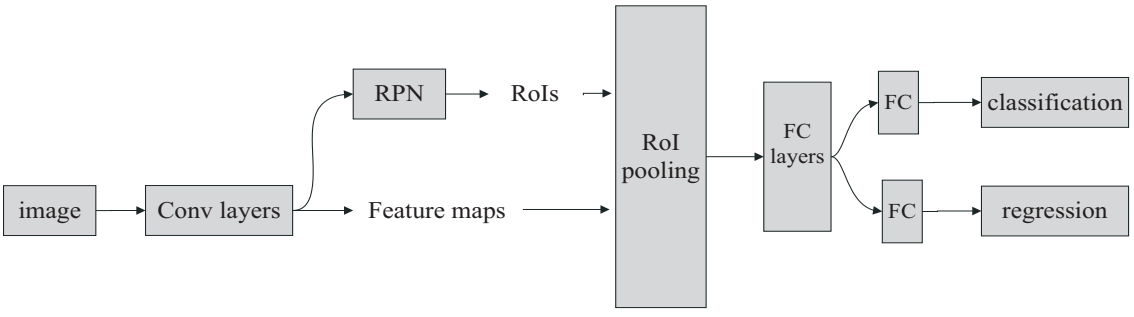


图 1　Faster R-CNN 结构

卷积层用于提取图像特征,可以选择不同结构的网络,输入为整张图片,输出为提取的特征称为 feature maps。文中采用 VGG16[7]的中间卷积层作为特征提取器。

RPN 网络用于推荐候选区域,这个网络是用来代替之前的 selective search[13]的,输入为图片,输出为多个矩形区域以及对应每个矩形区域含有物体的概率。首先将一个 3 * 3 的滑窗在 feature maps 上滑动,每个窗口中心点映射到原图并生成 9 种矩形框(9 anchor boxes),之后进入两个同级的 1 * 1 卷积,一个分支通过 softmax 进行二分类,判断 anchor boxes 属于 foreground 还是 background。另一分支计算 anchor

boxes 的 bounding box regression 的偏移量。Proposal 层结合两个分支的输出信息去冗余后保留 N 个候选框,称为 proposals。

传统的 CNN 网络,输入图像尺寸必须是固定大小的,但是 Faster R-CNN 的输入是任意大小的,RoI pooling 作用是根据候选区域坐标在特征图上映射得到区域特征并将其 pooling 成固定大小的输出,即对每个 proposal 提取固定尺寸的特征图。

分类和回归模块,一方面通过全连接层和 softmax 层确定每个 proposal 的类别,另一方面回归更加精确的目标检测框,输出候选区域在图像中的精确坐标。

1.2 微调 Faster R-CNN

微调,即用预训练模型重新训练自己的数据,现有的 VGG16 FasterR-CNN 预训练模型主要是基于 VOC2007 数据集中 20 类常见的物体预训练得到的。文中的数据集是建筑物,与 VOC2007 图片相似度和特征都相差较大,如果依然采用预训练模型,效果必然不好,再加上从头开始训练,需要大量的数据、时间和计算资源。而文中所选用的数据集较小,因此需要进行微调,这样不仅可以节省大量时间和计算资源,同时还可以得到一个较好的模型。微调主要分为以下三个步骤:

(1)数据预处理。

首先需要对数据集进行数据清洗,主要去除无效值和纠正错误数据,提高数据质量。其次,由于文中的数据集较小且每类样本分布不均衡,因此选择性地对每类图片作图像增强处理,其中图像增强方法包括水平翻转,增加高斯噪声,模糊处理,改变图像对比度。最后将每幅图片中的目标样例标记出来。

(2)建立训练集和测试集。

一般是按一定的比例进行分配,但是文中选用的数据集存在样本不均衡的问题,有些类别特别多,有些类别特别少。由于是将所有图片全部放入同一个文件夹,然后依次读取样本分配训练集和测试集,如果按比例分配,小类样本参与训练的机会就会比大类少,训练出来的模型将会偏向于大类,使得大类性能好,小类性能差。平衡采样策略就是把样本按类别分组,每个类别生成一个样本列表,制作训练集时从各个类别所对应的样本列表中随机选择样本,这样可以保证每个类别参与训练的机会比较均衡。

(3)修改相关网络参数,进行训练。

主要修改网络文件中的类别数和类名,然后不断调节超参数,使性能达到最好。

1.3 实例搜索

实例检索一般经过初次搜索,局部重排,扩展查询三部分完成。

初次搜索首先提取查询图和数据库所有图像的全局特征,然后计算特征之间的相似度,最后经过排序得到初步的检索结果。文中提取 VGG16 网络的最后一个卷积层(Conv5_3)特征。

局部重排是将初次搜索得到的前 K 幅图片作为新的数据库进行重排,基本思路是提取查询图和数据库的区域特征并进行匹配,根据匹配结果进行排序从而得到查询结果。这里查询图和数据库的区域特征提取方法不同,其中,查询图的区域特征是用 groundtruth 给定的边界框对整幅图像特征进行裁剪得到的,而数据库的区域特征是经过 RPN 网络和 RoI pooling 池化得到的特征(pool5)。

扩展查询(query expansion,QE)是取出局部重排返回的前 K 个结果,对其特征求和取平均作为新的查询图,再做一次检索,属于重排的一种。

2 实 验

2.1 数据集和评价指标

(1)实验环境。

文中所有实验均在 NVIDIA GeForce RTX 2080 上进行,所用系统为 Ubuntu 18.04,使用的深度学习框架为 Caffe,编程语言为 Python。

(2)数据集介绍。

在两个公开的建筑物数据集 Oxford[17] 和 Paris[18] 上进行实验。其中 Oxford 包含 5 063 张图片,Paris 包含 6 412 张图片,但是有 20 张被损坏,因此有 6 392 张图片可用。两个数据集都来自 Flickr,共有 11 类建筑物,同一种建筑物有 5 张查询图,因此每个数据集总共有 55 张查询图,除此之外,两个数据集相同类别建筑物的场景、拍照角度和光照都有所不同,而且有很多本来不是同一种建筑物但是从表面看上去却非常相似的图片。

(3)评价指标。

平均精度均值(mean average precision,mAP)是一个反映了图像检索整体性能的指标,如式(1)和(2)所示。

$$\mathrm{MAP}=\frac{\sum_{k=1}^{N}\mathrm{AP}}{\text{查询图的个数}} \tag{1}$$

$$\mathrm{AP}=\frac{\sum_{k=1}^{N}(P(k)\cdot \mathrm{rel}(k))}{\text{相关图片个数}} \tag{2}$$

其中,N 表示返回结果总个数,$P(k)$ 表示返回结果中第 k 个位置的查准率,$\mathrm{rel}(k)$ 表示返回结果中第 k 个位置的图片是否与查询图相关,相关为 1,不相关为 0。MAP 是多次查询后,对每次检索的平均精度 AP 值求

和取平均。这里对是否相关做进一步解释:两个数据集的 groundtruth 有三类,分别是 good,ok 和 junk,如果检索结果在 good 和 ok 中,则判为与查询图相关,如果在 junk 中,则判为不相关。

2.2　基于深度学习的实例检索

2.2.1　方　法

先尝试使用 VGG16 Faster R-CNN 预训练模型进行检索,在两个数据集上 MAP 值仅 0.5 左右,接着文中使用微调策略,只冻结了前两个卷积层,更新了之后所有的网络层权值,通过不断调参,训练一个精度尽可能高的模型。其中,微调过程中分别采用数据增强和平衡采样技术对数据进行处理。具体地,对于 Oxford 数据集,建筑物 radcliffe_camera 的数量高达 221 张,而建筑物 pitt_rivers 仅有 6 张,其他 9 类样本数量在 7 至 78 之间不等,数量差距相当大,因此,选择性地对数量小的 6 类样本通过上面提到的方法进行数据增强,使小类样本数量增大。对其进行数据增强之后,虽然样本数量差距缩小,但是依然存在不均衡的问题,如果将所有样本放入一个文件夹中,按比例分配训练集和测试集,则依然会导致训练出来的模型小类性能差的问题。因此,将每类样本生成一个列表,再从每个列表中随机选取一定数量的样本作为该类的训练样本。

除此之外,文献[15]在局部重排部分使用了两种特征匹配方法,一种是直接利用候选框对应得分(scores)进行排序。数据库中每幅图片经过 Proposal layer 会得到 300 个区域提议(proposal)的坐标和对应得分,找到查询图对应类的最高得分作为查询图和数据库每幅图片的相似度,再从高到低进行排序就可以得到检索结果。另一种是利用余弦距(cosine)进行排序。数据库中的每幅图片经过 RoI pooling 可以得到 300 个特征向量,计算查询图与数据库中每幅图片的 300 个区域特征的余弦距,最小距离对应的候选框即就是和查询图最相似的区域提议,然后把所有的最小距离再从小到大进行排序,就可以得到相似度排序。第一种方法虽然得到的边界框(bounding box regression)定位较准,mAP 值也很高,但是视觉相似度并不是很高。而且根据候选框得分进行排序,每类建筑物的得分和排序都是一定的,因此每次相同类别的不同查询返回结果都是相同的,不会根据查询图片的不同而返回不同的排序。第二种方法得到的检索结果,图像相似度很高,但是边界框定位不是很准确。文中将两种方法结合(scores+cosine),利用余弦距计算相似度并排序,选择得分最高的候选框进行目标定位,这样既解决了视觉相似度不够的问题,也解决了相同类别的不同查询返回结果相同的问题,同时又解决了目标定位不准的问题。

2.2.2　参数设置

本节主要讨论扩展查询中的参数 k 对实验结果的影响,并选取一个最优值作为本实验的默认值。在特征匹配方法选用余弦距的情况下,在两个数据集上分别测试了 k 等于 3、4、5、6、7 时的 mAP 值,实验结果如表 1 所示。

表 1　不同 k 值对 mAP 值的影响

k	Oxford	Paris
3	0.909	0.889
4	0.910	0.888
5	0.910	0.885
6	0.912	0.889
7	0.912	0.888

从表 1 综合来看,文中选用 6 作为 k 的默认值。

2.2.3　结果与分析

表 2 是文中与其他文献 mAP 值的对比,最后两项是文中三种方法的实验结果。可以看出,相比于其他文献,文中方法的检索性能得到很大的提升,比目前最好的方法分别高出 6.1% 和 4%,说明文中方法优于其他检索方法。文中方法与文献[15]所采用方法类似,但结果却得到了很大的改善,通过分析认为,虽然所用数据集都相同,但是生成的训练集和测试集完全不同,而且文中采用数据增强方法,使得样本的数量增加了 3~4 倍,使用平衡采样的方法保证小类样本可以得到和大类样本同样的训练机会。除此之外,网络超参数对于模型的影响非常大,因此文中对参数进行调优,使得训练出来的模型更好。

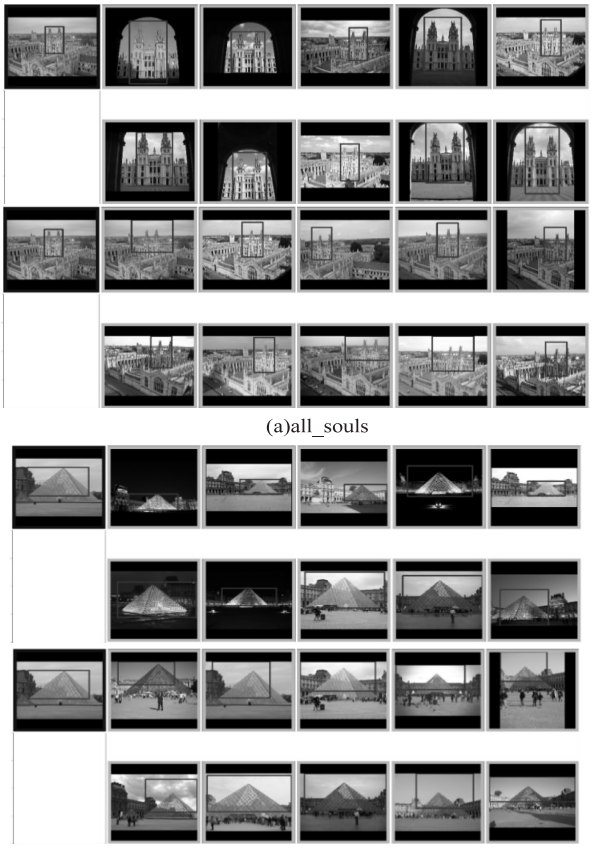
表 2　文中方法与其他方法的 mAP 值对比

方法	Oxford	Paris
R. Tao 等人[3]	0.765	-
Razavian 等人[10]	0.556	0.697
Eva 等人[11]	0.739	0.819
CA-SR[15]	0.772	0.824
CS-SR[15]	0.786	0.842
DRH[16]	0.851	0.849
Ours(scores)	0.903	0.890
Ours(cosine /cosine+scores)	0.912	0.890

图 2 是两组对比图,图(a)和图(b)是对建筑物 all_souls 和 louvre 分别用候选框得分和余弦距进行排序的结果,其中,左边 1 列是查询图,右边 5 列是查询返回结果,表示与查询相关(下同)。图 3 是 all_souls 的两个不同查询用得分进行排序得到的检索结果。图 4 是将两种匹配方法结合得到的检索结果。

从图 2 可以看出利用候选框得分排序得到的结果目标定位较准确,但是返回结果的背景、光照、拍照角

度、颜色、对比度和样例大小与查询图相差很大,而利用余弦距排序得到的结果候选框定位不是很准,但从背景、样例大小等视觉角度来看相似度较高。图 3 中 all_souls 的两个不同查询图返回结果中不仅图片一样,而且顺序也相同。因此可以看出两种方法各有缺陷。



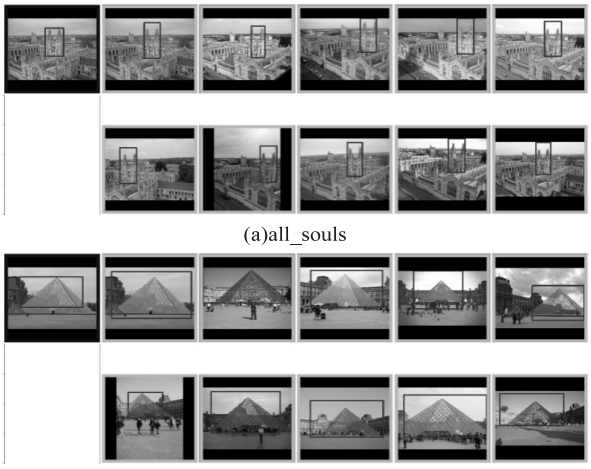
(a)all_souls

(b)louvre

图 2 不同建筑物的同一个查询分别利用得分和余弦距得到的排序

图 3 建筑物 all_souls 的两个不同查询根据得分得到的排序

图 4 是将两种方法结合得到的返回结果,与图 2 相比视觉相似度提高了,目标定位也更准确了,与图 3 相比,同一个建筑物的两个不同查询返回结果也会根据查询图片的不同而改变,且从表 2 最后一行可以看出该方法比使用候选框得分在 Oxford 上得到的 mAP 值高 0.009,与使用余弦距得到的 mAP 值相同。因此认为,以上提出的特征匹配方法得到的返回结果在不降低 mAP 值的基础上提高了检索的准确率。



(a)all_souls

(b)louvre

图 4 综合得分和余弦距两种方法得到的检索结果

2.3 基于深度学习的局部实例检索

2.3.1 方 法

本节局部图像的检索是建立在 2.2 节基础之上的,正是由于整幅图像检索采用候选区域特征实现,局部图像的检索才得以实现。

较之于整幅图像的检索,局部图像的检索具有同样重大的现实意义。生活中常会因为某个原因使图片变得残缺,且难以识别,那么此时就需要使用局部图像检索得到原始图像的完整信息。比如,通过某建筑物的顶部搜索得到整幅图像从而识别该建筑物。或者可以应用在刑侦工作中,当摄像机捕获到的是某犯罪嫌疑人的部分特征时,可以通过已有的部分特征在图像库或者其他摄像头下搜索得到该嫌疑人的完整信息。

由于目前没有一个现成的残缺图像库,因此本节利用截图工具对整幅图像作裁剪处理以模拟残缺图像,即从图像库选取不同实例通过裁剪得到不同大小,不同背景,不同角度,不同颜色的局部查询图。由于图像库图像都是整幅图像,在尺寸和包含的信息方面与局部查询图相差很大,因此局部检索最大的难点在于如何处理局部图像。2.2 节会对输入图片统一进行缩放,那么局部查询图片输入后,先进行放大,则会导致原始输入图像失真,提取特征后再对其进行裁剪又会进一步丢失大部分图像信息,因此根本无法得到正确的检索结果。文中对其进行以下处理:即输入查询图后,先将局部查询填充至与数据库图像相同大小(图像库的图像基本都是 1 024 * 768 或者 768 * 1 024 大小的),这样对图像进行统一缩放,提取特征,按比例

裁剪之后,得到的正是局部图像的特征,再与图像库匹配,则会输出正确的排序结果。本节输入为建筑物的局部图像,输出为局部查询所属的建筑物图像,并且会标记出局部查询在所属建筑物中的位置。

目前,很多文献(如[15])都比较注重算法的研究,基本都采用离线的形式实现图像检索,不仅离线建立特征库,查询图也是成批输入到网络中进行离线检索,得到的结果也是成批保存起来,可是实际应用中,一般都是将查询图逐幅输入进行实时检索,因此文中在前文基础之上,加入了在线检索功能,最后实现在线局部实例检索。

2.3.2　结果与分析

如图 5 是通过裁剪建筑物 radcliffe_camera 和 triomphe 的原图,得到的 5 个不同查询图,分别选取一幅进行检索,得到了完整的建筑物,且标记出了局部查询图像在整个建筑物中的位置,如图 6 所示。最终 mAP 值分别为 0.880 和 0.857。从检索结果可以看出返回结果的视觉相似度极高,目标定位准确,且 mAP 值高于其他文献中整幅图像的检索准确率。因此,可以证明文中提出的全局搜索算法在局部图像检索任务中亦能取得很好的效果。

(a)radcliffe_camera

(b)triomphe

图 5　两种建筑物的五个局部查询图

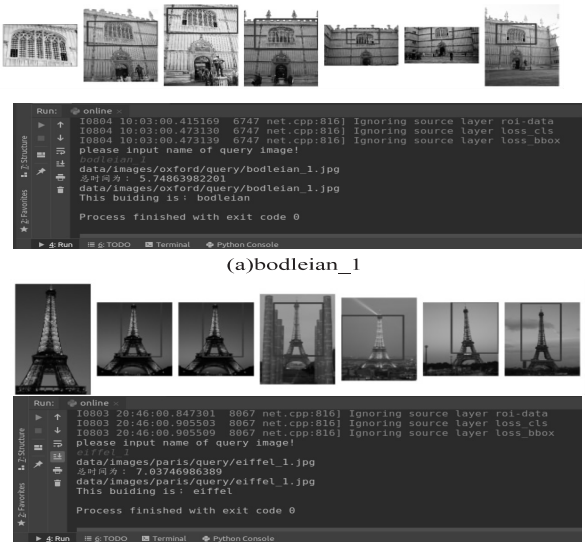
(a)radcliffe_camera_5

(b)triomphe_1

图 6　两组局部查询的检索结果

在此之前,只有文献[19]为了证明行人重识别系统的普适性,使用 CaffeNet 和 VGG16 两个网络模型在 Oxford 数据集上对局部建筑物图像进行了测试,得到的 mAP 值分别为 0.662 和 0.764,远低于文中的准确率。因此提出的局部实例搜索的性能良好。

在线检索功能按照输入图片,可得到查询结果和查询建筑物的名字,且文中在没有使用任何编码算法的情况下,在两个数据集上检索一幅图的平均耗时分别为 5.7 s 和 7 s,经检测,90% 的时间都耗费在利用特征向量计算相似度部分。如图 7 所示,分别是 bodleian 和 eiffel 的检索结果和总耗时。



(a)bodleian_1

(b)eiffel_1

图 7　在线检索结果

3　结束语

为了进一步提高实例检索性能,针对以往的利用候选框得分和余弦距进行特征匹配的不足,提出将两种方法结合,即利用余弦距计算相似度并排序,选择得分最高的候选框进行目标定位。并使用微调策略重新训练预训练模型从而使其适用于文中的实例检索。相比于其他方法,文中采用的方法在性能方面有明显的提升。在此基础之上,利用残缺图像搜索得到整幅图像,性能高于其他文献整幅图像的检索,且仅比文中整幅图像检索低 0.032,实验结果证明提出的全局搜索算法同样适用于局部图像检索任务。之后加入在线检索功能,在没有任何编码的情况下检索一幅图像平均耗时仅需 5.7 s ~ 7 s。在未来的工作中,可以进一步加入编码模块,以提高检索速度,并且可以在更大的数据集上进行测试。

参考文献:

[1] LOWE D G. Object recognition from local scale-invariant features [C]//International conference on computer vision.

Kerkyra, Greece: IEEE, 1999: 1150-1157.

[2] ZHU C, SATOH S. Large vocabulary quantization for searching instances from videos[C]//International conference on multimedia retrieval. New York, NY, United States: ACM, 2012: 1-8.

[3] TAO R, GAVVES E, SNOEK C G M, et al. Locality in generic instance search from one example[C]//Conference on computer vision and pattern recognition. Columbus, OH, USA: IEEE, 2014: 2099-2106.

[4] PERRONNIN F, LIU Y, SÁNCHEZ J, et al. Large-scale image retrieval with compressed Fisher vectors[C]//Conference on computer vision and pattern recognition. San-Francisco, CA, USA: IEEE, 2010: 3384-3391.

[5] GONG Y, WANG L, GUO R, et al. Multi-scale orderless pooling of deep convolutional activation features[C]//European conference on computer vision. Zürich, Switzerland: Springer, 2014: 392-407.

[6] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Neural information processing systems. New York, NY, United States: ACM, 2012: 1097-1105.

[7] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Computer vision and pattern recognition. Boston, MA, USA: IEEE, 2015: 1-14.

[8] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[9] WAN J, WANG D, HOI S C H, et al. Deep learning for content-based image retrieval: a comprehensive study[C]//International conference on multimedia. New York, NY, United States: ACM, 2014: 157-166.

[10] RAZAVIAN A S, SULLIVAN J, CARLSSON S, et al. Visual instance retrieval with deep convolutional networks[J]. ITE Transactions on Media Technology and Applications, 2016, 4(3): 251-258.

[11] MOHEDANO E, MCGUINNESS K, O'CONNOR N E, et al. Bags of local convolutional features for scalable instance search[C]//International conference on multimedia retrieval. New York, NY, United States: ACM, 2016: 327-331.

[12] UIJLINGS J R R, VAN DE SANDE K E A, GEVERS T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[13] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Computer vision and pattern recognition. Columbus, OH, USA: IEEE, 2014: 580-587.

[14] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Neural information processing systems. New York, NY, United States: ACM, 2015: 91-99.

[15] SALVADOR A, GIRÓ-I-NIETO X, MARQUÉS F, et al. Faster R-CNN features for instance search[C]//Conference on computer vision and pattern recognition workshops. Las Vegas, NV, USA: IEEE, 2016: 9-16.

[16] 何 涛. 基于深度学习和哈希的图像检索的方法研究[D]. 成都: 电子科技大学, 2018.

[17] PHILBIN J, CHUM O, ISARD M, et al. Object retrieval with large vocabularies and fast spatial matching[C]//Conference on computer vision and pattern recognition. Minneapolis, MN, USA: IEEE, 2007: 1-8.

[18] PHILBIN J, CHUM O, ISARD M, et al. Lost in quantization: improving particular object retrieval in large scale image databases[C]//Conference on computer vision and pattern recognition. Anchorage, AK, USA: IEEE, 2008: 1-8.

[19] ZHENG Z, ZHENG L, YANG Y. A discriminatively learned CNN embedding for person re-identification[J]. Multimedia Computing, Communications, and Applications, 2017, 14(1): 1-20.