分类号：TP181　　密　　级：公　开
U D C ：_____　　单位代码：10424

学 位 论 文

基于深度卷积神经网络的实例级别图像检索研究

梅 舒 欢

申请学位级别：硕士学位　　专业名称：计算数学
指导教师姓名：段　　华　　职　　称：副 教 授

山 东 科 技 大 学
二零一八年六月

论文题目：

基于深度卷积神经网络的实例级别图像检索研究

作者姓名：梅舒欢　　入学时间：2015年9月

专业名称：计算数学　　研究方向：计算理论和数据处理

指导教师：段华　　职　　称：副教授

论文提交日期：2018年5月

论文答辩日期：2018年6月

授予学位日期：

Research on Case-level Image Retrieval Based on Deep Convolutional Neural Network

A Dissertation submitted in fulfillment of the requirements of the degree of

MASTER OF SCIENCE

from

Shandong University of Science and Technology

by

Shuhuan Mei

Supervisor: Professor Hua Duan

College of mathematics and system sciences

June 2018

学位论文原创性声明

本人呈交给山东科技大学的这篇硕士学位论文，除所列参考文献和世所公认的文献外，全部是本人攻读学位期间在导师指导下的研究成果。除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

若有不实之处，本人愿意承担相关法律责任。

硕士生签名：

日　　期：

学位论文使用授权声明

本人完全了解山东科技大学有关保留、使用学位论文的规定，同意本人所撰写的学位论文的使用授权按照学校的管理规定处理。

作为申请学位的条件之一，学校有权保留学位论文并向国家有关部门或其指定机构送交论文的电子版和纸质版；有权将学位论文的全部或部分内容编入有关数据库发表，并可以以电子、网络及其他数字媒体形式公开出版；允许学校档案馆和图书馆保留学位论文的纸质版和电子版，可以使用影印、缩印或扫描等复制手段保存和汇编学位论文；为教学和科研目的，学校档案馆和图书馆可以将公开的学位论文作为资料在档案馆、图书馆等场所或在校园网上供校内师生阅读、浏览。

（保密的学位论文在解密后适用本授权）

硕士生签名：　　导师签名：

日　　期：　　日　　期：

学位论文审查认定书

研究生　　在规定的学习年限内，按照培养方案及个人培养计划，完成了课程学习，成绩合格，修满规定学分；在我的指导下完成本学位论文，论文中的观点、数据、表述和结构为我所认同，论文撰写格式符合学校的相关规定，同意将本论文作为申请学位论文。

导师签名：

日　期：

摘 要

与传统物体检索相比，实例级图像检索有一系列难点，如：相同类别之间差异大（例如，光照，旋转，遮挡，裁剪等），类别与类别之间差异不大（可口可乐瓶与雪碧瓶），图像含有大量的干扰信息（如背景图像）以及有大量的未经标注的干扰图像等。最近的进展表明，卷积神经网络（CNN）可以提供了一个比传统方法更加优秀的图像特征表示方法。但是，卷积神经网络从整个图像中提取的特征包含大量的干扰信息，会导致检索性能达不到预期效果。为了解决这个问题，本文提出了两种解决方法。

一是一种基于 FasterRCNN 检测的用于实例级图像检索的方法，它有两个阶段，即 Faster R-CNN 离线训练和在线实例检索。首先，训练 FasterR-CNN 模型以定位物体所在的区域。然后，提取物体所在区域的 CNN 特征并将这些特征融合成图像的整体特征，最后通过计算整体特征之间的欧式距离来得到检索结果。本文分别在 INSTRE 和 Oxford 数据集上进行了实验，实验结果验证了本文方法的有效性。

二是一个新的实例级图像检索框架。该框架由两个阶段组成。首先，本文通过区域提议网络（RPN）去检测图像，将其检测结果输入双损失正则化三连体网络（DLRTN）。其次，通过计算排名子网络和分类子网络的损失函数，并利用计算结果来优化该网络。然后，本文引入区域广义均值池化（RGMP）层，对来自双损失正则化三连体网络输出的特征映射进行池化并产生区域广义卷积激活（R-GAC）作为全局图像表示。最后，通过在图像检索数据集的实验证明了本文所提出的图像检索框架的有效性。

关键词：深度学习；实例级别物体检索；三连体网络；区域广义均值池化

ABSTRACT

This paper focuses on the problem of instance-level image retrieval. Compared with the traditional object retrieval, instance-level image retrieval has a series of difficulties, such as: the difference between the same categories (for example, lighting, rotation, occlusion, cropping, etc.), the difference between categories and categories is not great (Coca Cola bottle and Sprite bottle), the image contains a large amount of interference information (such as background images) and a large number of unlabeled interference images. Recent developments have shown that Convolutional Neural Networks (CNN) can provide an image feature representation that is superior to traditional methods. However, the features extracted from the entire image by the convolutional neural network contain a large amount of interference information, which may cause the retrieval performance to be less than expected. In order to solve this problem, this paper proposes two solutions.

The first is a method based on Faster R-CNN detection for instance-level image retrieval. It has two stages, namely Faster R-CNN offline training and online instance retrieval. First, the Faster R-CNN model is trained to locate the area where the object is located. Then, CNN features of the region where the object is located are extracted and integrated into the overall features of the image. Finally, the Euclidean distance between the overall features is calculated to obtain the search result. This paper carries out experiments on Instre and Oxford datasets respectively. The experimental results verify the effectiveness of the proposed method.

The second is a new instance-level image retrieval framework. The framework consists of two phases. First, this paper uses the Regional Proposal Network (RPN) to detect the image, and its detection result is input into the Dual Regularized Triple Network (DLRTN). Second, by calculating the loss function of the ranking subnetwork and the classification subnetwork, and using the calculation results to optimize the network. Then, this paper introduces the regional generalized mean pooling (RGMP) layer, pools the feature maps from the output of the dual regularized triplets network and generates the regional generalized convolution activation (R-GAC) as the global image representation. Finally, experiments on image

retrieval datasets demonstrate the effectiveness of the proposed image retrieval framework.

Keywords: Deep learning; Instance-level object retrieval; Triplet network; Regional generalized-mean pooling

目 录

Contents

1 绪 论

1.1 研究背景及意义

在 Web2.0 时代，伴随着腾讯，百度等不同的社交网站和搜索网站的流行，每天都会产生数以亿计的图像等信息。如何从这庞大的数据库中快速准确的查询并检索出用户所需的图像成为多媒体领域一个急需解决的问题。为了解决这个问题，人们提出了基于图像内容的图像检索方法，这种方法充分发挥了计算机擅长处理重复有规律问题的优势，大大减少了人工标注所需要的巨大的人力物力成本。经过十多年的发展，基于内容的图像检索技术已广泛应用于人们生活的各个方面[1]。

按描述图像内容方式的不同，图像检索可以分成两类，一类是基于文本的图像检索，另一类是基于内容的图像检索[2]。

基于文本的图像检索方法开始于上世纪 70 年代，它主要利用文本描述的方式描述图像的特征，如绘画作品的物体、作者、年代、尺寸等，但是，这种方法需要人工提供大量的标注信息，因此需要消耗巨大的人力物力。由于每个人的认知水平不同，所以人工标注也没办法客观统一，比如，有的人没有办法精确的描述自己看到的图像特征。而且随着图像数据量的增长，这种方法暴露了越来越多的问题，在 1992 年美国国家科学基金会就图像数据库管理系统新发展方向达成一致共识，即表示索引图像信息的最有效方式是基于图像内容。自此，基于内容的图像检索技术开始逐步建立，并在十多年里得到了迅速的发展。自 2003 年以来，由于 SIFT[3]在处理图像变化方面的优势，基于局部描述符的图像检索已经被广泛研究了十多年。最近，基于卷积神经网络（CNN）[4]的图像表示已经引起图像领域越来越多的关注，并且其他一系列图像任务中的表现令人印象深刻。这两种方法也使得基于内容的图像检索方法拥有实现的可能。

基于内容的图像检索技术在目前人们的生活及相关工业领域都具有广阔的应用前景。例如，日常生活中常用的阿里巴巴中，拥有允许用户抓拍上传服务器，通过服务器检索并返回商品链接和相似衣服的拍立得系统，在版权保护方面，可以快速的从数据库中确认注册的商标是否已经认证等等，基于内容的图像检索技术目前已经深入到了人们身边，为人们的生活生产提供了极大的便利。

1.2 研究内容

实例级图像检索不同于传统的物体检索，它更加注重相似但不同类别物体的区分，例如，尽管可口可乐瓶和米林达瓶具有相似的形状，算法也需要检测它们之间的差异。为了解决这个问题，本文提出了两套方案：

（1）利用区域生成网络先定位目标物体所在区域，再提取目标区域的特征。具体做法如下：首先，为了让用户搜索他们想要获得搜索结果，先预训练 Faster R-CNN[5]网络，然后用预训练好的网络去检测实例物体所在区域。之后，通过微调 VGG16[6]的网络来提取图像的区域特征，然后将区域特征融合成图像的整体特征，最后用图像的整体特征来进行检索。

（2）这种方法可以看成上一种方法的改进版，通过三连体网络[7]，可以更加精确的定位实例物体所在的区域。然后通过排名损失学习更好的图像表示。在本文中，针对从区域提议网络（RPN）中检测到的候选区域，首先提出一个双损失正则化三连体网络（DLRTN），它通过排名子网络和分类子网络来共同优化三连体网络。同时，在训练过程中会不断删除不相关的区域。然后引入三连体网络的广义均值池化（RGMP）层来学习更好的池化策略。在广义均值池化层，将每个区域的特征映射作为输入，综合生成一个图像整体特征（R-GAC）。在测试阶段，将图像直接经过训练的框架，以生成全局图像表示 R-GAC，该图像可以用点积与数据集图像进行比较。

1.3 论文结构安排

针对以上两个方面的研究，本文在后面章节给出详细阐述，论文结构安排如下：

第二章 首先回顾了实例级图像检索的总结，实例级图像检索的研究现状；然后介绍了这些研究工作中常用方法：（1）基于 SIFT 的图像检索方法和（2）基于深度卷积网络的图像检索方法。

第三章重点介绍了基于 Faster R-CNN 检测的图像检索方法，详细描述了具体方法的步骤，参数设置和算法原理，并通过实验验证了方法可行性，并对实验结果做了详细的分析。

第四章介绍了基于三连体网络的图像检索方法，详细描述方法的原理，步骤，以及实验过程中可能出现的问题及对应解决方案，最后通过实验验证方法的有效性，最后，对实验结果做了分析和总结。

第五章对全文进行总结，并对未来的研究工作进行了展望。

2 相关研究概述

2.1 图像检索

图像检索是计算机视觉领域一个基本任务，其目的是在大量未标注图像数据中检索出自己想要的图像数据。因为在现实生活中，随着互联网的发展，每天会产生越来越多的图像数据，而标注这些图像需要投入大量的人力，为了节约人工成本，基于内容的图像检索成了一个急需解决的问题。针对检索的目的，这个任务分为两个子任务，基于内容的图像检索和基于文本的图像检索，本文主要研究基于内容的图像检索。

物体图像检索是指对查询图像中的某一物体，从图像库中找出包含有该物体的图像，物体图像检索的目标是在给定查询图像的情况下，从大量无序图像集合中检索与查询图像相同的对象实例。实例级别图片的物体检索比物体检索有着更加苛刻的要求，在物体检索中，只针对物体类别做检索，但是在实例级图像检索中，会具体到物体的名称。例如：物体检索要求对杯子这个类别做检索，而实例级别物体检索要求检索出杯子的名称。

实例级图像检索有着广泛的应用前景，可视化的地理定位，组织个人的收藏相册和三维重构等等。在深度卷积神经网络出来之前，该方向主要依赖手工特征来完成的，比如 SIFT 及其各种变体。随着深度卷积网络的兴起，其在各个计算机研究方向取得了很好的性能。越来越多的研究者将深度学习的方法用于图像检索，并取得了很多成果。

2.2 国内外相关研究分析

基于内容的图像检索（CBIR）一直是计算机视觉领域的一个长期研究课题。在 20 世纪 90 年代初，CBIR 的研究真正开始。图像通过简单视觉线索（例如纹理和颜色）提取图像的描述符，通过这种描述符在数据库中进行物体检索。根据这种思想延伸出无数算法和图像检索系统。一个简单的策略是提取全局描述符。这个想法在 20 世纪 90 年代和 21 世纪初主导了图像检索领域。然而，这种方法有个很严重的缺陷，全局特征没有办法分辨对应图像的变化，如照明、平移、遮挡和截断。这些差异损害了检索的准确性并限制了全局描述符的应用范围。为了解决这个问题，又出现了基于局部特征的图像检索。这种方法着重用在解决实例级图像检索任务。在这个任务中，给定描述特定对象、场景、体系结构的查询图像，其目的是检索包含可以在不同视角、光照或遮挡下检测的相同对象、场景、体系结构的图像。实例检索不同于类别检索[8]，为后者旨在检索与查询相同

的类别的图像。也就是说，如果没有指定，可以将 “图像检索”和“实例检索”方法视为一样。这些传统方法都被记录在 Smeulders 等人 2000 年提交的一份关于 CBIR 的综合调查报告。

2003 年，词袋（BoW）模型被引入图像检索领域[9]，并在 2004 年被应用于图像分类，这种方法主要依赖于 SIFT 描述符。自那以来，检索领域见证了 BoW 模型十多年来的突出表现，并出现了一系列改进的方法。其中，影响比较大的改进如下，2006 年，Stewénius and Nistér 提出了基于层次聚类的 SIFT 算法[10]，次年，Philbin 提出了改进版年的 Approximate 聚类方法[11]，2008 年，Jegou 等人提出了基于汉明嵌入的改进 SIFT 方法[12]，针对 SIFT 特征的融合策略，Perronnin 等人于 2010 年提出了一种改进的特征向量方法[13]，此后，Jegou 在 2010 年提出了 SIFT 领域最优秀的算法—VLAD（Vector of Aggragate Locally Descriptor）[14]。2012 年，Krizhevsky 等人使用 AlexNet 在 ILSRVC 2012 中实现了最先进的识别精度，大幅超越了以前的最佳结果。从那以后，研究焦点开始转移到基于深度学习的方法[15]，特别是卷积神经网络（CNN）。

基于 SIFT 的方法主要依赖 BoW 模型[16]。 BoW 最初是为建模文档而提出的，因为文本自然被解析为单词，它通过将单词响应累加到全局矢量中来为文档建立单词直方图。在图像领域，引入尺度不变特征变换（SIFT）使得建立 BoW 模型可行。最初，SIFT 由探测器和描述符组成，但彼此孤立的使用。在本文中，如果没有指定，SIFT 通常会引用 128 维描述符，这是检索方法中的一种常见做法。使用预先训练的码本（单词集合），局部特征量化为整体的视觉词汇。图像因此可以获得一个与文档类似的形式表示，并且可以利用经典的加权和索引方案进行检索。近年来，基于 SIFT 模型的流行趋势已经被卷积神经网络（CNN）所取代，这种深层结构已被证明在许多视觉任务中胜过手工制作的 SIFT 特征。在检索任务中，已经报道了与 BoW 模型相比的性能表现，即使用经过未微调微调的 CNN 特征。基于 CNN 的检索模型通常计算紧凑表示，并采用欧氏距离或近似最近邻（ANN）搜索方法进行检索。当前的文献可以直接使用预先训练的 CNN 模型或对特定的检索任务进行微调。这些方法中的大多数仅将图像馈送到网络中以获得描述符。比如 Razavian 等人于 2014 年左右提出了 CNN off-the-shelf[17]和 Babenko[18]等人于同年提出的 Neural codes 算法。但是这些方法在性能上基本与基于 SIFT 特征的算法持平，针对这一点，Ng 在 2015 年的工作中将 VLAD 特征和 CNN 特征相融合，实现了当时最好的检索性能[19]。直到 2016 年 Tolias 提出了 R-MAC[20]算法，CNN 相关算法开始取得了远远超过 SIFT 算法的性能。

2.3 研究方法介绍

为了实现有针对性实例级检索方法，需要针对现有的 CNN 方法进行改进。由于目标数据集具有类别内部差异大而类别与类别之间差距小，背景复杂而混乱的特点。本文的改进方法要从去除背景干扰和学习更好的目标区域图像描述符这两方面着手。

本文尝试使用两种不同的思路去改进现有的 CNN 算法：

（一）使用 Faster R-CNN 对图像进行探测，然后用图像去微调 CNN 网络从而得到更好的特征。通过 Faster R-CNN 可以得到实例物体所在的区域，提取区域的特征相当于把背景干扰信息去除了，从而可以大大提高准确率。而用实验数据集去微调 CNN 网络可以帮助图像特征更好的区别类间差，从而达到提高检索性能的效果。

（二）通过三连体网络去直接学习图像的描述符。首先通过区域提议网络中检测到的候选区域，然后使用双损失正则化三连体网络去学习优化图像的特征。通过这一步，可以直接得到图像的整体描述，并且也去除了背景干扰，从而达到提高检索性能的效果。

后面第三章和第四章将给出详细的描述。

3 基于深度区域卷积神经网络的实例级图像检索

3.1 引 言

图像检索是计算机视觉中一项重要任务。近年来，研究人员在图像检索领域取得了巨大成功。例如，Jegou 等人[21]结合汉明嵌入和弱几何一致性提取视觉特征对象检索。Albert 等人[22]为每幅图像提取了全局和局部的特征表示，最后将这些特征融合为整张图像的特征。

由于检索对象不同，图像检索可以粗略地分为两组：第一组是类别级图像检索，其中数据集中的图像被认为与查询图像相似且他们有相同的物体类别。另一组是实例级别的物体检索，只有图像包含于查询图像相同的物体，才认为图像与查询匹配。实例级别的物体检索更加困难，因为检索方法需要对局部的物体信息进行编码，以便检索出所需的图像，例如，尽管可口可乐瓶与雪碧瓶有着相似的形状，算法应该需要检测它们之间的差异。本文研究的是实例级别图像检索。

在实例级别上有一些现有的图像检索方法，例如基于 SIFT 的相关工作。然而，最近的进展表明卷积神经网络网络（CNN）在多个计算机视觉任务中拥有远超 SIFT 工作的性能。这种方法的流行归功于 GPU 的计算能力和拥有非常大的标注数据集[23]。然而，CNN 特征是从全局图像中提取的，该特征包含大量的背景信息，会导致对象检索的性能下降。

为了解决这个问题，本章采用检测实例所在区域，然后从实例所在的区域提取特征这种方法。实际应用过程中，针对实例级别的图像检索提出了一种 Faster R-CNN 方法。如图 3.1 所示，本章提供的实例级图像检索系统主要由两部分组成：Faster R-CNN 训练和在线实例检索。对于 Faster R-CNN 训练，实例检索系统需要两个基本的结构：首先，为了得到索引图像中实例物体的区域，需要训练 Faster R-CNN 以更好地定位实例检索的对象。其次，用检测到的物体图像区域提取 CNN 特征。对于在线实例检索，使用 Faster R-CNN 来检测实例物体所在的区域，然后从检测到的区域中提取 CNN 特征。索引图像的检索结果是通过这些特征的视觉相似性计算得到的。此外，本章采用并比较三种不同的策略，包括连接，均值池化和最大池化，用来融合实例图像所在区域的特征。

3.2 方法框架

本文探讨了使用由物体检测方法检测图像实例物体所在区域，并通过深度卷积网络提取实例物体所在区域的特征并实现在线检索的检索方法。本章提供的实例级对象检索系统的框架如图 3.1 所示，它分为两个阶段，即 Faster R-CNN 提取区域阶段和在线实例检索阶段。 接下来详细描述这两个阶段。

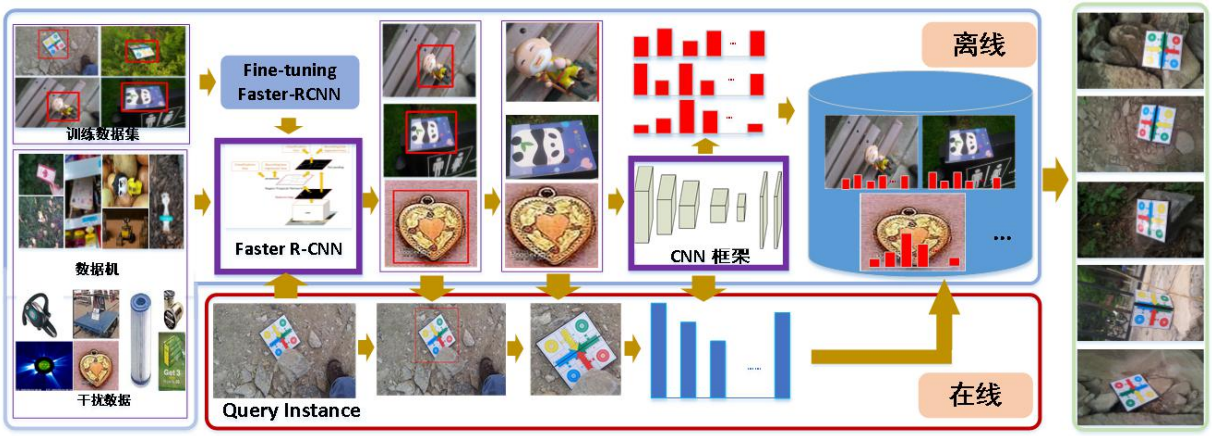


图 3.1 系统的框架

Fig.3.1 The framework of' system

3.2.1 离线的 Faster R-CNN 训练

目标数据集具有类别内部差别小，背景复杂而混乱的特点。为了克服这些困难，通过 Faster R-CNN 方法去获得实例物体所在的区域，并通过这些区域的特征进行检索从而获得更好的实例检索性能。在实际操作过程中，选择以下微调模式。最初的两个卷积图层具有不变的权重，只更新所有后续图层的权重。通过改变卷积特征，使 RPN 提议和全连接层更适应查询实例。由此产生的微调网络将用于提取更好的用于检索的特征。训练 RPN 先要给每个区域一个二进制的标签（不是最终结果）。为以下两种类型的区域分配正例的标签：（i）与具有人工标注的边界框重叠超过 0.7 的区域;（ii）目标区域点用任何人工标注的边界框重叠大于 0.7 的区域。请注意，一个人工标记的边界框可能会为多个目标区域点分配正向标记。为与人工标注的边界框重复比率小于 0.3 的目标区域分配一个副标签。负标签对训练目标不产生任何影响。有了这些定义，可以遵循 Fast R-CNN 中的多任务损失来最小化目标函数。将图像的损失函数定义为

$$L(p_i,t_i)=\frac{1}{N_{cls}}\sum_i L_{cls}\left(p_i,p_i^*\right)+\lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}\left(t_i,t_i^*\right) \tag{3.1}$$

这里 i 是一个训练批次中区域的索引，p_i 表示的是区域为 i 的预测概率，如果区域类别判定正确，p_i^* 为 1， 反之 p_i^* 为 0。t_i 表示预测边界框的 4 个参数坐标，t_i^* 是与正确区域相对应的人工标注的边界框的坐标向量，分类损失 L_{cls} 是两个类别（目标与非目标）的对数损失。

经过微调 Faster R-CNN 之后，利用训练好的模型来获取图像的边界框以及每个边界框的对应分数。 图 3.2 显示了一些例子，包括检测到的两个数据集（a）Oxford 和（b）Instre 的前 5 个边界框及其相应得分。 接下来，使用 CNN 模型从检测到的对象区域中提取深度视觉特征。

图 3.2（a）Oxford 和（b）Instre 的检测到的前 5 个边界框和它们对应的得分

Fig 3.2 (a) The first 5 bounding boxes detected by Oxford and (b) Instre and their corresponding scores

3.2.2 基于深度卷积神经网络的图像表示

CNN 主要用于识别位移，缩放和其他形式的空间不变性的二维图像。 CNN 的特征是通过大量训练数据学习，从浅层特征到深层的模拟人的视觉方式，通过神经元学习从物体边界开始学习，直到形成物体的整体轮廓的表示，这是 CNN 的主要优势。 如上所述，将利用深度卷积网络对图像进行特征表示，本文中采用经典的 VGG16 架构作为基本框架。详细的说，为数据库中的每个图像选择分数较高的 top-K 边界框。根据这些区

域边界框的坐标们使用在 Imagenet 上训练的 VGG16 网络从全连接层（FC7）提取 4096 维特征。

3.2.3 在线的图像检索

给定查询实例图像，根据 Faster R-CNN 提取区域并用 CNN 提取区域的视觉特征，最后计算相似度以返回检索结果。一般情况下，选择每个图像得分最高的边界框作为目标区域，为之后的实例检索提取 CNN 特征。但是，得分最高的检测区域往往与实际实例对象区域不同。所以，单纯使用这种方法可能会失去一些有用的信息。图 3.2 显示了一些例子。对于 oldman 和 parchis 这两个类别，他们的得分最高的区域不是需要检索的对象，从而会导致性能下降。之后，通过观察得分最高的三个检测结果，可以发现实例对象通常位于顶级 K 分数的区域，因此增加这些分数的相应区域将尽可能地增加正确的信息。此外选择一定数量的边界框之后，采用以下三种策略来融合每个实例图像的顶部 K 区域的特征，分别是（1）直接连接（2）平均池化和（3）最大池化。对于直接连接，通过简单地连接其相应的 4096 维特征来融合来自不同区域的特征。通过上述方法，为每个查询实例图像获取相应的搜索结果。通过 Faster R-CNN，能有效地减少了图像背景的干扰。同时，通过 CNN 方法得到了检测区域的高层语义判别信息,最后，通过选择 top-K 边界框并尽可能地保留图像的信息，以减少由于某些类别的检测结果较差带来性能损失。

3.3 实验评估

在本节中，首先描述基本的实验设置，包括数据集和实现细节。然后，定性和定量的评估所提出方法的性能。

3.3.1 数据集

本章在两个实例数据集上验证本文提出的方法(F-R-CNN+CNN)，f 分别是 Oxford105k 和 Instre。

Oxford105k：该数据集[24]由 5062 张牛津地标图像和从 Flickr 收集的额外 100,000 张图像组成。这 5062 张地标图像已经过人工注释，为 11 个不同的地标框定了检索区域，每个地标有 5 个查询图像。10 万张图像与 5062 张图像不相交。

Instre：这个数据集[25]由两个子集组成：Instre-S 和 Instre-M。 Instre-S 包含 200 个单标签类，Instre-M 专门为多个对象设计。在本章工作的实验中选择 Instre-S。 Instre-S 数

据集总共包含 23070 个图像，每个图像都提供有对象位置注释。此外，还有一百万个从 Flickr 抓取下来的干扰图像，用于当数据集扩充到较大规模时测试检索性能。

3.3.2 实验设置

对于 Oxford105k 数据集，只有查询图像有对应的人工标注，并且为了与其他 CNN 方法进行比较，本文也使用该数据集提供的 55 个图像进行查询。对于 Instre 数据集，从每个类中随机选择 75 个图像来形成训练集，以便进行更 Faster R-CNN 的微调。在 Faster R-CNN 网络微调之后，使用训练好的 Faster R-CNN 从数据集中检测区域，然后使用 CNN 提取具有更高分数的检测区域的特征。 本节的实验选择使用 VGG16 网络提取 4096 维特征。 所有的实验都在 Nvidia Titan X GPU 上运行。 类似文献[25]，本文选择平均均值精度（mAP）作为评估指标。作为每个查询的平均精确度得分的均值，mAP 被证明是更好的评价指标。

3.4 评估方法

3.4.1 Oxford105k 评估方法

为了将本节的方法与此数据集上的现有方法进行比较，本节考虑以下标准进行比较：

（1）CNN 的方法(CNN)：直接使用 VGG16 网络为所有图像提取 4096 特征。

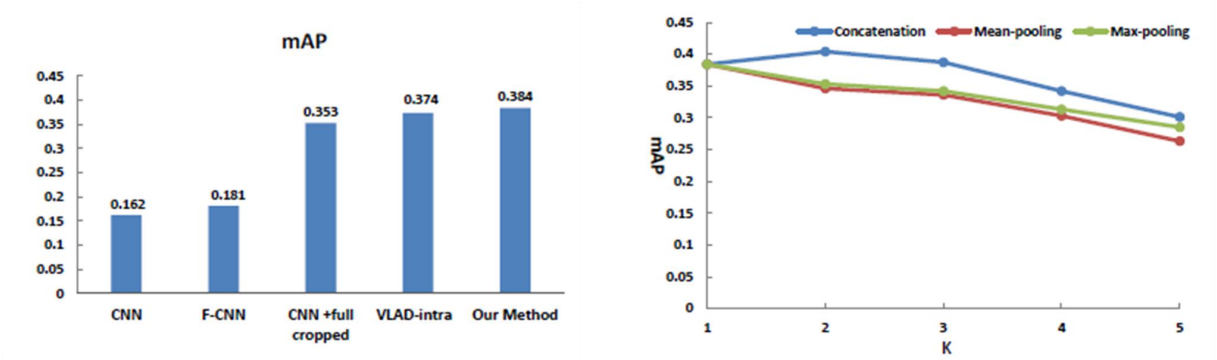
（2）微调 CNN 方法(F-CNN)：在这个方法中，首先对 VGG16 模型进行调整，使用训练数据集对 VGG16 模型进行微调。对 Oxford105k 数据集而言，修改网络中的输出层以返回 11 类概率。经过微调后，按照 CNN 中描述的步骤从所有图像中提取视觉特征。

3.4.2 Oxford105k 实验结果

实验结果如图 3.3（a）所示,本章使用训练的 Faster R-CNN 从数据集中检测区域，然后提取每个图像得分最高的检测区域的 4096 特征。从这些比较结果可以看出：(1)F-CNN 的性能优于 CNN。这是因为经过调整的 CNN 适合当前的任务。 （2）本文的方法表现达到最佳性能。这是因为 F-R-CNN+CNN 可以更准确地提取对象的信息，从而提高检索性能。

由于 Faster R-CNN 中得分最高的检测区域可能不准确，因此将分数较高的区域中的特征相加可以提高检索结果的性能。为了验证它，根据每幅图像中候选区域的不同数量显示不同的融合策略。从图 3.3（b）可以看出：（1）连接策略的性能优于均值池化和最

大池化方法，原因是连接方法比这两种策略保留更多的对象信息。（2）当 K = 2 时，这些方法都达到最佳性能。在这种情况下，连接策略达到最佳 mAP，即 0.404。这意味着 K = 2 实现了正确的对象信息和背景噪声之间的平衡。



(a)不同方法的性能对比　　(b)针对K个候选区域不同融合策略的mAP

图 3.3 Oxford105k 中不同方法的比较

Fig 3.3 Comparison of different methods in Oxford 105k

3.4.3 Instre 评估标准

考虑以下标准进行比较：

•空间编码（SC）[26]：通过 SIFT 特征组成的三维空间图，通过将查询图像旋转 180 度来减少对图像旋转的敏感度，以生成查询扩展的新查询。

•几何编码（GC）[27]：改善了旋转不变性的空间编码。

•组合方向 - 位置一致性（COP）[28]：COP 采用图模型来对每两个候选 SIFT 匹配的相互空间一致性进行建模。

•汉明嵌入+弱几何一致性(HE + WGC)：HE 为每个 SIFT 分配一个二进制签名来编码它在 Voronoi 单元内的位置，并且 WGC 利用 Hough 方案投票进行量化变换。在本章的实验中，使用 64 的签名长度和 22 的汉明阈值。

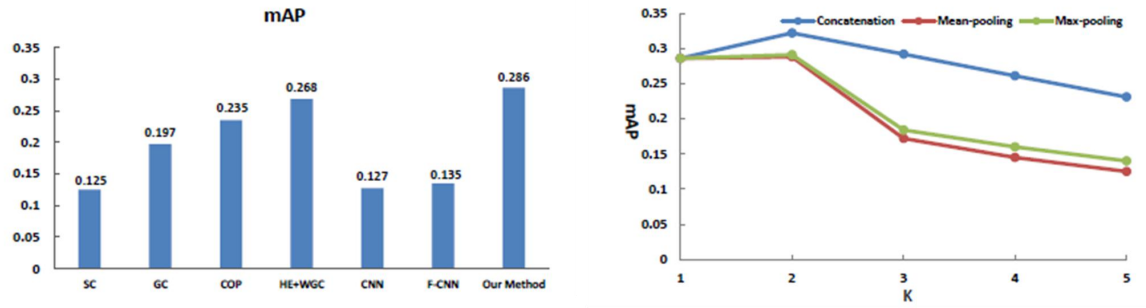
对于基于 CNN 的实验，使用两种方法。一种是传统的 CNN 特征方法，另一种是经过微调的 CNN 方法。

•传统 CNN 的方法(CNN)：在本章中，将使用 VGG16 网络评估性能，以进行实例检索。即使用 VGG16 网络为所有图像提取 4096 特征。

•微调 CNN 方法(F-CNN)：在这个方法中，首先对 VGG16 模型进行调整，使用训练数据集对 VGG16 模型进行微调。对 Instre 数据集而言，修改网络中的输出层以返回 200 类概率。经过微调后，按照 CNN 中描述的步骤从所有图像中提取视觉特征。

3.4.4 Instre 实验结果

本文使用训练的 Faster R-CNN 从数据集中检测区域，然后提取每个图像得分最高的检测区域的 4096 特征。实验结果如图 3.4 所示，从实验结果可以得到：(1) COP 和 HE+WGC 的性能表现优于 CNN。原因是来自 Instre 数据集中图像的背景是一个重要的干扰因素。通过局部特征提取，构造视觉字典，生成原始 BOF 特征，引入 TF-IDF 权重 HE+WGC 可以从目标区域提取特征，大大减少背景干扰。（2）F-CNN 的表现优于 CNN，这是因为经过微调的 VGG16 模型可以更准确地提取实例对象的信息，从而提高检索性能。(3)F-R-CNN+CNN 的性能达到最佳性能。由于筛选方法通过构造特征点的向量来构造向量，然后匹配向量以使图像必须满足足够的纹理，否则构造的向量判别不会太大并且会导致错误匹配。CNN 从整个目标区域提取特征，并且没有这样的限制。考虑到 HE + WGC 和 F-R-CNN+CNN 方法有比其他标准方法具有更好的性能，对这两种方法的实验结果进行了更详细的分析。图 3.5 给出了随机选择的 30 个对象类别的 mAP 性能。在大多数情况中，F-R-CNN+CNN 提供了最佳的表现。由于 Faster R-CNN 中对象得分最高的检测区域可能不准确，因此从分数较高的区域添加特征都可以提高检索结果的性能。为了验证它，根据每幅图像中候选区域数量的不同用不同的融合策略。从图 3.6 可以看出：(1) 连接策略的性能优于均值池化和最大池化方法，原因是连接方法比另外两种策略保留更多的对象信息。(2) 当 K = 2 时，这些方法都达到最佳性能。这意味着 K = 2 实现了正确的对象信息和背景噪声之间的平衡。最后，当 K = 1,2 和 3 时，定性评估 F-R-CNN+CNN 方法的检索结果，比其他标准方法获得更好的检索结果。图 3.7 显示了一些示例结果。正如预期的那样，与 Top1 和 Top3 相比，使用 Top2 能够获得更好的结果，这表明将前两个区域的功能与更高的分数相连是一个有效的解决方案。



(a)不同方法的性能对比　　(b)针对K个候选区域不同融合策略的mAP

图 3.4 Instre 中不同方法的比较

Fig 3.4 Comparison of different methods in Instre

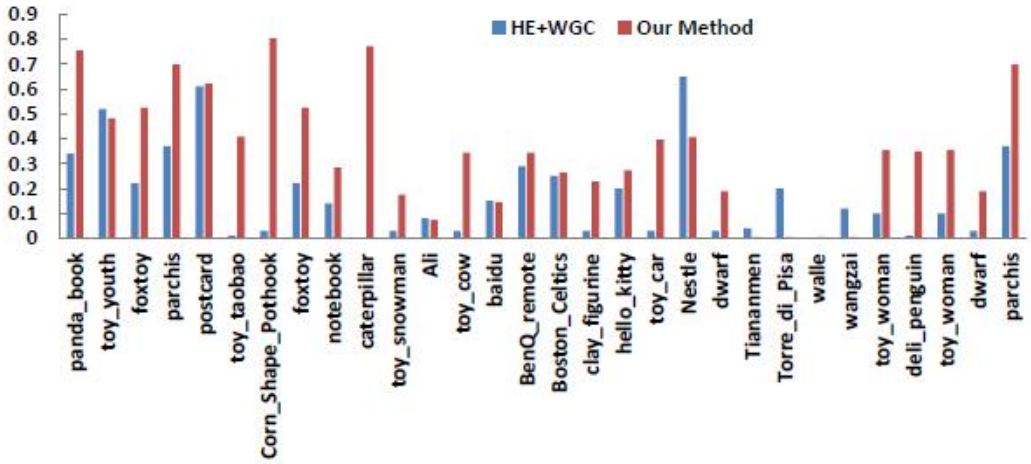


图 3.5 所选 30 个类别的 mAP 性能

Fig 3.5 Selected mAP Performance for 30 Categories

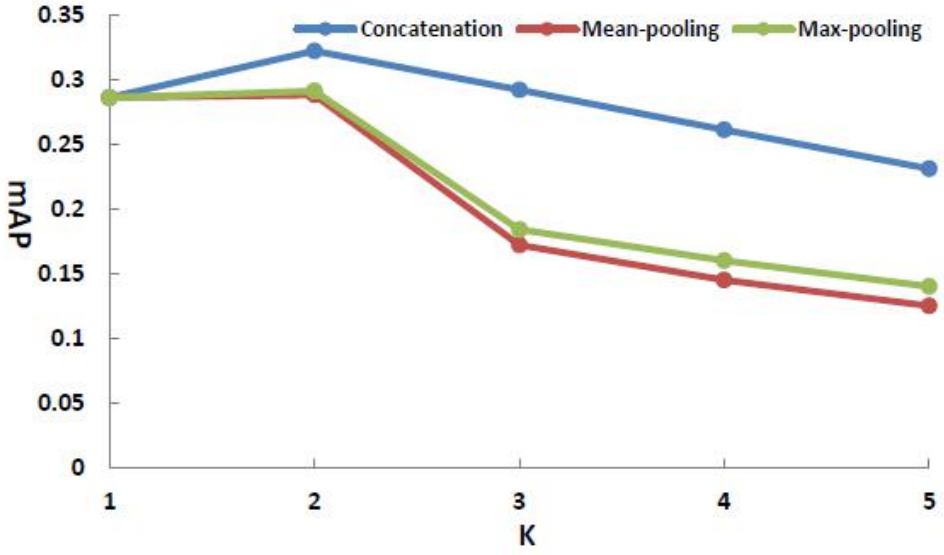


图 3.6 不同融合方法的 mAP 性能对比

Fig 3.6 mAP performance of different fusion method

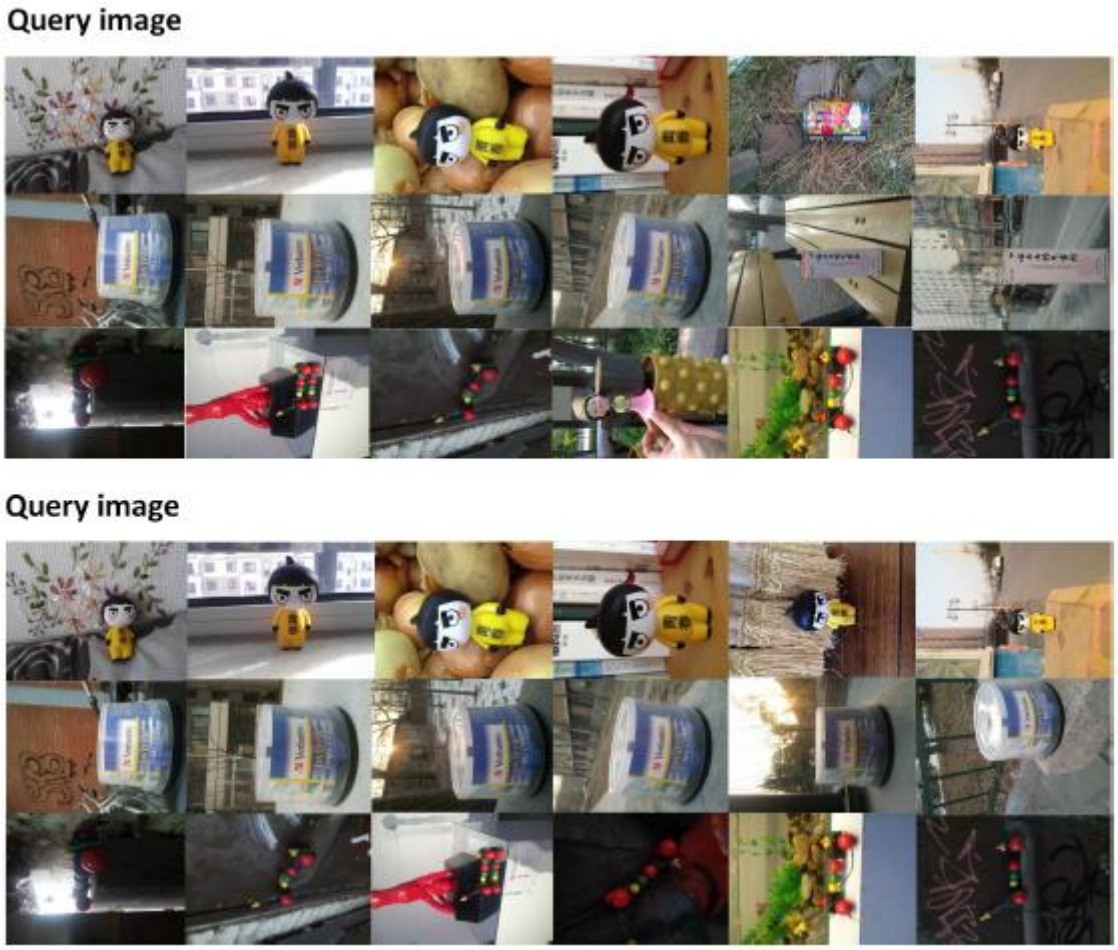


图 3.7 当 K = 1，2 和 3 时，图像检索示例。

Fig 3.7 Image retrieval example when K = 1, 2 and 3, respectively.

3.5 讨论

考虑到图像的背景噪声会影响实例级图像检索的性能。 为了减少背景的影响，首先检测物体区域，并直接从物体区域提取特征进行检索。 但是，并不能保证得分最高的地区是物体区域。 因此，通过结合得分最高的 K 个区域结果的特征来获得最佳性能。但是，区域数量和表现之间存在平衡。在实验中，当 K> 2 时，随着 K 的增加，检测到的边界框中包含更多的噪声。因此，导致其性能有所下降。 另一点要注意的是，在实验中，选择了 VGG16 网络来提取视觉特征，而不是使用 Faster R-CNN 的 ZF 网络。因为 VGG16 网络层数比 ZF 多，且具有更好的分类性能。在 Faster R-CNN 网络微调之后，使用的层数更多的 VGG16 网络通常比 ZF 网络更好，从而可以提高了实例级检索的性能。进行了实验。采 ZF 网络的检索结果为 26.9％，而采用的 VGG16 网络检索结果为 28.6％。与 ZF 网络相比，有大约 2％的改进。此外，设计了一种方法，利用区域视觉特征和预测的

类别特征相结合构成图像特征的方法，并用该图像特征进行检索，准确率为 32.2%，与 F-RCNN+CNN 方法相比，有大约 1%的提高。该实验验证了引入预测类别信息的有效性。

3.6 本章小结

本文提出了一种使用 CNN 特征的对象检测 CNN 的实例级对象检索方法。这个方法的创新点可以总结如下：

（1）本章提出了一种结合 Faster R-CNN 和 CNN 进行实例级对象检索的方法。

（2）对比了融合区域特征的不同策略，选出最优的策略。

（3）对两个不同数据集进行了实验，实验结果验证了方法的有效性。

与传统的基于 SIFT 的方法和基于 CNN 的全局特征提取方法相比，它具有提高性能的能力。 这项工作可以在以下三个方面扩展。

（1）使用本章现有的框架来实现一个图像中多个对象的实例检索。

（2）调整现有的实例检索框架。 例如，通过使用 R-MAC 架构动态选择物体区域的实例检索方法[29]。

（3）将本章的方法应用到不同的领域，例如实例级的食物检索和衣服检索。

4 用于图像检索的两阶段三连体网络

4.1 引 言

实例级图像检索旨在从大型无序图像集合中检索包含与查询图像相同的对象实例的所有图像。由上一章可以看出，如果简单采用先提取区域然后提取区域的特征的方式，在引入干扰数据集之后，检索性能大大降低。其中主要原因有以下两点：1.通过 Faster R-CNN 探测区域的准确率不是 100%，一旦出错会导致查询图片性能大大下降。2.使用微调后的训练模型作为通用特征提取器，用所提取的高层特作为用于图像检索的整体特征表示，但是这种特征不能区别未经训练的图片。

为了解决这类方法的不足，Tolias[20]等人提出了一种名为 R-MAC 的网络结构，它可以使用排名损失去学习一个更适合检索的整体表示。R-MAC 虽然可以改进现有的图像检索方法，但是这种选择检索区域的方法更加适合对图像整体的描述，不一定适用于实例级检索任务。因此，选择更好的实例物体所在区域是另一种选择。最近，Gordo 等人[30]提出了一种三连体网络，结合了三个分支来调整 CNN 以产生更好的特征表示。但是，他们没有充分利用监督信息，因此不能有效地利用不同类型的损失函数来更好地进行网络训练，特别是在更复杂的三连体网络中。多种类型的损失函数可以针对不同方面的特定检索任务约束神经网络的参数，以提高学习的深度网络的判别能力。例如，Softmax 损失函数可用于最小化所有训练样本的交叉熵损失。成对的排名损失可以解释所有训练图像的有序排名，以理解图像对之间的细微差异。以前的工作都没有对网络进行训练，特别是对于图像检索而言，具有不同类型损失的更复杂的三连体网络可以获得更好的检索性能。

本文针对基于 CNN 的实例级图像检索，通过引入分类子网络，提出了一种具有双损失正则化三连体网络，扩展了网络体系结构，该网络共同使用三连体排名损失和分类损失来进行调整 CNN。卷积层从这种微调中池化和聚合从而产生更多的判别性全局特征。

以前，有一系列池化策略可以使用。这些策略从最大池化、平均池化、区域池化到区域广义均值池化[31]。区域广义均值池化方法包含最大池化和平均池化，可以通过学习得到更适合数据集的池化参数从而提高检索性能。因此，为三连体网络引入区域广义均值池化层，不同于文献[31]，本文使用整个图像卷积层的空间做区域广义均值池化，具体步骤如下：

（1）利用广义均值池化层为来自区域提议的每个检测图像区域特征映射网络（RPN)

（2）将来自不同区域的特征映射合并到区域广义卷积激活（R-GAC）作为最终的图像表示。

由上述两个步骤的组合产生了一个新的框架，能够将一个图像编码成用于图像检索的全局特征表示。图 4.1 展示了整个框架的架构，主要由两个部分组成：双损失正则化三连体网络（DLRTN）和区域广义均值池化（RGMP），下面是更详细地介绍。

针对从区域提议网络（RPN）中检测到的候选区域，首先提出一个双损失正则化三连体网络（DLRTN），它通过附加排名子网络和分类子网络来扩展基本三连体网络。因此，DLRTN通过同时优化两种损失函数来进行训练。同时，在训练过程中不相关的区域被删除。作为副分支，额外的分类损失导致更快的收敛。考虑卷积层和DLRTN剩余区域的提取特征映射，然后引入三连体网络的区域广义均值池化（RGMP）层来进一步学习池化参数。通过RGMP，将每个区域的特征映射作为输入，并为每个区域生成一个合并特征向量，然后将不同区域的特征向量作为全局图像表示集合到区域广义卷积激活（R-GAC）。在测试阶段，将图像直接经过训练的框架，以生成全局图像表示R-GAC，该图像可以用点积与数据集图像进行比较。最后，对六个图像检索数据集进行了全面的实验，验证了框架的有效性。

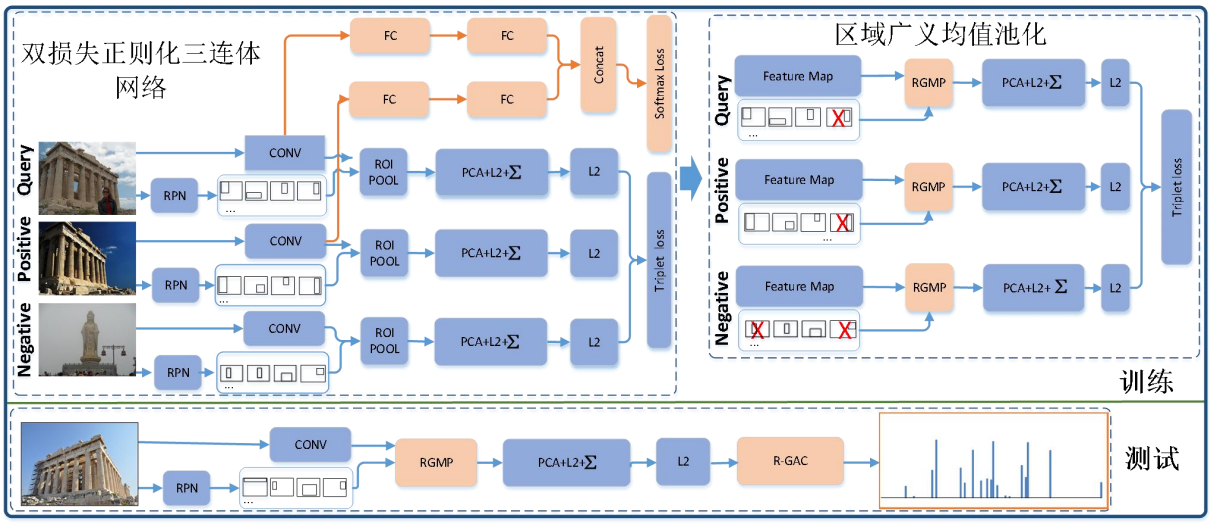


图 4.1 系统框架

Fig 4.1 The framework of system

4.2 方法框架

4.2.1 双损失正则化三连体网络

现有方法通常仅使用基于三连体网络的图像三元组的排名损失。 它明确规定，在给定查询的情况下，选择一个正例的图像和一个负例的图像，正例是一张比负例的更接近查询图像的图像。 令I_q为视觉表示q的查询图像，I^+为描述符d^+的相关图像，I^-为描述符d^-的非相关图像。排名损失定义为

$$L_r(I_q, I^+, I^-) = \max(0, m + \left\| q - d^+ \right\|^2 - \left\| q - d^- \right\|^2), \tag{4.1}$$

其中m是控制边距的参数。给定一个具有非零损失的三元组，梯度通过三连体网络的三个分支反向传播。

仅对排名损失进行训练可能不足以训练如此复杂的三连体组网络。 因此，通过附加分类子网络来微调这个基本的三连体组神经网络。将查询和正样本的特征提取层连接一个融合子网络，它由一组全连接的层和具有 Softmax 损失L_c的 Concat 层组成（图 4.1）。对于 Softmax 损失，第j类中第i个图像对$[q, d^+]$的归一化概率可以通过

$$p_{i,j} = \frac{\exp(Q_j([q, d^+]_i))}{\sum_{j=1}^{C} \exp(Q_j([q, d^+]_i))}, \tag{4.2}$$

其中$[q, d^+]$表示融合的子网络中q和d^+之间联合表示，$Q_j([q, d^+]_i)$是指在j类中，第i个图像对$[q, d^+]$的离散概率分布，C是可能的类别的数量。

Softmax 损失的形式是：

$$L_c = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \overline{p_{i,j}} \log(p_{i,j}), \tag{4.3}$$

其中N是查询样本及其正样本的图像对的数量。

为了训练三连体网络，需要将总损失函数最小化，这是排名损失和分类损失的加权组合，如下所示：

$$L = L_r(I_q, I^+, I^-) + \lambda L(I_q, I^+), \tag{4.4}$$

这里的λ是一个超参数。

本文的目标是结合不同区域的特征产生全局图像表示，如 R-MAC [2,48]中已经证明了这种方法对图像检索任务的有效性。因此，如图 4.1 左上部分也采用了类似的体系结构。所提出的三连体网络的每个分支还包含预先训练的网络（VGG16)和区域池化层（ROI），用于不同区域中特征的池化。将这些池化的区域特征归一化，再用 PCA 降维再归一化。然后 L_2 归一化以产生整张图像的特征向量。还使用区域提议网络（RPN）来提取图像中可能存在实例的区域。 RPN 的主要思想是通过分数预测每张图像中实例所在位置的可能性。通过 RPN，可以从每个图像中获得感兴趣的候选区域。

尽管 DLRTN 采用了文献[32]中提出的一些基本网络层次，但有两个重要差异，首先，与以前的三连体网络仅使用排序损失不同，增加了一个融合子网络利用排名损失和分类损失的联合优化。通过三连体网络的不同损失函数共同可以优化三连体网络的参数，并有助于提高其判别能力。此外，额外的分类损失可以加快收敛速度。其次，文献[32]的目标是直接学习视觉特征表达,而本章提出的网络主要侧重于卷积层的特征映射学习。同时，在训练过程中收集更加有挑战性的图像区域，并通过选择的图像区域和提取的特征用于进一步学习三连体网络。

4.2.2 区域广义均值池化

通过 DLRTN 从每个训练图像的检测区域提取特征映射。 然后，池化来自每个区域的特征映射，并通过区域广义均值池化（RGMP）将来自不同区域的特征映射池化为全局特征向量。

使用的是每个图像区域卷积层的特征映射 χ，其维度是 $W\times H\times K$，其中 K 是特征映射的数量。用一组 2D 张量来表示 3D 张量。 $\chi=\{\chi_{\mathrm{i}}\},\mathrm{i}=1,...,K$，其中 χ_i 是 2D 张量表示第 i 个特征信道对有效空间位置的集合的响应的 2D 张量。来自最后一个卷积层的特征比全连接层具有更好的检索性能，因此在本文中选择预训练网络的最后一个卷积层。

将来自每个区域的特征映射 χ 作为输入，并使用共用层来产生最终特征的 f 表示作为池化过程的输出。通常，这个特征向量是通过最大池化或平均池化来生成的：

$$f_{\Omega}^{m}=[f_{\Omega,1}^{m},...,f_{\Omega,i}^{m},...,f_{\Omega,K}^{m}],f_{\Omega,i}^{m}=\max_{p\in\Omega}\chi_{p}, \tag{4.5}$$

$$f_{\Omega}^{a}=[f_{\Omega,1}^{a},...,f_{\Omega,i}^{a},...,f_{\Omega,K}^{a}],f_{\Omega,i}^{a}=\frac{1}{\Omega}\sum_{p\in\Omega}\chi_{p}, \tag{4.6}$$

本文使用以下给出的区域广义均值池化（RGMP）去替代刚才的池化方法：

$$f_{\Omega}^{g}=[f_{\Omega,1}^{g},...,f_{\Omega,i}^{g},...,f_{\Omega,K}^{g}],f_{\Omega,i}^{g}=(\frac{1}{|\Omega|}\sum_{p\in\Omega}\chi_{p}^{\xi})^{\frac{1}{\xi}}, \tag{4.7}$$

该池化方法的输出为通用卷积激活（GAC），这与最大池中的最大卷积激活（MAC）类似。 池化操作参数ξ可以学习的，因为这个参数是可微的，并且可以通过反向传播来调整。 如图 4.1 的右上部分所示，输入是来自 DLRTN 最后一个卷积层的特征映射，仅用具有排名损失的三连体网络去学习参数ξ。

在学习了 RGMP 的三连体网络后，可以从每个选定区域获得 GAC，然后聚合一幅图像的所有区域的特征以获得区域广义卷积激活（R-GAC）。

4.2.3 生成用于图像检索的 R-GAC

在测试阶段，向预训练好的 DLRTN 发送一个图像，并通过预训练好的 DLRTN 从该图像中提取卷积层的特征映射。然后通过 RGMP 池化来自每个选定区域的特征映射。然后，分别通过 PCA 与L_2归一化和求和来结合来自不同区域的特征，以获得最终图像表示，即区域广义卷积激活（R-GAC）。最后通过点积来计算查询与数据集中图像之间的相似度并得到最后的检索结果。

4.3 实验评估

4.3.1 数据集

基于以下六个图像检索数据集来验证方法的有效性，包括 Oxford5k，Paris6k[33]，Oxford105k，Paris106k，Instre-S 和 Instre-S + Flickr1M。 前四个数据集是标准图像检索数据集，而最后两个数据集是最近推出的实例搜索数据集，它包含各种日常三维或平面物体，从建筑物到具有多种变化的徽标，例如不同比例、旋转和遮挡。有些物体只显示一小部分，使它们更具挑战性。

4.3.2 实验设置

对于三连体网络，选择图像三元组对于确保快速收敛有着至关重要的作用。 首先，在数据集中创建三元组$(q,m(q),N(q))$，其中q表示查询图像，$m(q)$是匹配查询图像的正例， $N(q)$是一组与查询不匹配的负例图像。这些三元组用于构成训练图像对，类似于，

应该选择容易区分的正例和容易混淆的负例。对于困难的正例，选择与正例相同类别并具有最大描述符距离的图像：

$$m(q)=\left\|\bar{f}(q)-\bar{i}(q)\right\|_{i\in M(q)}^{\max}, \tag{4.8}$$

其中 $M(q)$ 是数据集中与查询具有相同类标签的图像集合，$\bar{f}(q)$ 和 $\bar{i}(q)$ $\bar{i}(q)$ 表示图像特征。

同样，选择来自不同类别的容易混淆的负例，就是不同类别的图像中与查询特征距离最小的图像：

$$m_1(q)=\left\|\bar{f}(q)-\bar{j}(q)\right\|_{j\in \mathrm{N}(q)}^{\min}, \tag{4.9}$$

其中 $N(q)$ 是数据集中类别标签与查询不同的图像集合，$\bar{f}(q)$ 和 $\bar{j}(q)$ 表示图像特征。

生成三元组的过程如下：对于每个类，随机选择一个图像作为查询，然后根据公式找到一个正例图像，再根据公式 4.9 在每个剩余类别中选择负例图像。特别是对于 Oxford 和 Paris 数据集，使用官方提供的查询图像，并在标签为正常和一般的类别图像集合中为每个查询选择一个正例，然后从其余各个类别中选择一个负例。为了与以前的方法进行公平比较，在 Instre-S 数据集中使用所有图像作为查询图像。由于每个类别中的图像非常相似，因此在构建三元组的时候，从每个类别中选择一个图像作为构建三元组的查询。对于每个查询，选择一个正例，并在其他 199 个类别中每个类别选择一张作为反例。

通过上述方法，可以获得初始训练三元组集合，但是，在学习三连体网络时，应该根据实际情况考虑。类似于文献[30]，使用最近的训练好的网络检查三元组，每隔 5 轮去做一次检查，即计算这些三元组特征的损失，并选择具有非零损失的三元组。

框架在 Caffe 平台上实现。本文选择流行的架构 VGG16，该架构在 Imagenet 数据集上预先训练好，应用 adam 算法来训练网络。初始学习率为 $l_0=1\times10^{-6}$，并使用指数衰减为 $l_0\exp^{-0.1i}$ 的变化规则，动量是 0.9，衰减权重为 5×10^{-4}。所有训练图像的大小调整为 256×256。三元组的批量大小为 8，设置 margin = 0.75 和 $\lambda=1.5$。

4.4 评估指标

对所有数据集使用标准评估指标是 mAP。 正如在 Oxford 和 Paris 的标准做法一样，只使用标注为查询图片的图片来做查询，而对于 Instre，使用整个查询图像与其他标准结果进行公平比较。

4.4.1 评估提案框架

在本节中，将定量和定性的评估框架中每个组成部分的有效性。由于本章是对 Learned R-MAC 的神经结构进行扩展，所以将框架中的每个改进与文献[2]进行了比较。表 4.1 显示了实验结果，其中 L-R-MAC 表示学习的 R-MAC。从实验结果可以看出（1）DLRTN 的性能比 L-R-MAC 好。这是因为与 L-R-MAC 相比，DLRTN 进一步引入了分类子网。DLRTN 的两种不同类型的损失函数可以共同作用于 CNN 以提高其判别能力。（2）引入 RGMP 进一步提高了 DLRTN 的性能。验证了本章方法在利用图像中不同区域的广义均值池化的有效性。通过基于 RGMP 的网络训练，可以为训练数据集学习最佳融合参数。来自不同区域的卷积激活的更好的融合会得到更具有区别性的特征表示。在六个标准数据集中，DLRTN 平均提高了 1.5％左右。（3）双损失正则化与 RGMP 相辅相成，共同提高了图像检索的性能。从表 4.1 可以看出，在所有六个不同数据集上的方法一直比 L-R-MAC 的性能要高。平均而言，六个数据集有大约 2.5％的性能改善。另外，图 4.2 中给出检索结果的一些例子。

图 4.2 本章的方法（DLRTN + RGMP）和 L-R-MAC 检索图像的例子

Fig 4.2 Example of the top search image between our method (DLRTN + RGMP) and L-R-MAC

表 4.1 添加不同分支方法的的性能（mAP）比较（%）

Tab 4.1 Performance (mAP) Comparison of Adding Different Branching Methods (%)

Method	Dim.	Oxford5k	Oxford105k	Paris6k	Paris106k	INSTRE-S	INSTRE-S+1M
L-R-MAC[12]	512	83.1	78.6	89.1	79.7	75.6	32.8
DLRTN	512	84.2	79.8	90.8	80.7	77.2	33.6
DLRTN+RGMP	512	85.3	81.4	91.6	82.4	79.6	34.2

4.4.2 DLRTN 的收敛性分析。

额外的分类损失不仅有助于提高深度网络的判别能力，而且还会带来更快的收敛速度。为了探索所提出的 DLRTN 的收敛性，在图 4.3 分别给出了在所有三个数据集 Oxford5k，Paris6k 和 Instre-S 上的前 80 轮训练的 DLRTN 模型和 L-R-MAC 的训练排序损失。从图中可以看到 DLRTN 的收敛速度比 L-R-MAC 更快。在相同的迭代次数下，DLRTN 的损失比 L-R-MAC 小。

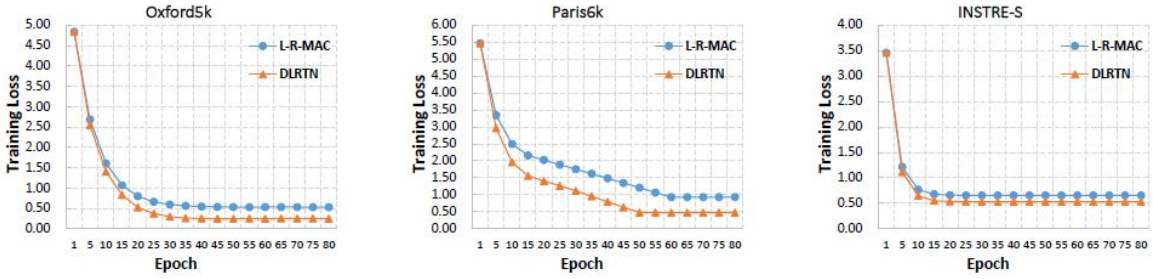


图 4.3 对于三个数据集不同迭代次数时的损失

Fig 4.3 Loss for different number of iterations for three data sets

4.4.3 RGMP 中ξ的影响

类似于文献[31]，在 CNN 与 RGMP 一起调整过程中的不同学习参数ξ与平均池化和最大池化进行了比较。图 4.4 中给出了实验结果，从图中可以看出，RGMP 层在三个数据集上始终优于传统的平均池化和最大池化，特别是在用 RGMP 对三连体网络进行精确调整之后，Oxford 的学习参数ξ为 2.72，Pairs 的学习参数ξ为 2.65，而 Instre-S 的学习参数ξ为 2.85。

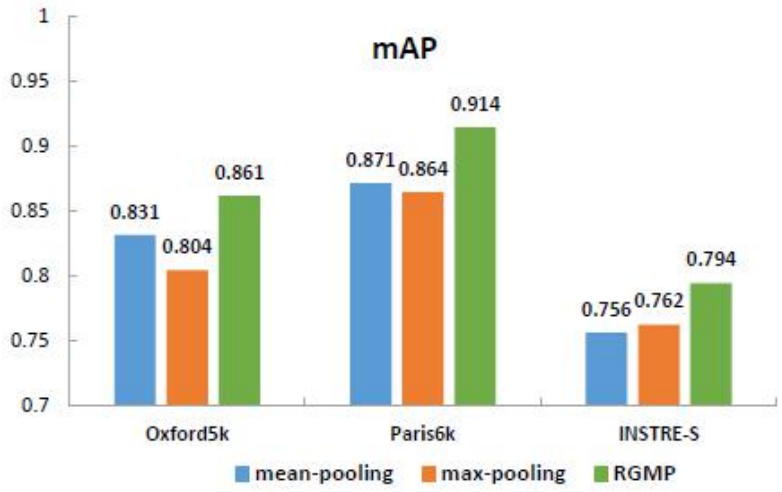


图 4.4 RGMP 与其他池化方法的比较

Fig 4.4 Comparison of RGMP and other pooling methods

4.4.4 与现有方法的比较

在本节中，将本章提出的框架与图像检索任务中的最新方法进行了全面的比较。首先将实验结果与当前的现有技术的性能进行比较。表 4.2 显示了在四个标准图像检索数据库上的实验结果并列出了最近提出了的一些方法，如 UFT-MAC [34]，L-R-MAC 和 Mask-MAC [35]。从表 4.2 可以看到，除了 Pairs106K 数据集上的 Mask-MAC 方法之外，本章的方法在所有标准数据集上实现了最佳性能，同时，本章的方法具有比其他方法更低的特征维度。表 4.3 显示了 Instre-S 的实验结果，可以看到，本章的方法始终超越所有这些方法。

然后，将本章的方法与增加了查询扩展(QE)的其他方法进行比较，其中 QE 是一种相对简单的策略，能够显著提高最终准确度。与文献[11]类似，用平均 QE [36]进行实验，其成本可以忽略不计(前 10 个返回结果被使用)，表 4.4 和表 4.5 显示了实验结果。从表 4.4 可以看出，本章的方法优于很多最新技术。本章方法的性能比最近提出的区域融合方法在数据集[37]Oxford5k 和 Oxford105k 上提高性能。其他两个数据集的更高性能来自其更加高维的特征。虽然他们在增加特征维度时的性能更高，但这种方法成本太高，因为他们需要大量的存储空间。表 4.5 还显示 Instre 数据集上的实验结果。实验设置与 Instre 数据集的不同：本章仅仅使用 Instre-S 数据集。

表 4.2 在 Oxford 和 Paris 数据库中与最先进的方法基于 mAP 的比较(%)

Tab 4.2 Based mAP comparison with state-of-the-art methods in the Oxford and Paris databases (%)

Method	Dim.	Oxford5k	Oxford105k	Paris6k	Paris106k
MAC [39]	512	56.4	47.8	72.3	58.0
Patch-CKN [32]	256K	56.5	-	-	-
FM-VLAD[30]	128	59.3	59.0	-	-
SPOC [3]	256	68.1	61.1	78.2	68.4
Crow [21]	512	70.8	65.3	79.7	72.2
R-MAC [48]	512	66.9	61.6	83.0	75.7
NetVLAD-Off-Shelf [1]	4096	66.6	-	77.4	-
NetVLAD-Finetune [1]	4096	71.6	-	79.7	-
Bow-CNN [29]	n/a	73.9	59.3	82.0	64.8
SIAM-FV [31]	512	81.5	76.6	82.4	-
UFT-MAC [36]	512	79.7	73.9	83.8	76.4
L-R-MAC [12]	512	83.1	78.6	87.1	79.7
Mask-MAC [15]	**4096**	83.8	80.6	88.3	**83.1**
Ours	**512**	**85.3**	**81.4**	**91.6**	82.4

表 4.3 在 Instre 数据库中与最先进的方法基于 mAP 的比较(%)

Tab 4.3 Based mAP comparison with state-of-the-art methods in the Instre database (%)

Method	Dim.	INSTRE-S	INSTRE-S+1M
Bow [46]	n/a	48.1	9.1
RANSAC [28]	n/a	46.5	7.7
SC [55]	n/a	53.4	12.5
GC [54]	n/a	58.6	19.7
COP [6]	2048	63.8	23.5
HE+WGC [18]	2048	67.2	26.8
L-R-MAC [12]	512	75.6	32.8
Ours	512	**79.6**	**34.2**

表 4.4：在 Oxford 和 Paris 数据库中与最先进的方法基于 mAP 的比较(增加查询扩展)(%)

Table 4.4:Based mAP comparison with state-of-the-art methods in the Oxford and Paris databases (increasing query expansion) (%)

Method	Dim.	Oxford5k	Oxford105k	Paris6k	Paris106k
BoW(1M)+QE [7]	n/a	82.7	76.7	80.5	71.0
CroW+QE [21]	512	72.2	67.8	85.5	79.7
R-MAC+AML+QE [48]	512	77.3	73.2	86.5	79.8
UFT-MAC [36]	512	85.0	81.8	86.5	78.8
L-R-MAC+QE [12]	512	89.1	87.3	91.2	86.8
Global diffusion [17]	512	85.7	82.7	94.1	92.5
Region diffusion [17]	5×512	91.5	84.7	**95.6**	**93.0**
Ours+QE	**512**	**91.8**	**85.7**	95.5	89.8

表 4.5：在 Instre 数据库中与最先进的方法基于 mAP 的比较(增加查询扩展)(%)

Tab 4.5: Based mAP comparison with state-of-the-art methods in the Instre database (increasing query expansion) (%)

Method	Dim.	INSTRE-S	INSTRE-S+1M
L-R-MAC+QE[12]	512	79.8	35.8
Global diffusion [17]	512	70.3	-
Region diffusion [17]	5×512	77.5	-
Ours+QE	512	82.4	38.2

4.5 本章小结

本章提出了一个双损失正则化三连体网络来学习更强大的图像检索视觉表示的方法。主要工作可以归纳如下：

(1) 提出了一种新的图像检索框架，它可以将三连体网络与不同类型的损失函数和区域广义均值池化方法相结合，以提高图像检索的性能。

(2) 提出了一个双损失正则化三重连体网络（DLRTN），它包含排名子网络和分类子网络，用于调整网络参数以进行图像检索。两种类型的损失函数同时通过反向传播。

(3) 在六个图像检索数据集进行了全面的实验，包括标准数据集（Oxford 和 Paris）和最近引入的 Instre 数据集。实验结果验证提出的框架的有效性[38]。

今后，计划将 RGMP 和双重损失结合在一个三连体网络中，并为该网络设计端到端的学习方法。另外，可以在框架中引入了最先进的哈希方法，例如 Deepbit [39]或宽松二进制自编码器[40,41]，以支持大规模实例级图像检索。

最后，最近提出的方法如文献[42]和文献[43]采用更复杂的神经结构（例如 ResNet [44]），新的多尺度图像表示和新的查询扩展方法[45,46,47]，以实现更好的图像检索性能。因此，可以将这些模块纳入本章的框架，以进一步提升性能。

5 总结与展望

5.1 总结

实例检索是多媒体领域的一个基本问题。与传统图像检索方法相比，它更具挑战性，这种挑战性主要体现在实例物体类别之间差距大（光照，旋转，遮挡，裁剪等），类类之间差别不大（可口可乐瓶与雪碧瓶），并含有大量的背景干扰信息。最近的进展表明，卷积神经网络（CNN）提供了一个更加优秀的图像特征表示方法。然而，单纯的使用 CNN 方法会从整个图像中提取特征，因此提取的特征中包含大量的背景噪声信息[48,49,50,51]，从而导致较差的检索性能。为了解决这个问题，本文提出了两种技术路线：

（1）一种基于 Faster R-CNN 检测的用于实例级图像检索的方法，它有两个阶段，即 Faster R-CNN 离线训练和在线实例检索。首先，训练 Faster R-CNN 模型以更好地定位物体的区域。其次，从检测到的物体图像区域中提取 CNN 特征，然后根据这些特征的视觉相似性检索相关图像。此外，利用三种不同的策略来基于检测到的来自 Faster R-CNN 的实例区域特征进行特征融合。

（2）一个新的实例级图像检索框架。与上一种方法相比，这种基于三连体网络的方法，可以同时通过分类损失和排序损失训练网络，为图像检索产生更具分辨度的全局特征表示。提出的框架主要由两个阶段组成。首先，提出了一个双重正则化三连体网络（DLRTN），它将来自经过预训练的区域提议网络（RPN）的检测区域作为输入，通过附加排名子网络和分类子网络。基于双重损失函数，通过同时优化两种损失函数和反向传播来训练 DLRTN。其次，接着引入三连体的区域广义均值池化（RGMP）层，以有效地池化来自 DLRTN 的每个区域的卷积激活，并汇总来自每个图像的不同区域的特征映射激活，产生区域广义卷积激活（R-GAC）作为全局图像表示[52,53,54]。

5.2 工作展望

虽然对图像检索的研究取得了一定的成果，但是在当前社会媒体数据爆炸的时代下需要我们亟待解决和深入研究不断出现的新问题。展望未来的工作，可以从以下两个方面继续探索和研究：

（1）可以将RGMP和双重损失结合在一个三连体网络中，并为该网络设计端到端的学习方法。此外，还可以在框架中引入了最先进的哈希方法以支持大规模图像检索。并且，目前还有更复杂的神经网络可以用于改进目前的方法[55,56,57]。

（2）通过以往的实验，通过Faster R-CNN或者双重损失的三连体网络都是通过探测目标区域的方法来构建对图像的描述，一旦探测结果不好，会使得性能大幅下降。同时，探测需要大量的人工标注信息。最近，一种显著性（attention）模型结构可以用于解决这个问题。因此，如何将显著性模型融入到本文的方法是一个值得研究的课题[58,59,60]。

参考文献

[1] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[2] Kalia R, Lee K D, Samir B V R, et al. An analysis of the effect of different image preprocessing techniques on the performance of SURF: Speeded Up Robust Features[C]. Frontiers of Computer Vision. IEEE, 2011:1-6.

[3] Lowe D G. Object recognition from local scale-invariant features[C]. The Proceedings of the Seventh IEEE International Conference on Computer Vision. IEEE, 2002:1150.

[4]Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutionalneural networks[C]. International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[5] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]. International Conference on Neural Information Processing Systems. MIT Press, 2015:91-99.

[6] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[7] Tolias G, Sicre R, Jégou H. Particular object retrieval with integral max-pooling of CNN activations[J]. Computer Science, 2015.

[8] Smeulders A W M, Worring M, Santini S, et al. Content-based image retrieval at the end of the early years[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2000, 22(12):1349-1380.

[9] Sivic J. Video Google : A Text Retrieval Apporach to Object Matching in Viedeos[C]. Proc. IEEE International Conference on Computer Vision. 2003.

[10] Razavian A S, Azizpour H, Sullivan J, et al. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society, 2014:512-519.

[11] Babenko A, Slesarev A, Chigorin A, et al. Neural Codes for Image Retrieval[C]. European Conference on Computer Vision. Springer, Cham, 2014:584-599.

[12] Ng Y H, Yang F, Davis L S. Exploiting local features from deep networks for image retrieval[J]. 2015:53-61.

[13] Jegou H, Douze M, Schmid C. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search[C]. European Conference on Computer Vision. Springer-Verlag,

2008:304-317.
[14] Jiang Y G, Wang J, Xue X, et al. Query-Adaptive Image Search With Hash Codes[J]. IEEE Transactions on Multimedia, 2013, 15(2):442-453.
[15] Sharma G, Schiele B. Scalable Nonlinear Embeddings for Semantic Category-Based Image Retrieval[C]. IEEE International Conference on Computer Vision. IEEE, 2015:1296-1304.
[16] Razavian A S, Sullivan J, Maki A, et al. A Baseline for Visual Instance Retrieval with Deep Convolutional Networks[J]. Computer Science, 2014.
[17] Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115(3):211-252.
[18] Donahue J, Jia Y, Vinyals O, et al. DeCAF: a deep convolutional activation feature for generic visual recognition[C]. International Conference on International Conference on Machine Learning. JMLR.org, 2014:I-647.
[19] Razavian A S, Azizpour H, Sullivan J, et al. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society, 2014:512-519.
[20] Gong Y, Wang L, Guo R, et al. Multi-scale Orderless Pooling of Deep Convolutional Activation Features[J]. 2014, 8695:392-407.
[21] Babenko A, Slesarev A, Chigorin A, et al. Neural Codes for Image Retrieval[C]. European Conference on Computer Vision. Springer, Cham, 2014:584-599.
[22] Kalantidis Y, Mellina C, Osindero S. Cross-Dimensional Weighting for Aggregated Deep Convolutional Features[C]. European Conference on Computer Vision. Springer, Cham, 2016:685-701.
[23] Yandex A B, Lempitsky V. Aggregating Local Deep Features for Image Retrieval[C]. IEEE International Conference on Computer Vision. IEEE, 2016:1269-1277.
[24] Philbin J, Chum O, Isard M, et al. Object retrieval with large vocabularies and fast spatial matching[C]. IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2007:1-8.
[25] Wang S, Jiang S. INSTRE: A New Benchmark for Instance-Level Object Retrieval and Recognition[J]. Acm Transactions on Multimedia Computing Communications & Applications, 2015, 11(3):1-21.
[26] Zhou W, Lu Y, Li H, et al. Spatial coding for large scale partial-duplicate web image search[C]. ACM International Conference on Multimedia. ACM, 2010:511-520.
[27] Zhou W, Li H, Lu Y, et al. SIFT match verification by geometric coding for large-scale partial-duplicate web image search[J]. Acm Transactions on Multimedia Computing

Communications & Applications, 2013, 9(1):4.
[28] Chu L, Jiang S, Wang S, et al. Robust Spatial Consistency Graph Model for Partial Duplicate Image Retrieval[J]. IEEE Transactions on Multimedia, 2013, 15(8):1982-1996.
[29] Panda J, Brown M S, Jawahar C V. Offline Mobile Instance Retrieval with a Small Memory Footprint[J]. 2013:1257-1264.
[30] Gordo A, Almazán J, Revaud J, et al. Deep Image Retrieval: Learning Global Representations for Image Search[C]. European Conference on Computer Vision. Springer, Cham, 2016:241-257.
[31] Radenović F, Tolias G, Chum O. Fine-tuning CNN Image Retrieval with No Human Annotation[J]. 2017.
[32] Azizpour H, Razavian A S, Sullivan J, et al. From generic to specific deep representations for visual recognition[C]. Computer Vision and Pattern Recognition Workshops. IEEE, 2015:36-45.
[33] Philbin J, Chum O, Isard M, et al. Lost in quantization: Improving particular object retrieval in large scale image databases[J]. Proc Cvpr, 2008:1-8.
[34] Radenović F, Tolias G, Chum O. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples[C]. European Conference on Computer Vision. Springer, Cham, 2016:3-20.
[35] Arandjelovic R, Gronat P, Torii A, et al. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition[C]. Computer Vision and Pattern Recognition. IEEE, 2016:1-1.
[36] Schonberger J L, Radenovic F, Chum O, et al. From single image query to detailed 3D reconstruction[C]. Computer Vision and Pattern Recognition. IEEE, 2015:5126-5134.
[37] Iscen A, Tolias G, Avrithis Y, et al. Efficient Diffusion on Region Manifolds: Recovering Small Objects with Compact CNN Representations[C]. IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2017:926-935.
[38] Lowe D G. Object recognition from local scale-invariant features[C]. The Proceedings of the Seventh IEEE International Conference on Computer Vision. IEEE, 2002:1150.
[39] Lin K, Lu J, Chen C S, et al. Learning Compact Binary Descriptors with Unsupervised Deep Neural Networks[C]. Computer Vision and Pattern Recognition. IEEE, 2016:1183-1192.
[40] Li Y, Kong X, Zheng L, et al. Exploiting Hierarchical Activations of Neural Network for Image Retrieval[C]. ACM on Multimedia Conference. ACM, 2016:132-136.
[41] Schonberger J L, Radenovic F, Chum O, et al. From single image query to detailed 3D reconstruction[C]. Computer Vision and Pattern Recognition. IEEE, 2015:5126-5134.
[42] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition

and clustering[C]. IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015:815-823.

[43] Shen X, Lin Z, Brandt J, et al. Spatially-Constrained Similarity Measurefor Large-Scale Object Retrieval[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 36(6):1229-1241.

[44] Bergamo A, Sinha S N, Torresani L. Leveraging Structure from Motion to Learn Discriminative Codebooks for Scalable Landmark Classification[C]. Computer Vision and Pattern Recognition. IEEE, 2013:763-770.

[45] Chum O, Mikulik A, Perdoch M, et al. Total recall II: Query expansion revisited[C]. IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2011:889-896.

[46] Do T T, Tan D K L, Pham T T, et al. Simultaneous Feature Aggregating and Hashing for Large-scale Image Search[J]. 2017: 411-423.

[47] Dubey A, Naik N, Parikh D, et al. Deep Learning the City: Quantifying Urban Perception at a Global Scale[J]. 2016:196-212.

[48] Gong Y, Wang L, Guo R, et al. Multi-scale Orderless Pooling of Deep Convolutional Activation Features[J]. 2014, 8695:392-407.

[49] Gordo A, Almazán J, Revaud J, et al. End-to-End Learning of Deep VisualRepresentations for Image Retrieval[J]. International Journal of Computer Vision, 2016:1-18.

[50] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2016:770-778.

[51] Hoang T, Do T T, Tan D K L, et al. Selective Deep Convolutional Features for Image Retrieval[C]. ACM, 2017:1600-1608.

[52] Ke Y, Sukthankar R. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors[C]. Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. IEEE, 2004:506-513.

[53] Kingma D, Ba J. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014:190-205.

[54] Ji X, Wang W, Zhang M, et al. Cross-Domain Image Retrieval with Attention Modeling[C]. ACM on Multimedia Conference. ACM, 2017.

[55] Noh H, Araujo A, Sim J, et al. Image Retrieval with Deep Local Features and Attention-based Keypoints[J]. 2016:180-194.

[56] 任夏荔，陈光喜，曹建收,等. 基于深度学习特征的图像检索方法[J]. 计算机工程与设计，2018(2).

[57] 张杰. 基于卷积神经网络的图像检索[J]. 卷宗，2016(7):1-12.

[58] 姜磊，赵汉理，吴承文. 基于卷积神经网络的鞋类图像检索研究[J]. 现代计算机，2016(6):39-43.

[59] 郑启财. 基于深度学习的图像检索技术的研究[D]. 福建师范大学，2015.

[60] 万吉. 基于深度学习的大规模图像检索技术研究[D]. 中国科学院大学，2016.

硕士阶段主要成果

[1] 梅舒欢，闵巍庆，刘林虎,段华等.基于 Faster R-CNN 的食品图像检索和分类[J]. 南京信息工程大学学报(自然科学版), 2017, 9(6):635-641.

[2] Shuhuan Mei, Weiqing Min, Hua Duan, Shuqiang Jiang.A Deep Region CNN Method with Object Detection for Instance-Level ObjectRetrieval[C].China Multimedia,2017.

[3] Shuhuan Mei, Weiqing Min, Hua Duan, Shuqiang Jiang.Instance-Level Object Retrieval via Deep Region CNN[J]. Multimedia Tools and Applications 2018(accepted).

[4] Shuhuan Mei,Weiqing Min,Hua Duan,Shuqiang Jiang. A Two-Stage Triplet Network Training Framework for Image Retrieval. Submit to ACM MM2018.

参与的工作：

[1] Weiqing Min, Bing-Kun Bao, Shuhuan Mei. Yaohui Zhu, Yong Rui, Shuqiang Jiang. You Are What You Eat:Exploring Rich Recipe information for Cross-Region Food Analysis. IEEE Trans. Mutimedia 20(4):950-964(2018).

[2] Shuqiang Jiang, Weiqing Min, Shuhuan Mei.Hierarchy-Dependent Cross-Platform Muti-View Feature Learning for Venue Category Prediction submit to TMM2018.

致 谢

行文至此，意味着我的硕士生涯已至谢幕时刻。回首三年的研究生生活，百感交集，苦辣酸甜集结于心头，但心中充盈最多的仍是感激之情。

首先要感谢的是我的导师段华副教授。段老师不仅给了我在山东科技大学读研的机会，还让我有机会参与山东科技大学与中国科学院计算技术研究所联合培养。在我研究生期间，段老师不仅传授了我做学问的技巧，还传授了我做人的准则，这些必将让我受益终身。

其次衷心感谢蒋树强研究员。没有他的悉心教导和帮助，就不会有我今天的成果。他严谨的科学态度，精益求精的工作作风，诲人不倦的高尚师德，严以律己、宽以待人的崇高风范，朴实无华、平易近人的人格魅力深深地感染和激励着我。在此谨向蒋老师致以诚挚的谢意和崇高的敬意。

另外，我还要衷心感谢闵巍庆师兄。在实验室的学习生活中，闵师兄教会了我科学的研究方法和准则，在我遇到科研方面的困难时，会不断鞭策我，并提出宝贵的意见和建议，循序渐进的引导我找到正确的解决方法。

感谢实验室的宋新航师兄、黎向阳师兄、朱耀辉师兄、吕雄师兄、王华阳师兄、乔雷先师兄、李雪师姐、朱永清、刘培培、陈恭巍、陈程鹏、刘林虎、罗正东、梁思思、吕永强等同学。感谢在这里没有一一列举的同学和朋友。感谢你们在平时实验和学术研讨方面给予的热情帮助。另外还要感谢实验室秘书王晓彪老师对于平时工作和生活的关心和帮助。

此外，我要感谢我的父母以及女朋友，真挚的感谢他们对于我学习生涯的理解和支持，以及生活上对我给予的照顾。没有他们背后的付出，我也不会安心完成研究生的学业。