 密级：

硕 士 学 位 论 文



论文题目 基于目标对筛选和联合谓语识别的

视觉关系检测

作者姓名

指导教师

学科(专业) 计算机科学与技术

所在学院 计算机科学与技术学院

提交日期

A Dissertation Submitted to Zhejiang University for the Degree of

Master of Engineering



TITLE: Visual relationship detection based on object pair sifting and joint predicate recognition

Author:

Supervisor:

Subject: Computer Science and Technology

College: Computer Science and Technology

Submitted Date:

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 **浙江大学** 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名： 签字日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解 **浙江大学** 有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 **浙江大学** 可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名： 导师签名：

签字日期： 年 月 日 签字日期： 年 月 日

学位论文作者毕业后去向：

工作单位： 电话：

通讯地址： 邮编：

摘要

随着深度学习尤其是卷积神经网络的飞速发展，计算机视觉在最近几年成为了十分活跃的研究领域。在目标检测与识别算法已经相对比较成熟的今天，进一步对图像内容进行理解和探索成了大势所趋，视觉关系检测应运而生。视觉关系检测的任务是从图片中找出用<主语-谓语-宾语>三元组表示的视觉关系，并标出其相应的位置。

由于可能组成的关系数量十分巨大，因此如何从有限的样本中学习出大量可能出现的关系成为了该任务的难点。近几年来，学者们围绕这个问题展开了许多研究，包括基于语言先验的方法、基于知识表征学习的方法和基于统计依赖的方法等。本文考虑到同种谓语在不同场景下可能表现差异较大的情况，以及一幅图中可以组成的目标对数量可能远大于实际存在关系的目标对数量的现象，分别基于视觉关系检测的不同子任务，提出了基于多特征联合统计预测的谓语识别算法及基于目标对筛选和联合谓语识别的视觉关系检测算法，并展开研究，本文的主要工作包括：

1）提出一种基于多特征联合统计预测的谓语识别算法

针对已知主语和宾语标签与位置情况下的谓语识别任务，考虑到同种谓语的表现差异可能较大的问题，本文提出了一种基于多特征联合统计预测的谓语识别算法。该算法同时考虑了多种信息来共同进行预测，融合了视觉特征、位置特征和语义特征得到联合特征，再结合统计依赖关系进行联合预测。在进行联合统计预测时，现有的基于统计的方法没有考虑到不同统计分量的重要程度，因此本文对不同分量的重要程度做出区分，把不同的分量作为不同的通道，并利用1x1卷积的通道融合功能将该思想融入到网络中。本文还使用了交叉熵损失和中心损失联合训练的方法，旨在使不同谓语之间的联合特征差异更大，而同种谓语的联合特征更加相似。通过在公开数据集上的实验和对比，结果表明该方法的Recall@100指标可以达到85.82%，相较于现有的DR-Net算法有3.92%的提升，说明该方法可以得到更好的谓语识别结果。

2）提出了一种基于目标对筛选和联合谓语识别的视觉关系检测算法

针对视觉关系检测中的短语检测和关系检测任务，考虑到图片中检测出的目标之间可能很多都不存在关系的现象，本文提出了一种基于目标对筛选和联合谓语识别的视觉关系检测算法，把视觉关系检测分为三个阶段——目标检测、目标对筛选和谓语识别阶段。先对输入的图片进行目标检测，再使用目标对筛选模型判断两个目标之间是否可能存在关系，目标对筛选模型主要从目标对的相对位置信息和语义信息两个方面出发来进行筛选。最后再利用之前提出的谓语识别算法对有关系的目标对进行谓语识别，以判断出它们之间是什么关系。本文首先对目标对筛选模型进行实验和评估，再对整体的视觉关系检测算法进行实验，并在视觉关系检测公开数据集上进行测试，通过与现有方法的比较发现该算法可以取得一定的效果提升。

**关键词：**视觉关系检测，谓语识别，统计预测，特征中心化，中心损失

Abstract

With the rapid development of deep learning, especially convolutional neural networks, problems related to computer vision have become very active research fields in recent years. Nowadays, the object detection and recognition algorithm has been relatively mature. Further understanding and exploration of image content has become the trend of the times, and visual relationship detection has emerged. The task of visual relationship detection is to find the visual relationship represented by the <subject-predicate-object> triplet from the picture and mark its corresponding position.

Since the number of possible relationships is very large, how to learn a large number of possible relationships from a limited sample becomes a difficult task. In recent years, scholars have carried out a lot of research on this issue, including methods based on language prior, methods based on knowledge representation learning and methods based on statistical dependence. This thesis considers that the same predicate may vary greatly in different scenarios, and the phenomenon that the number of object pairs in a graph may be much larger than the number of object pairs in the actual existence relationship. For the different subtasks of visual relationship detection, this thesis proposes a predicate recognition algorithm based on multi-feature joint statistical prediction and a visual relationship detection algorithm based on object pair filtering and joint predicate recognition, respectively. The main work of this thesis includes:

1) A predicate recognition algorithm based on multi-feature joint statistical prediction

Aiming at the predicate recognition task in the case of known subject and object labels and positions, considering the problem that the performance difference of the same predicate may be large, this thesis proposes a predicate recognition algorithm based on multi-feature joint statistical prediction. The algorithm considers a variety of information to jointly predict, combines visual features, location features and semantic features to obtain joint features, and then combines statistical dependencies for joint prediction. In the joint statistical forecasting, the existing statistical-based method does not take into account the importance of different statistical components. Therefore, this thesis makes a distinction between the importance degree of different components, uses different components as different channels, and uses 1x1 convolution, which can make channel fusion to incorporate this idea into the network. This thesis also uses the method of joint training of cross entropy loss and center loss, which aims to make the difference of joint features between different predicates larger, and the joint features of the same predicate are more similar. Through experiments and comparisons on the public dataset, the results show that the Recall@100 index of the method can reach 85.82%, which is 3.92% higher than the existing DR-Net algorithm, indicating that the method can get better result of predicate recognition task.

2) A visual relationship detection algorithm based on object pair sifting and joint predicate recognition

For the phrase detection and relationship detection tasks in visual relationship detection, considering the phenomenon that there is no relationships between many detected object pairs in the picture, this thesis proposes a visual relationship detection algorithm based on object pair sifting and joint predicate recognition. The visual relationship detection is divided into three stages - object detection, object pair sifting and predicate recognition stage. Firstly, the objects in image are detected, and then the pair sifting model is used to determine whether there is a possible relationship between two objects. The pair sifting model is mainly based on the relative position information and semantic information of the object pair. Finally, the proposed predicate recognition algorithm is used to identify the predicate of related pairs to determine the relationship between them. This thesis firstly conducts experiments and evaluations on the pair sifting model, and then tests the overall visual relationship detection algorithm and tests it on the visual relationship detection public data set. By comparing with the existing methods, it can be found that the algorithm can achieve certain effect improvement.

**Keywords：**visual relationship detection, predicate recognition, statistical judgment, feature centralization, center loss

目录

[摘要 i](#_Toc535186476)

[Abstract I](#_Toc535186477)

[第1章 绪论 1](#_Toc535186478)

[1.1 研究背景与意义 1](#_Toc535186479)

[1.1.1 课题研究背景 2](#_Toc535186480)

[1.1.2 课题研究意义 4](#_Toc535186481)

[1.1.3 课题研究难点 5](#_Toc535186482)

[1.2 国内外研究现状 6](#_Toc535186483)

[1.3 本文主要工作 9](#_Toc535186484)

[1.4 本文的组织结构 10](#_Toc535186485)

[第2章 相关工作 12](#_Toc535186486)

[2.1 引言 12](#_Toc535186487)

[2.2 视觉关系检测算法 12](#_Toc535186488)

[2.2.1 基于语言先验的视觉关系检测 12](#_Toc535186489)

[2.2.2 基于知识表征学习的视觉关系检测 13](#_Toc535186490)

[2.2.3 基于统计依赖的视觉关系检测 14](#_Toc535186491)

[2.3 目标检测算法 15](#_Toc535186492)

[2.3.1 基于区域建议的方法 15](#_Toc535186493)

[2.3.2 无区域建议的方法 17](#_Toc535186494)

[2.4 语义嵌入 18](#_Toc535186495)

[2.4.1 词向量 18](#_Toc535186496)

[2.4.2 Word2vec 19](#_Toc535186497)

[2.5 本章小结 19](#_Toc535186498)

[第3章 基于多特征联合统计预测的谓语识别 21](#_Toc535186499)

[3.1 研究问题 21](#_Toc535186500)

[3.2 基于多特征联合统计预测的谓语识别算法 22](#_Toc535186501)

[3.2.1 整体网络框架 22](#_Toc535186502)

[3.2.2 特征模块介绍 23](#_Toc535186503)

[3.2.3 联合统计预测 28](#_Toc535186504)

[3.2.4 损失函数 33](#_Toc535186505)

[3.3 实验与分析 35](#_Toc535186506)

[3.3.1 数据集 35](#_Toc535186507)

[3.3.2 评价指标 37](#_Toc535186508)

[3.3.3 实验设置 38](#_Toc535186509)

[3.3.4 实验结果分析 39](#_Toc535186510)

[3.4 本章小结 41](#_Toc535186511)

[第4章 基于目标对筛选和联合谓语识别的视觉关系检测 43](#_Toc535186512)

[4.1 研究问题 43](#_Toc535186513)

[4.2 基于目标对筛选和联合谓语识别的视觉关系检测 44](#_Toc535186514)

[4.2.1 算法框架 44](#_Toc535186515)

[4.2.2 目标对筛选模型 46](#_Toc535186516)

[4.3 实验与分析 48](#_Toc535186517)

[4.3.1 数据集 48](#_Toc535186518)

[4.3.2 评价指标 48](#_Toc535186519)

[4.3.3 实验设置 49](#_Toc535186520)

[4.3.4 结果分析 50](#_Toc535186521)

[4.4 本章小结 52](#_Toc535186522)

[第5章 总结与展望 54](#_Toc535186523)

[5.1 本文工作总结 54](#_Toc535186524)

[5.2 未来工作展望 55](#_Toc535186525)

[参考文献 56](#_Toc535186526)

图目录

[图 1.1 视觉关系样例 3](#_Toc535186527)

[图 1.2 根据视觉关系进行图像描述和视觉问答 4](#_Toc535186528)

[图 1.3 根据视觉关系构造知识图谱 5](#_Toc535186529)

[图 1.4 根据视觉关系进行图像检索 5](#_Toc535186530)

[图 1.5 视觉关系中的同谓语不同表现和同目标对间的不同谓语表达 6](#_Toc535186531)

[图 2.1 基于语言先验的视觉关系检测算法框架 13](#_Toc535186532)

[图 2.2 VTransE网络结构图 13](#_Toc535186533)

[图 2.3 DR-Net网络结构图 14](#_Toc535186534)

[图 2.4 R-CNN算法流程图 15](#_Toc535186535)

[图 2.5 Fast R-CNN算法流程图 16](#_Toc535186536)

[图 2.6 RPN网络基本结构示意图 17](#_Toc535186537)

[图 2.7 word2vec例图 19](#_Toc535186538)

[图 3.1基于多特征联合统计预测的谓语识别网络框架 23](#_Toc535186539)

[图 3.2 VGG16网络结构图 24](#_Toc535186540)

[图 3.3 主语图像，宾语图像和联合区域图像 24](#_Toc535186541)

[图 3.4 “人骑着马”和“人在马旁边”对比图 25](#_Toc535186542)

[图 3.5 位置模块结构图 26](#_Toc535186543)

[图 3.6 语义模块结构示意图 28](#_Toc535186544)

[图 3.7 ResNet的BottleNeck结构 31](#_Toc535186545)

[图 3.8 利用1×1卷积进行通道融合 32](#_Toc535186546)

[图 3.9 联合统计预测模块结构图 32](#_Toc535186547)

[图 3.10 中心损失函数效果示意图 34](#_Toc535186548)

[图 4.1 基于目标对筛选和联合谓语识别的视觉关系检测整体算法框架 45](#_Toc535186549)

[图 4.2 目标对筛选模型结构示意图 47](#_Toc535186550)

表目录

[表 2.1 One-hot词向量实例 18](#_Toc535186551)

[表 3.1 person和table的One-Hot表示 27](#_Toc535186552)

[表 3.2 person和table的word2vec词向量表示 28](#_Toc535186553)

[表 3.3 VRD数据集数据描述 36](#_Toc535186554)

[表 3.4 VRD数据集中的目标样例 36](#_Toc535186555)

[表 3.5 VRD数据集谓语举例 37](#_Toc535186556)

[表 3.6 本文谓语识别方法与现有算法在VRD数据集上效果对比 40](#_Toc535186557)

[表 3.7 本文谓语识别算法不同模块比较 41](#_Toc535186558)

[表 4.1 实际分类和预测分类的四种情况 48](#_Toc535186559)

[表 4.2 目标对筛选模型训练结果 50](#_Toc535186560)

[表 4.3 本文视觉关系检测算法与现有算法在VRD数据集上对比 52](#_Toc535186561)

# 绪论

* 1. 研究背景与意义

计算机视觉技术（Computer Vision）又名机器视觉，从首次提出至今，已有几十年的发展历程。近年来，随着机器学习尤其是深度学习的不断发展和阶段性的成功，计算机视觉已成为了非常活跃的研究领域，每年国内外都有大量关于计算机视觉的研究成果公诸于世。

计算机视觉是一个跨学科领域，从人工智能领域的角度来说，计算机视觉可以被看作是“赋予机器自然视觉能力”的科学，涉及如何使计算机从数字图像或视频中获得高层次的理解[1]。从信息处理领域的角度来说，计算机视觉则包括图像获取、图像处理、图像分析等技术，是一种从现实世界中提取高维数据以产生数字或符号信息的方法[2]，图像数据可以采用多种形式，例如视频序列、来自多个相机的视图或来自医学扫描仪的多维数据等。

综合来说，计算机视觉是以图像处理技术、计算机技术、统计分析、模式识别、神经网络乃至生理学、语言学等为基础，研究如何对输入的图像信息进行合理的提取组织、对物体或场景等进行识别、进而对图像的内容进行理解和解释的科学。常见的计算机视觉子领域包括图像分类、图像识别、图像分割、图像检索、场景理解、三维重建、视频追踪等内容。无论在学术领域还是在工程应用领域，计算机视觉都是一个富有价值和挑战性的重要方向。

传统的计算机视觉技术主要以传统的图像特征提取为基础，利用统计方法或机器学习方法来完成视觉任务。常见的图像特征包括颜色特征、形状特征、纹理特征等。颜色特征包括颜色直方图等；形状特征包括点特征、角特征、边缘特征和线条特征等[3]，常见的特征描述方法有方向梯度直方图（Histogram of Oriented Gradient, HOG），尺度不变特征变换（Scale-invariant feature transform, SIFT），以及傅里叶变换等；常见的纹理特征包括局部二值模式特征（Local Binary Pattern, LBP）和灰度共生矩阵（Gray-level Co-occurrence Matrix, GLCM）等。

随着深度学习尤其是卷积神经网络（Convolutional Neural Networks, CNN）技术的重大突破及GPU硬件的飞速发展，深度学习方法一跃成为计算机视觉技术中的主流方法。与传统方法不同的是，深度学习方法使用卷积神经网络自动从图像中提取特征，取代了传统的手工设计特征方法。

这一突破不但在计算机视觉的学术领域中大放异彩，也大力推动了计算机视觉技术在产业界的应用与发展，如安防领域的人脸识别、车牌识别、视频监控，智能与自动驾驶，医疗影像计算机辅助诊断，文字识别，手机人像美颜，图像及视频编辑、创作、风格化等。

* + 1. 课题研究背景

目标识别作为计算机视觉中一个重要研究内容，在过去的几年里取得了重大的突破，也促进了整个计算机视觉的大发展。从2012年AlexNet[4]在全球最权威的计算机视觉图像识别竞赛ILSVR（ImageNet Large Scale Visual Recognition Competition）上崭露头角开始，卷积神经网络在利用大数据进行监督学习的领域中展现出令人印象深刻的能力，这使计算机视觉的研究重心从传统方法转移到深度学习方法上。从2012年到2017年，不断有新的卷积网络被提出如VGG[5], GoogLeNet[6], ResNet[7]等，它们将图像识别的错误率从28%降低到了2.3%。在攻克图像识别的难题的征途上，计算机视觉的研究者们共同努力，取得了非凡的成绩。目标识别任务在现阶段已经达到很高的水平，甚至超越了人类的一般水平。

然而，目标识别仅仅是计算机视觉技术在“理解图像”道路上的第一步，想要让计算机真正做到对图像的进一步理解，仅仅识别出目标还远远不够，故继续进行图像之上的探索成了大势所趋。近两年来，对图像场景的进一步理解、图像与语言的结合等领域越来越受到人们的关注，而这也是搭建人类和计算机之间沟通桥梁的必经之路，对科学研究和工程实践都有重要的意义。

在对图像进一步理解的方向上，首先出现的任务是图像描述（Image Caption）[8]，即输入一张图像，根据算法自动生成对应的描述性文字。这类算法大多先使用卷积神经网络对图像提取特征，在使用循环神经网络（Recurrent Neural Network, RNN）生成描述的文字。

然而直接让计算机通过图像对图像内容做出描述却不是一个简单的任务，由于图像中目标之间关系的复杂性，很容易出现对目标之间关系描述不准确的情况。

于是“视觉关系”的概念渐渐被研究者们提炼出来，从最早的简单预测空间关系，如“above（上方）”、“below（下方）”、“inside（内部）”和“around（周围）”，到视觉短语（Visual Phrase）[9]，如“a person riding a horse”等。在2016年的计算机顶级会议——欧洲计算机视觉国际会议（European Conference on Computer Vision, ECCV）上，视觉关系检测任务被斯坦福大学的李飞飞团队正式定义[10]，于此同时，该团队也公开了他们为此任务构建的视觉关系数据集，加速了视觉关系检测任务的发展。

一般来说，图像中的一组视觉关系可以被定义为<object1 – predicate – object2> （<对象1 – 谓语 – 对象2>）的三元组形式，也可以被更准确的定义为<subject –predicate - object> [11]（即<主语 - 谓语 - 宾语>）的三元组。视觉关系检测任务需要找到图像中的这些三元组关系，并用方形边界框（Bounding Box, BBox）标出位置。如下图1.1所示[10]，这幅图中的对象包括“person（人）”，“motorcycle（摩托车）”，“helmet（头盔）”和“wheel（轮子）”，则它们之间可以包含3种关系，即“person on motorcycle（人在摩托车上）”, “person wear helmet（人戴着头盔）”和“motorcycle has wheel（摩托车有轮子）”。而人和轮子、摩托车和头盔、头盔和轮子之间则没有明显的关系。

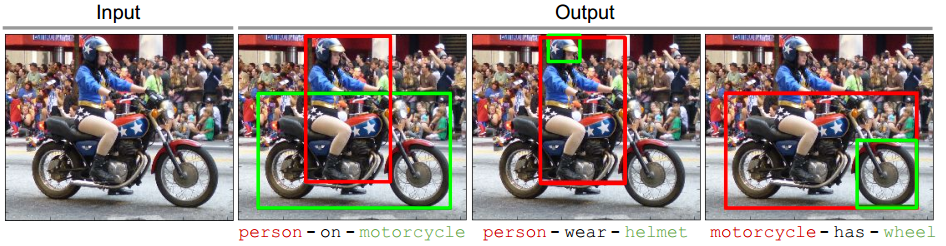


图 1.1 视觉关系样例

视觉关系检测被定义了三个子任务[10]，分别是：

1. 谓语识别任务（Predicate Detection），给定每对主语和宾语的标签以及Bounding Box（边界框），预测它们之间的谓语应该是什么。
2. 短语检测任务（Phrase Detection），输入一张图像，需在准确预测出三元组<subject - predicate - object>的同时，将关系三元组所在整体区域用一个Bounding Box标出来，并要求其与真实Bounding Box的重叠率（Intersection over Union, IoU）在50%以上。
3. 关系检测任务（Relationship Detection），与任务（2）相似，输入一张图像，需要准确预测出三元组<subject – predicate - object>，并要求分别找出主语与宾语的Bounding Box位置，并保证与真实Bounding Box的重叠率在50%以上。
   * 1. 课题研究意义

研究图像中的视觉关系，是图像理解的基础，是连接低层目标检测和高层场景理解之间的重要桥梁，是一种中层学习任务。如果已知图像中目标之间的关系，在进行图像描述时就可以更加准确、更加结构化，同时也可以根据视觉关系进行更加准确地视觉问答，如图1.2所示。

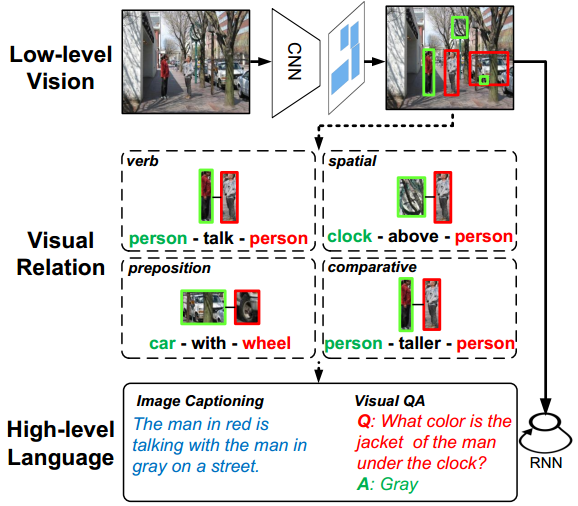


图 1.2 根据视觉关系进行图像描述和视觉问答

除了图像描述以外，视觉关系对知识图谱构建、知识问答和图像检索等也有很大的意义。比如当检测出了图像中所有目标之间的关系和属性，就可以构建这幅图的知识图谱，把图像中的信息结构化，如下图1.3所示[11]。在图像检索任务中，我们需要根据一张图片从大量的图片库中搜索与其相似的图片，如下图1.4所示[10]，假设提供的图片的内容是“人骑着马”，如果没有判断人和马之间是“骑”的关系，则很可能搜索出来的图片是“人在马旁边”、“人牵着马”、“人在马后面”等，而如果确定了人和马之间的关系是“骑”，搜索出来的图片则会更加符合。

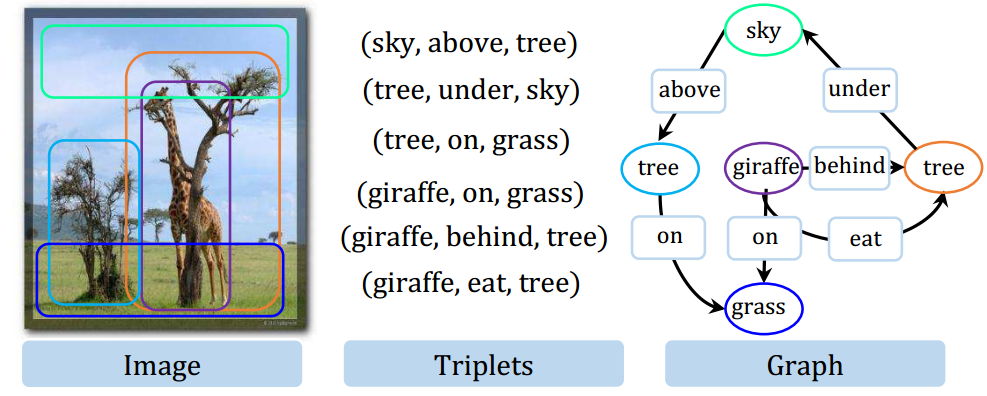


图 1.3 根据视觉关系构造知识图谱

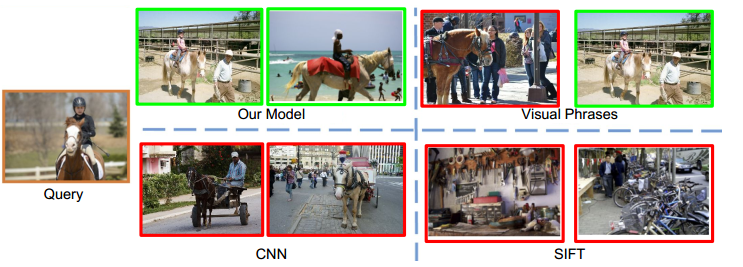


图 1.4 根据视觉关系进行图像检索

* + 1. 课题研究难点

视觉关系检测同样是一项富有挑战性的任务，它的难点可以总结为以下几点：

1. 在整个样本空间中，可能出现的关系非常多，假设object（对象）的种类数为*N*，predicate（谓语）的种类数为*K*，则可能出现的关系数量为，这么庞大的数量无疑对准确预测视觉关系带来了极大的挑战；
2. 同一个谓语的表现形式可能差异很大，如下图1.5中，同样是谓语“in”， “surfboard in hand（冲浪板在手上）”和“person in street（人在街道上）”却呈现了完全不同的表现形式，这使视觉关系的检测变得更加困难；
3. 同一对目标间可能存在多个合理的谓语，同样如图1.5，“person（人）”和“surfboard（冲浪板）”之间的关系可以用三个谓语来表示——“has”、 “hold”、“carry”都可以表达出“人拿着冲浪板”的意思，因此找出同一对目标间的不同谓语表达也成为了一个难点。
4. 一幅图中可以组成的目标对数量，可能远大于实际存在关系的目标对数量的现象。

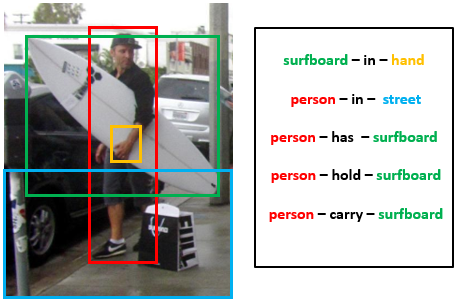


图 1.5 视觉关系中的同谓语不同表现和同目标对间的不同谓语表达

因此，想要视觉关系检测达到满意的效果，这些问题都需要被解决和克服，国内外的学者们也围绕这个任务进行了一系列的研究，提出了各自的解决方法。

* 1. 国内外研究现状

近几年来，有不少研究者们展开了在视觉关系领域的研究，并取得了一定的成果。

最早出现的视觉关系检测相关研究工作聚焦于预测目标之间的空间位置关系，Galleguillos等人在2008年提出了对“above（上方）”、“below（下方）”、“inside（内部）”和“around（周围）”四种空间关系的预测方法[12]，而Gould等人则在同年进一步探索这种空间关系预测并把它用于改善分割任务[13]。

而空间位置关系仅仅是众多关系中很小的一个分支，因此，研究者们开始关注与“人”相关的关系。Chao等[14]研究了人与其他目标之间的交互关系，即固定人为主语，需要找到“谓语-宾语”，如“ride – bicycle（骑自行车）”、“feed – giraffe（喂长颈鹿）”等。Yao等人[15]的工作则不但包括了与动作相关的谓语，还同时将介词纳入并加以区分，如“play – violin（弹奏小提琴）”和“with – violin（在小提琴旁边，但没有弹奏）”。而Maji等人[16]则更加聚焦于动作识别，即单纯对动作进行识别分类，如“running（跑）”、“phoning（打电话）”等。

2011年视觉短语（Visual Phrase）[17]被提出，此时关系的主语不再限制为“人”也可以是其他对象，算是真正意义上定义了“视觉关系”的概念。Visual Phrase方法把一组关系看作一个整体，如“a person riding a horse”等，预测时不对主语、谓语、宾语进行切分，且把每种关系当作一种分类。该工作使用的数据集中只包含了8种对象和17种常见的视觉短语，8种对象分别是——“person（人）”，“bike（自行车）”，“car（汽车）”，“dog（狗）”，“horse（马）”，“bottle（瓶子）”，“sofa（沙发）”，和“chair（椅子）”；17种视觉短语为——“person riding horse（人骑着马）”，“person sitting on sofa（人坐在沙发上）”，“person sitting on chair（人坐在椅子上）”，“person lying on sofa（人躺在沙发上）”，“person lying on beach（人躺在沙滩上）”，“person riding bicycle（人骑自行车）”，“horse and rider jumping（马和骑手跳跃）”，“person next to horse（人在马旁边）”，“person next to bicycle（人在自行车旁边）”，“bicycle next to car（自行车在车旁边）”，“person jumping（人在跳跃）”，“person next to car（人在车旁边）”，“dog lying on sofa（狗躺在沙发上）”，“dog running（狗在跑）”，“dog jumping（狗在跳）”，“person running（人在跑）”和“person drinking from a bottle（人从瓶子喝水）”。

然而，这种方法的局限性也很明显，假设object（对象）的种类数为*N*，predicate（谓语）的种类数为*K*，则可能出现的关系数量为，当对象种类数或谓语种类数很大时，这种方法将需要训练种分类器，这是一个相当大的数字。为了解决这个问题，研究者们的解决方案可以大致划分为两类，一类是把主、谓、宾分别进行预测，而不作为一个整体一起进行预测，这类方法是目前最主流的方法；另一类是依旧把视觉短语作为整体预测，但通过一些方法缩小可能出现的关系空间。

在第一类基于主谓宾分别预测的工作中，代表工作是Lu等人在2016年的ECCV上提出的基于语言先验的视觉关系检测方法[10]，事实上也是从这篇工作开始，视觉关系检测任务被正式提出，同时该工作的作者也公开了它们构建的视觉关系数据集（包含100种对象和70种谓语），使得视觉关系检测相关研究可以更快的发展。Lu等人先通过外观模块把对象（包括主语和宾语）和谓语分别作为单独的模块进行训练，并把训练结果组合在一起用于预测最后的关系。再进一步通过语言模块把关系映射到向量空间中，并利用语言先验知识来进行约束和优化，最终得到预测出的关系。这种方法的缺陷是，它将外观提取模块和最终的关系预测分了两个阶段来进行训练，并不是“端到端（end-to-end）”的训练方式。

随后，“端到端”的视觉关系训练方式在2017年被相继提出。Zhang等人在2017年的CVPR上提出了VTransE网络（Visual Translation Embedding Network）[18]，它通过将主语和宾语映射到共同的低维关系空间来对视觉关系建模，从而使谓语被认为是主语和宾语之间的翻译向量，即“主语+谓语≈宾语”。 并把知识表征学习的思想应用到其中，设计了一种全新的特征提取层，以全卷积的方式完成对象到关系的知识迁移。

Liang等人在2017提出了一个基于深度异结构强化学习框架的视觉关系及属性检测方法[19]，它通过利用全局上下文线索来顺序检测关系和属性实例。

Dai等人则从统计学的角度出发，同年在CVPR上提出了他们的DR-Net（Detecting Visual Relationships with Deep Relational Networks）[11]， 这是一种新的深度关系网络，基于谓语、主语和宾语之间的统计依赖关系，并把统计模型通过条件随机场理论和深度学习结合起来，大大提升了性能效果。

Liang和Guo等人着重考虑了同一组目标对之间可能出现多种合理的谓语的情况、以及真实标签中标注不完整的情况，在2018年的AAAI上提出了一种基于深度结构化排序的视觉关系检测方法[20]，充分利用图片中已标注出的关系，构造了结构化排序损失函数，使得图片中标注出的视觉关系的相关分数更高。

在第二类基于视觉短语整体的工作中，代表工作是Li等人在2017年提出的ViP-CNN（Visual Phrase Guided Convolutional Neural Network）[21]，它是一个视觉短语引导的卷积神经网络，利用整个视觉短语来进行学习，具体做法是通过消息传递同时学习三个相互连接的识别问题（主语-谓语-宾语），并利用该消息传递策略来减少参数数量。

上述工作虽然已经取得了不错的效果，但是仍然存在一些不足和局限性。

（1）上述工作没有很好的考虑到同一谓语在不同场景下的可能存在外观形式差异较大的问题，如上文所提到的，同样是谓语“in”， <surfboard – in – hand>（冲浪板在手上）和<person – in – street>（人在街道上）却呈现了完全不同的表现形式。

（2）基于谓语和主语、宾语统计依赖关系的算法，虽然考虑到了他们之间的统计依赖，却没有考虑到这种统计依赖相互之间的重要程度和所占的比重。

（3）上述方法大多是直接对目标对之间的谓语进行分类，没有先判断目标对之间是否会存在关系。

* 1. 本文主要工作

本文基于上文所提到的视觉关系检测任务的难点，以及现有工作的局限性，在前人工作的基础上，围绕1.1节中提到的视觉关系检测的三个子任务，首先针对任务一提出了一种基于多特征联合统计预测的谓语识别算法，再针对任务二和任务三提出了一种基于目标对筛选和联合谓语识别的视觉关系检测。同时对提出的算法进行了试验和分析，并与现有算法进行比较。研究的主要内容包括：

1. 基于多特征联合统计预测的谓语识别算法

针对已知主语和宾语标签与位置的情况下的谓语识别任务（任务一），考虑到同种谓语的表现差异可能较大的问题，本文提出了一种基于多特征联合统计预测的谓语识别算法。该算法同时考虑了多种信息来共同进行预测，融合了视觉信息、位置信息和语义信息共同得到联合特征，再结合统计依赖关系进行联合预测。

在进行联合统计预测时，现有的基于统计的方法没有考虑到不同统计分量的重要程度，因此本文对不同分量的重要程度做出区分，把不同的分量作为不同的通道，并利用1x1卷积的通道融合功能将该思想融入到网络中。本文还使用了交叉熵损失和中心损失联合训练的方法，旨在使不同谓语之间的特征差异更大，而同种谓语的特征更加相似，以缓解同种谓语在不同场景下外观形式差异较大的问题。通过在公开数据集上的实验和对比，结果表明该方法可以得到更好的谓语识别结果。

（2）基于目标对筛选和联合谓语识别的视觉关系检测

针对视觉关系检测中的短语检测任务（任务二）和关系检测任务（任务三），考虑到图片中检测出的目标之间可能很多都不存在关系的现象，本文提出了一种基于目标对筛选和联合谓语识别的视觉关系检测算法的算法，把视觉关系检测分为三个阶段——目标检测、目标对筛选和谓语识别阶段。先对输入的图片进行目标检测，再使用目标对筛选模型判断两个目标之间是否可能存在关系，最后再利用之前提出的基于多特征联合统计预测的谓语识别算法对有关系的目标对进行谓语识别，以判断出它们之间是什么关系。目标对筛选模型的设计主要用到了目标对的相对位置信息和语义信息。

本文首先对目标对筛选模型进行实验和评估，再对整体的视觉关系检测算法进行实验，并在视觉关系检测公开数据集上进行测试，通过与现有方法的比较可以发现该算法可以取得一定的效果提升。

* 1. 本文的组织结构

本文共分为五章，各章节的内容安排如下：

第一章是绪论部分。主要阐明了本文研究工作的相关背景和视觉关系检测任务的定义，并介绍了该任务的研究意义。然后对国内外的研究现状做了总结，分析了现有工作的存在的不足和局限性。最后介绍了本文的主要工作和组织结构。

第二章介绍了一些相关的研究工作。首先介绍了视觉关系检测任务的几个经典算法，包括基于语言先验的方法、基于知识表征学习的方法和基于统计依赖的方法等。接着，介绍了目标检测相关主流算法。最后介绍了语义嵌入和词向量的相关知识。

第三章研究了基于多特征联合统计预测的谓语识别算法。首先对研究问题进行了定义，然后详细介绍了基于多特征联合统计预测的谓语识别算法，包括整体的网络框架，每个子模块的功能和结构，损失函数的选用等。最后是实验部分，介绍了数据集和评价指标，并对实验结果进行分析、与其他算法进行比较。

第四章研究了基于目标对筛选和联合谓语识别的视觉关系检测。首先介绍了研究的问题，然后详细介绍了基于目标对筛选和联合谓语识别的视觉关系检测，包括整体算法流程，判断目标对之间是否存在关系的目标筛选模型的网络框架等。然后是实验部分，介绍了数据集和评价指标，并进行实验，最后分别对目标对筛选模型和视觉检测算法的结果进行分析，并对比了该算法与现有算法的效果。

第五章是总结与展望部分。主要回顾和总结了本文的研究工作，并对未来可以研究的内容进行了展望。

# 

# 相关工作

* 1. 引言

在目标检测与识别算法已经相对比较成熟的今天，进一步对图像内容进行理解和探索成了大势所趋，故视觉关系检测应运而生。近几年来，学者们围绕这个问题展开了许多研究，包括基于语言先验的方法、基于知识表征学习的方法和基于统计依赖的方法等，本章将对这些视觉关系检测算法作详细的介绍。

此外，本章还将介绍本文提出的算法中会涉及的目标检测算法及语义嵌入方法。

* 1. 视觉关系检测算法

近几年来，有不少研究者们对视觉关系检测算法展开了研究，并取得了一定成果。在这些工作中，最经典的三种算法是基于语言先验的视觉关系检测、基于知识表征学习的视觉关系检测和基于统计依赖的视觉关系检测，下面将对这三种算法进行详细介绍。

* + 1. 基于语言先验的视觉关系检测

基于语言先验的视觉关系检测算法[10]由Lu等人在2016年提出，正是在这个工作中，视觉关系检测的任务被正式定义，同时一个包含100种对象和70种谓语的视觉关系数据集也被公开，使得视觉关系检测相关研究可以更快的发展。

相较于传统的把<主语-谓语-宾语>三元组关系视作一个整体进行训练的方法，该算法做了一个改进，即把目标（包括主语和宾语）和谓语拆分开来单独训练。这么做的原因是，可能组成的三元组关系数量十分巨大，且很多关系在数据集中出现的很少，而目标和谓语分别出现的情况却更加频繁。

该算法的整体算法框架如图2.1所示，先对图片进行目标检测，产生目标的集合，并组成目标对。接着，每对目标分别通过视觉模块和语言模块，并结合这两模块的结果，最终得到可能组成的视觉关系的分数。

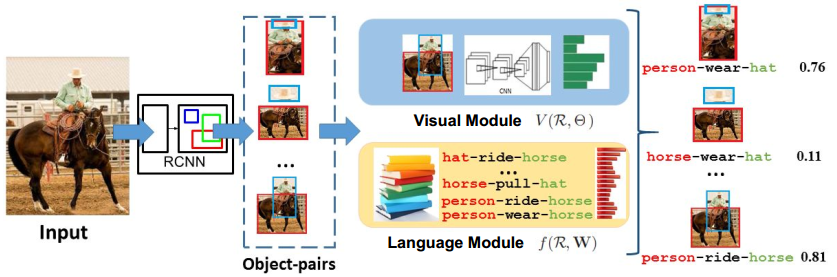


图 2.1 基于语言先验的视觉关系检测算法框架

在视觉模块中，训练了两种分类器，分别对目标和谓语进行分类，并把它们的结果融合到一起。在语言模块中，使用了语言先验把主语和宾语映射到词向量空间中，再把这两个词向量结合起来映射到关系向量空间，并使相似的关系在向量空间中尽量相似，利用语言先验知识来进行约束和优化。最后使用了一个排序损失函数来合并视觉模块和语言模块的结果，得到预测的每种关系的最终分数。

该算法相较于传统的基于视觉短语的方法有了巨大的提升，将任务一的谓语识别任务提升了45%，而任务二的短语检测则从0.07%提升到了17.03%。

* + 1. 基于知识表征学习的视觉关系检测

继视觉关系数据集在2016年被公开之后，2017年有更多的学者投入到视觉关系检测的研究中，一种基于知识表征学习的视觉转换嵌入网络（Visual Translation Embedding Network，VTransE）[18]即在2017年的CVPR会议上被Zhang等人提出。

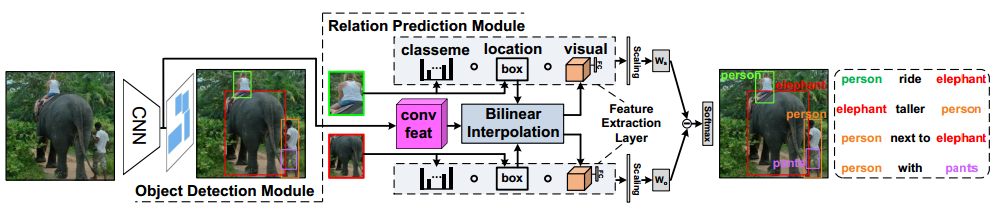


图 2.2 VTransE网络结构图

作者受到知识表征学习的启发，将主语和宾语映射到共同的低维关系空间来对视觉关系建模，从而使谓语被认为是主语和宾语之间的翻译向量，即“主语+谓语≈宾语”。并把目标和关系之间知识转换结合起来，设计了一种全新的特征提取层，提取了转换嵌入中的三种对象特征，分类、位置和ROI视觉特征，以全卷积的方式完成目标和关系的知识转换。VTransE的网络结构如图2.2所示。

该方法相较于Lu等人基于语言先验的算法有一定的提升，在短语检测任务和关系检测任务上Recall@100分别有5.39%和0.5%的提升。

* + 1. 基于统计依赖的视觉关系检测

同样在2017年的CVPR上被提出的，还有Dai等人的基于统计依赖的深度关系网络（Deep Relational Networks, DR-Net）[11]。

该网络从统计学的角度出发，认为谓语、主语和宾语之间是有统计依赖关系的，比如“person ride horse（人骑着马）”是一个很常见且合理的关系，但是“horse ride person（马骑着人）”、“person wear horse（人穿着马）”或“person ride mouse（人骑着老鼠）”则并不容易出现，因此可以看出“ride - horse”间有强烈的依赖关系，而“horse - ride”、“ride - person”、“wear - horse”和“ride - mouse”间的却并没有。作者通过条件随机场理论推导了这种统计依赖和视觉关系检测之间的联系，并把这种关系设计到网络中，迭代的进行学习。

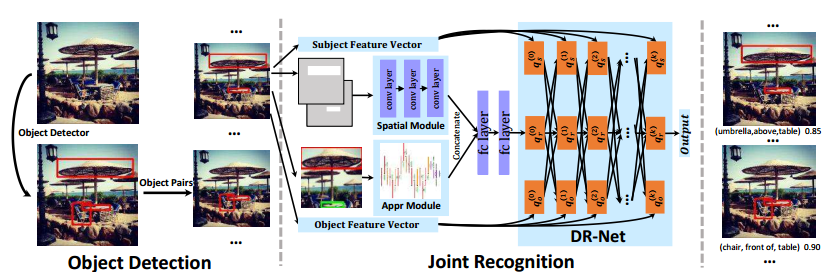


图 2.3 DR-Net网络结构图

如图2.3所示，DR-Net首先进行目标检测得到图像中的目标，再把目标配对送入网络进行联合识别。网络首先经过一个空间模块和一个外观模块得到相应的特征，最后再送入DR-Net统计模块进行预测，整个网络可以进行“端到端”的学习。

DR-Net网络在视觉关系公开数据集上取得了大幅的效果提升，将谓语识别任务从47%提升到了81%，为视觉关系检测问题设立了新的性能标杆。

* 1. 目标检测算法

目标检测是计算机视觉中的一个重要领域，它将目标的分割和识别合二为一，可以同时获取目标的位置信息和类别信息。目标检测在视频监控、自动驾驶等方面具有重要的研究意义和应用价值。

接下来，本节将从基于区域建议的方法和无区域建议的方法两方面来介绍目标检测的几个主流算法。

* + 1. 基于区域建议的方法

Girshick 等在2014年提出了R-CNN（Region-based convolutional neural networks）方法[22]，该方法将目标区域建议（Region proposal）和卷积神经网络相结合，流程如图2.4所示。首先，使用选择性搜索对输入的图像提取候选区域（即可能包含目标的区域），然后利用AlexNet[4]卷积神经网络对每个区域提取特征，接着利用支持向量机（Support Vector Machine, SVM）进行分类。同时该方法还使用非极大抑制来处理IoU (Intersection-over-Union) 过大的交叠区域，并用坐标框回归（Bounding-Box Regression）来改善检测结果。

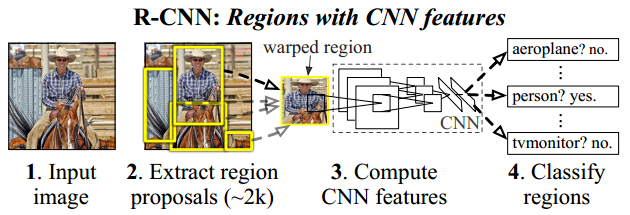


图 2.4 R-CNN算法流程图

R-CNN方法在PASCAL VOC 2012数据集上的平均均值精度（Mean average precision, mAP）为53.3 %，相比于传统方法有50%的提升。但是，R-CNN 也存在一些问题，如需要训练三个不同的模型（区域建议模型，分类模型和回归模型），由于候选区域存在重合而产生的重复计算，边框回归无法处理被遮挡的目标等。

于是不久后，Girshick 等又在2015年对R-CNN 进行了改进，提出了Fast R-CNN 算法[23]。算法的流程如图2.5所示，与R-CNN 类似，Fast R-CNN先利用选择性搜索提取图片中的感兴趣区域（Regions of Inter-est, RoI），然后通过卷积层得到每个RoI对应的特征图，再送入RoI池化层，这个网络层可以把不同大小的输入映射为一个固定长度的特征向量。最后，Fast R-CNN利用了多任务（Multi-Task）的思想，同时训练目标分类和边框回归，这两个任务可以共享卷积特征，提高了空间和时间效率。然而Fast R-CNN 也仍然存在速度上的瓶颈，无法实现实时检测，主要原因是区域建议步骤需耗费大量的时间。

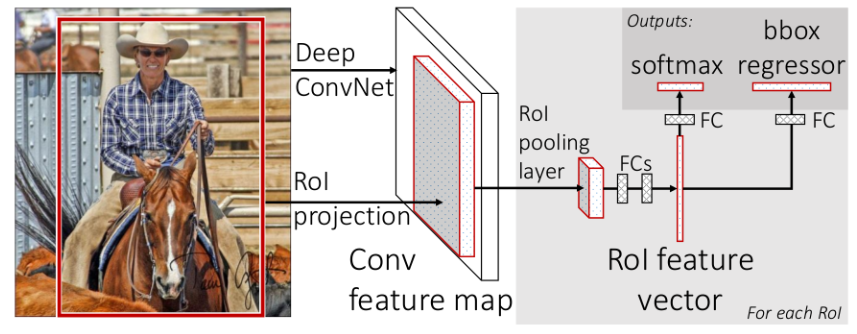


图 2.5 Fast R-CNN算法流程图

因此，为了突破目标检测在速度上的瓶颈，使目标检测可以实时的进行，Ren等在同年提出了区域建议网络（Region proposal network, RPN），并将RPN 和Fast R-CNN结合到一起，称为Faster R-CNN[24]。它用RPN取代了选择性搜索算法，使检测任务可以由神经网络端到端的完成。

RPN将区域建议的步骤建模为二分类的问题，即是否存在目标，故可以利用RPN直接训练得到候选区域。RPN网络的结构图如图2.6所示，它被接在最后一个卷积层之后，它在卷积层输出的特征图上滑动窗口，并将每个滑动窗口映射为一个低维向量，再分别输入分类层和边框回归层。对于每个窗口，以中心点为基准点选取k（文中使用k = 9）个不同尺度、不同长宽比的Anchor，对每个Anchor，分类层将判断其是否存在目标，边框回归层则会输出它的坐标位置。

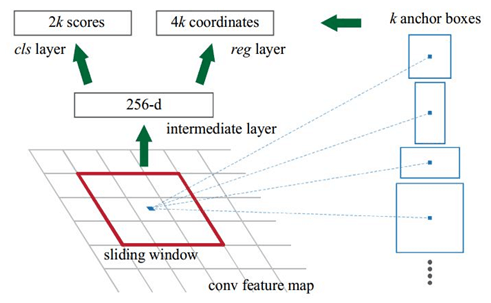


图 2.6 RPN网络基本结构示意图

* + 1. 无区域建议的方法

基于区域建议的目标检测方法存在一定的局限性，它们忽视了目标在整张图片中的空间信息，因此，学者们开展了无区域建议的目标检测研究。

Redmon等在2016年提出了YOLO (You only look once) [25]方法，它将检测任务定义为一个统一的、端到端的回归问题，是一种无区域建议的目标检测方法。

YOLO将输入的图片划分为个网格，每个格子负责中心在该格子中的目标检测。YOLO网络结构类似GoogLeNet[6]，由24个卷积层和2个全连接层构成，卷积层主要用来提取特征，全连接层则用于预测类别和边框位置坐标。网络的输入为448448，输出是一个维度为的张量，其中*B*为每个网格所负责的目标数量，*C*为类别数。

YOLO的优点是检测速度很快，可以达到45 fps，主要是因为它减少了对同一目标的重复检测，大大减少了时间复杂度，以达到实时的效果。但是YOLO也有一定的局限性，它的整体性能不如Fast R-CNN，且对小目标和相邻目标检测效果不好。

针对YOLO 存在的不足，Liu等在同年提出了SSD（Single Shot MultiBox Detector）方法[26]，在提高效果的同时兼顾实时性的要求。相比于YOLO，SSD同样使用了卷积神经网络对图片提取特征，但是它的特征是从不同的卷积层提取出来的多尺度特征图，对它们组合后再进行分类和回归。另外SSD也使用了更多的Anchor，每个网格点生成不同大小和长宽比例的边框，并将类别预测概率基于边框预测。在公开数据集上的测试结果显示，SSD与基于区域建议的方法相比速度提高了很多，且可以保证精度；与YOLO 相比，SSD 能取得更高的精度，且对小图像的表现更好。

* 1. 语义嵌入

在自然语言处理中，需要理解句子中的每个词组的语义信息，故有学者提出了语义嵌入的概念，而语义嵌入的方法中，词向量（Word Embedding）[27]是应用最为广泛的方法，本节将对其进行介绍。

* + 1. 词向量

在自然语言处理中，首先需要考虑词如何在计算机中表示，最直观的方法是One-Hot表示，该方法把每个词看作一个原子符号，表示为一个高维度的长向量。该向量的维度即词表的大小，其中只有一个维度的值等于1，其它维度值均等于0，值为1的维度即代表了当前的词，如表2.1中实例所示。

表 2.1 One-hot词向量实例

|  |  |
| --- | --- |
| 词 | One-hot表示 |
| 北京 | [0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 ...] |
| 京城 | [0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 ...] |
| 纽约 | [0 0 0 0 0 0 0 0 0 0 ... 0 0 0 0 1 0] |

One-Hot表示的优点是简单明了，但是因为该方法采用了稀疏存储，容易造成维度灾难问题。另外该方法也无法表示词与词之间的关系，即词之间是相互独立的，如“北京”和“京城”实际是相近的含义，但是却无法表示出来。

因此，可以将词转化为一种分布式表示（Distributed Representation），又称词向量（Word Embedding）。词向量可以将词表示成一个定长的连续稠密的实数向量，这种表示可以体现词之间的相似关系，使相似的词在词向量空间中距离更近，同时也可以包含更多的信息，通常可以用欧式距离或余弦距离来衡量词向量之间的相似性。

* + 1. Word2vec

表示词向量的方法目前有很多，其中word2vec[28]是最常用的工具，它以将词转化为低维度的连续稠密的向量，如[0.324 0.233 0.998 … ]，利用这种工具，就可以把语义相似的词映射到向量空间中相近的位置上。图2.7列举了几个词在向量空间中的例子。

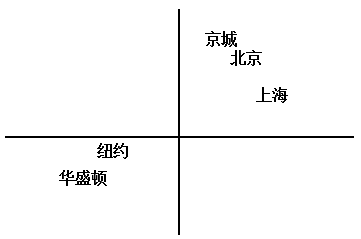


图 2.7 word2vec例图

如图2.7所示，图中词与词的距离体现了它们在语义上的相似性，比如“京城”和“北京”的含义相近，因此它们向量空间中也很接近。“上海”是和“北京”一样也是中国的城市，故它们的距离也比较相近。而“纽约”和“华盛顿”都是美国的城市，因此它们之间离得更近，而和其他三个词则离得较远。

word2vec主要分为CBOW（Continuous Bag of Words）和Skip-Gram两种模式。CBOW是给定上下文，预测词；而Skip-Gram则相反，是根据词来预测上下文。

* 1. 本章小结

本章介绍了本文的相关工作，首先介绍了视觉关系检测任务的几个经典算法，包括基于语言先验的方法、基于知识表征学习的方法和基于统计依赖的方法等。接着，介绍了目标检测的主流算法，包括R-CNN, Fast R-CNN, Faster R-CNN, YOLO和SSD等。最后，介绍了语义嵌入和词向量的相关知识。本章内容是后续章节研究内容的基础，是后续研究工作的铺垫。

# 基于多特征联合统计预测的谓语识别

* 1. 研究问题

随着视觉关系检测任务被正式提出，其包含的三个子任务也被定义[10]，分别是：

任务一：谓语识别任务（Predicate Detection），给定每对主语和宾语的标签以及Bounding Box（边界框），预测它们之间的谓语应该是什么。

任务二：短语检测任务（Phrase Detection），输入一张图像，需在准确预测出三元组<subject – predicate - object>的同时，将关系三元组所在整体区域用一个Bounding Box标出来，并要求其与真实Bounding Box的重叠率（Intersection over Union, IoU）在50%以上。

任务三：关系检测任务（Relationship Detection），与任务（2）相似，输入一张图像，需要准确预测出三元组<subject – predicate - object>，并要求分别找出主语与宾语的Bounding Box位置，并保证与真实Bounding Box的重叠率在50%以上。

不难看出，任务一的谓语识别是任务二和任务三的基础，也是整个视觉关系检测的关键。

因此，本章要研究的问题就是任务一中所定义的，如何在已知图片中每对主语和宾语标签及位置的情况下，预测它们之间的谓语是什么。可以将该问题抽象为一个数学问题：

公式（3.1）

即在已知图片、主语类别、宾语类别、主语位置和宾语位置的条件下，谓语的概率。其中表示需要求的谓语类别，表示谓语可能的类别的候选空间，表示输入的图像，表示主语的类别，表示宾语的类别，表示主语和宾语的类别空间，表示主语的位置，表示宾语的位置。

* 1. 基于多特征联合统计预测的谓语识别算法

在已知图片中每对主语和宾语标签及位置的情况下，预测它们之间的谓语可以被抽象为一个分类问题。然而它和普通的目标分类却有很大的区别，主要原因是，同一谓语在不同场景下的外观形式差异可能较大，比如同样是谓语“in”，“surfboard in hand（冲浪板在手上）”和“person in street（人在街道上）”却呈现了完全不同的表现形式。

针对这个问题，本文从两个角度出发，尝试缓解这种同类谓语之间的表现差异。

首先，本文利用多信息融合的方法，结合了图片中主语和宾语之间的视觉信息、位置信息和语义信息得到联合特征，并利用谓语和主语、宾语之间的统计关系信息来进行联合预测。特别的，在引入统计关系信息时，进一步考虑到了各部分统计信息的重要程度，提出了一种基于1x1卷积的统计信息引入方式。

另外，本文基于特征中心化的思想，利用中心损失（Center Loss）进行训练，希望同一分类的类内特征能够更加向中心靠拢，即类内距离更小，从而使得同一种谓语的联合特征可以在关系空间中变得更加相近。

基于以上两点，提出本文的整体网络框架，并对具体模块和损失函数进行详细介绍。

* + 1. 整体网络框架

本文提出的谓语识别算法的整体网络结构图如图3.1所示。

网络的输入为主语和宾语的联合区域图像，主语和宾语各自的类别及位置（用Bounding Box表示）。

首先，将主语和宾语的联合区域图像通过视觉模块，得到视觉特征信息；

同时，根据主语和宾语的Bounding Box位置得到双遮罩模板，通过位置模块，得到位置特征信息；

另外，根据主语和宾语的类别通过语义模块，得到语义特征信息。

将视觉特征、位置特征和语义特征结合到一起，通过全连接层得到联合特征，送入统计预测模块。

与联合特征一起被送入统计预测模块的还有主语和宾语各自的类别，统计预测模块利用它们之间的统计信息最终预测得到谓语的分类结果。

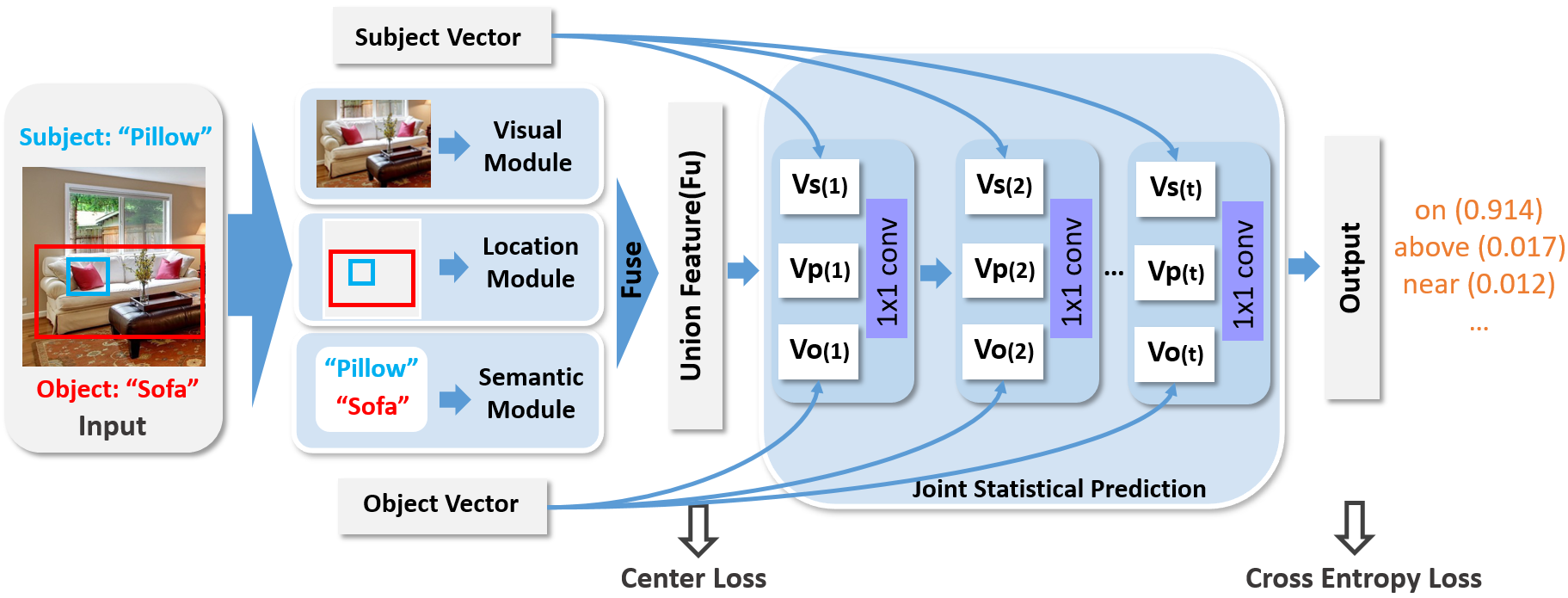


图 3.1基于多特征联合统计预测的谓语识别网络框架

这就是整个网络的整体结构框架，网络需要使用到的输入包括主语类别，宾语类别，主语和宾语的联合区域图像，主语位置和宾语位置。则该方法可以抽象为函数

公式（3.2）

下面对网络中的每个模块做具体介绍。

* + 1. 特征模块介绍
       1. 视觉模块

当判断图片中一组主语和宾语之间的关系时，它们所展现的图片视觉内容往往是最为直观的信息。同时，通过视觉内容判断关系也是与人类的判断思路最接近的方式，比如当人看到“person wear shoes（人穿着鞋子）”的时候，可能不用去深入思考人和鞋子分别代表着什么，也不用仔细思考人和鞋子的位置关系，而仅仅是通过这个画面就可以得出“穿”这个谓语关系。

因此，视觉模块是整个网络的基础，扮演着十分重要的角色。

本文利用卷积神经网络（[Convolutional](file:///E:\SETUP\%E6%9C%89%E9%81%93%E8%AF%8D%E5%85%B8\7.5.0.0\resultui\dict\?keyword=Convolutional)[Neural](file:///E:\SETUP\%E6%9C%89%E9%81%93%E8%AF%8D%E5%85%B8\7.5.0.0\resultui\dict\?keyword=Neural)[Network](file:///E:\SETUP\%E6%9C%89%E9%81%93%E8%AF%8D%E5%85%B8\7.5.0.0\resultui\dict\?keyword=Network), [CNN](file:///E:\SETUP\%E6%9C%89%E9%81%93%E8%AF%8D%E5%85%B8\7.5.0.0\resultui\dict\?keyword=CNN)）作为视觉模块，这里具体使用的是VGG16网络[5]，其具体网络结构如下图3.2所示。

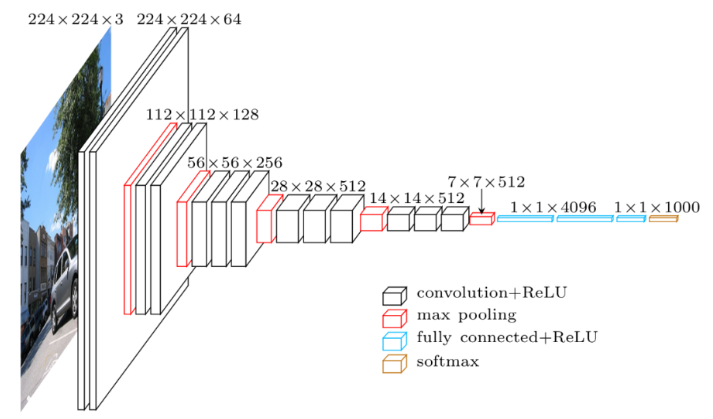


图 3.2 VGG16网络结构图

与原始的VGG16网络结构稍有不同的是，在卷积层之后，Feature Map将先后通过两个4096输出的全连接层和一个256输出的全连接层，最终得到一个长度为256的特征向量。

在视觉模块中，输入网络的图片为主语和宾语的联合区域图像，即主语目标图像和宾语目标图像的并集，并向外稍作边缘扩展，如图3.3红色方框所示。

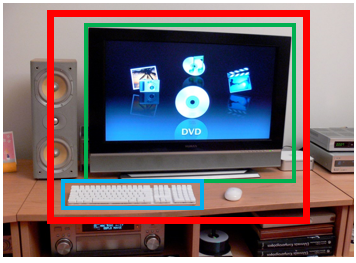


图 3.3 主语图像，宾语图像和联合区域图像

图片经过视觉模块后，最终会得到一个长度为256的视觉特征向量，它将作为视觉信息，与其他信息进行融合得到联合特征，一起作用于最后的预测。

* + - 1. 位置模块

虽然对于人来说，通过视觉内容可以很快的判断出关系的类别，但是对于神经网络来说，仅仅通过视觉信息往往还不足以准确地判断出两个目标之间的关系。

比如“person ride horse（人骑着马）”和“person near horse（人在马旁边）”，从视觉特征的角度来看，卷积神经网络可能很难判断这两者的区别，因为它们拥有的视觉特征可能十分类似，如图3.4。

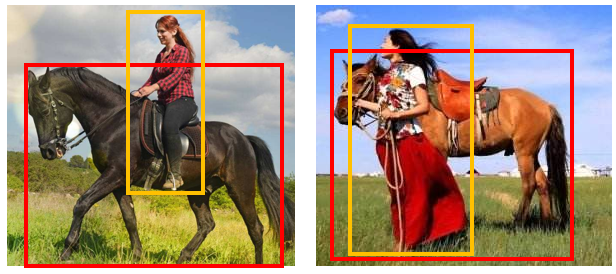


图 3.4 “人骑着马”和“人在马旁边”对比图

但是如果同时考虑进来它们的位置，则可以看出在“人骑着马”的图像中，人的相对位置和马的相对位置在中上方重合，且人的位置高于马；而“人在马旁边”的图像中，人的相对位置则和马的相对位置在左方重合，人的位置并没有很明显的高于马。

因此结合这种“位置信息”，可以更好的对主语和宾主之间的关系做出判断。除了刚刚所列举的这种关系以外，还有很多与“方位”相关的谓语尤其需要位置信息的补充，比如“on（上面）”、“above（上方）”、“below（下方）”、“inside（内部）”、“in the front of（前面）”和“behind（后面）”等等。

在本文中，位置模块的输入采用主语和宾语两个目标的图像掩模（Picture Mask）来构成，“掩模”起源于半导体工艺中的芯片制造，而图像掩模则可以用于提取图片中感兴趣的区域、或得到其空间信息，是图像处理方法中的一种。

在这里，图像掩模即一个单通道的图片，原图中目标所在的区域令其值为1，其余区域值为0。由于有主语和宾语两个目标，故需要提取两个图像掩模，最后把它们叠在一起送入位置模块。

位置模块的结构参考了Dai等人[11]的方法，如图3.5所示。

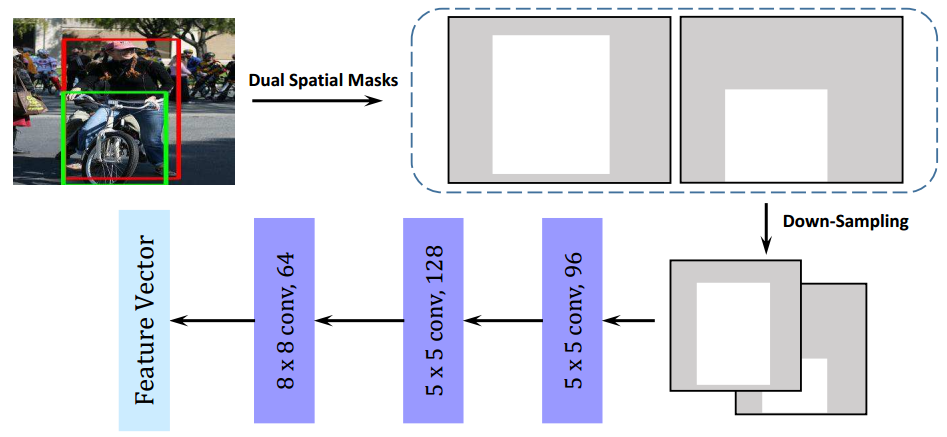


图 3.5 位置模块结构图

首先得到主语和宾语的图像掩模，在通道方向上叠在一起，并降采样到的大小，送入位置模块。在位置模块中，使用了三个卷积层，分别是96核的卷积层、128核的卷积层和64核的卷积层。最终将得到一个的Feature Map，将其展开即可得到一个长度为64的位置特征向量。它将作为位置信息，与其他信息进行融合得到联合特征，一起作用于最后的判断。

* + - 1. 语义模块

对于可能作为主语和宾语的目标来说，目标之间可能存在一定的相似性，比如“jacket（夹克衫）”和“coat（外套）”就比较相近，因为都是可以穿的上衣；再比如“horse（马）”和“elephant（大象）”也比较相近，因为都是动物且都是可以载人的动物。

因此，“jacket”和“coat”都可以与“person”组成“person wear jacket（人穿着夹克衫）”和“person wear coat（人穿着外套）”，因为它们都适用于“人穿着…”这个关系；同样的，它们也都可以与“person”组成“jacket on person（夹克衫在人身上）”和“coat on person（外套在人身上）”，因为它们都适用于“...在人身上”这个关系。

与此类似，“horse”和“elephant”也都可以与“person”组成“person ride …（人骑着…）”的关系，因为“马”和“大象”都是可以被人骑的大型动物。

由此可见，理解一个目标的“语义信息”可以更好的帮助我们来判断关系，对于神经网络来说也是如此。而人类对于目标的语义知识很多时候来自于人类的先验认知，而神经网络却没有这种“认知”，因此我们需要通过其他手段来让神经网络理解这种“语义”。

一般来说，主语和宾语的目标分类可以用一个One-Hot向量作为分类向量来最直观的表示，即假设目标的类别数为*N*，则类别向量是一个长为*N*的向量，如果目标的分类为第*i*个，则向量中第*i*位为1，其他值均为0。以“person（人）”和“table（桌子）”为例，假设目标类别总数为100个，它们分别是第1类和第6类，则它们的One-Hot向量为下表3.1所示。但是这种表示方法十分稀疏，且无法体现目标之间的关系，也无法体现我们需要的语义。

表 3.1 person和table的One-Hot表示

|  |  |
| --- | --- |
| 目标类别（词） | One-Hot向量 |
| person | [1 0 0 0 0 0 0 0 0 … 0 0 0] |
| table | [0 0 0 0 0 1 0 0 0 … 0 0 0] |

因此，在本文中，语义模块利用了语义嵌入的思想，把目标的类别转换为一种语义嵌入表达形式，这里使用了最常用的语义嵌入表达方式——词向量（Word Embedding Vector）**错误!未找到引用源。**。即把较为稀疏的One-Hot向量映射到更加稠密的实数向量空间中，这个得到的稠密实数向量即为词向量。

为了使训练出来的词向量表示更加准确和符合语义关系，这里利用word2vec[28]工具，使用其预训练好的参数来当作语义嵌入层的初始参数。同样以“person（人）”和“table（桌子）”为例，它们的word2vec词向量表达形式如表3.2所示。这样，两个向量之间的几何距离就可以反映出两个目标之间的语义相似程度。

表 3.2 person和table的word2vec词向量表示

|  |  |
| --- | --- |
| 目标类别（词） | One-hot向量 |
| person | [0.1208, -0.1084, 0.0076, 0.0737, … , 0.0326, -0.0405] |
| table | [-0.0539, -0.0099, 0.0439, 0.0798 , … , -0.0366, -0.0046] |

在该模块中，本文使用了一个300维输出的语义嵌入层来训练词向量，每个目标可以得到一个长为300的词向量，再把主语和宾语的词向量连接在一起，最终得到一个长度为600维的语义特征向量。语义模块的结构如下图3.6所示。

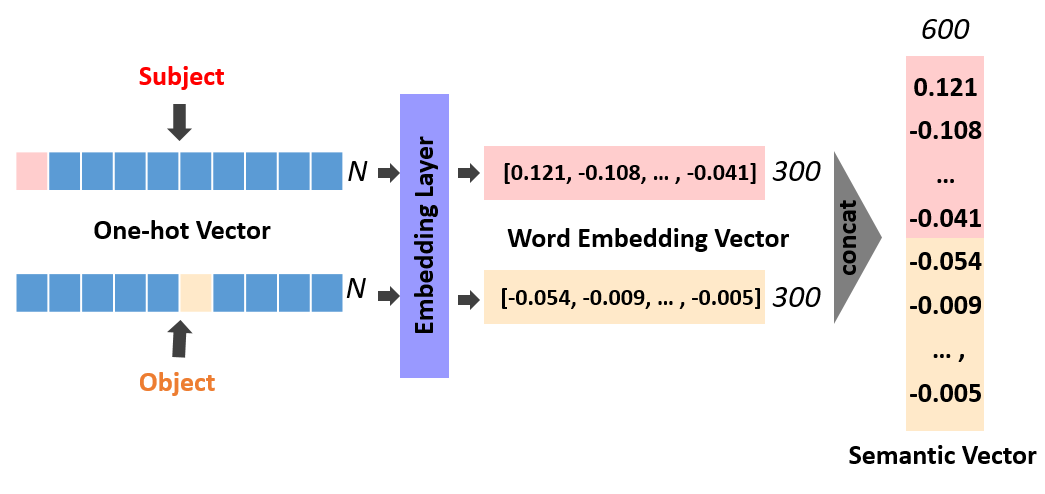


图 3.6 语义模块结构示意图

该语义模块得到的长度为600的语义特征向量将作为语义信息，与其他信息进行融合得到联合特征，一起作用于最后的预测。

* + 1. 联合统计预测
       1. 统计依赖的意义

利用上文介绍的三个模块——视觉模块、位置模块和语义模块，可以得到视觉特征、位置特征和语义特征，将它们连接在一起，通过全连接层进行融合即可以得到一个联合特征。

对于该联合特征，可以简单的再通过一个输出为谓语类别数量的全连接层得到谓语分类，但是如果直接这么做，效果并不理想。因此，这里将引入DR-Net[11]中的统计依赖思想，利用统计信息，结合之前得到的联合特征来一起进行综合预测，最终得到谓语的类别。

在一组关系中，主语、宾语和谓语之间往往有着很强的统计依赖关系。比如“person ride horse（人骑着马）”是一个很常见且合理的关系，但是“horse ride person（马骑着人）”、“person wear horse（人穿着马）”或“person ride mouse（人骑着老鼠）”则并不容易出现，因此可以看出“ride - horse”间有强烈的依赖关系，而“horse - ride”、“ride - person”、“wear - horse”和“ride - mouse”间的却并没有。基于这种直觉，可以把这种统计依赖应用到网络中。

根据条件随机场理论，可以把视觉关系中的统计关系表示为

公式（3.3）

其中，表示谓语的特征，和表示主语和宾语各自的特征，表示模型参数，表示正则化项。而联合势函数可以被表示为各部分的势之和

公式（3.4）

其中，单元势把目标（主语和宾语）和它的特征联系起来，把谓语和联合特征联系起来；而双元势把，和则用于获取主语、宾语和谓语之间的统计关系。

* + - 1. 利用统计依赖预测谓语

为了把这种推理应用到神经网络中，可以得到谓语*p*的后验概率

公式（3.5）

令和为矩阵且, ，则*p*的后验概率向量可以表示为

公式（3.6）

其中，表示*softmax*函数，和分别表示主语和宾语的One-Hot向量表示。

在DR-Net中，作者用主语和宾语的概率向量和代替和，因此可以得到更新方程

公式（3.7）

其中关系矩阵, , 。并使用神经网络中的网络层来完成这个公式的计算和迭代。

然而，这个公式中有两个局限性：

1. 主语和宾语的概率向量也要在网络中迭代，一方面大大增加了计算量，另一方面这种迭代对主语和宾语本身的修正作用并不大；
2. 在考虑主语跟谓语的统计依赖、宾语跟谓语的统计依赖、以及联合特征跟谓语的统计依赖这三个分量时，没有考虑到它们之间的区别和贡献的重要程度。

因此本文对统计依赖的引用做出了改进，一方面不再对主语和宾语的概率向量进行迭代，另一方面对上述三个分量的重要程度做出区分，并利用1x1卷积的通道融合能力来实现这一功能。

改进后，谓语*p*的后验概率向量的迭代过程可以表示为

公式（3.8）

其中的初始值为

公式（3.9）

其中，表示上文三个模块得到的联合特征，和分别表示主语和宾语的目标分类特征，, 和分别控制三个分量的重要程度，为一个偏执项，表示联合特征跟谓语之间的关系，和分别表示主语和宾语跟谓语之间的统计依赖关系。

* + - 1. 利用1x1卷积进行通道融合

在卷积神经网络中，卷积层是整个网络中最重要的网络层，是整个网络的核心。卷积层基于局部感知思想，对输入的图像或特征图进行“滑窗”计算来完成“卷积”操作。可以理解为使用一个过滤器（即卷积核）来过滤图像或特征图中的每个小区域，从而得到这些小区域的特征值。

而这些“小区域”的大小，则由卷积核的大小来决定，因此卷积核大小（Kernel Size）是卷积层中的一个重要参数。在最早的AlexNet[4]网络中，卷积层中卷积核的大小包括1111，5×5和3×3；而在VGG[5]网络中，卷积层的卷积核大小则都设为3×3。事实上，3×3的卷积核大小是卷积神经网络中使用最为广泛的尺寸，因为相比于更大的卷积核来说，3×3的卷积核大小是能捕获“上下左右”和“中心”概念的最小尺寸，且多个3×3卷积核大小的卷积层可以在保持相同感受野大小的同时，为网络带来更多的非线性，且参数更少。

然而在2014年提出的一种新型深度网络结构（Network In Network, NIN）[29]中，1×1卷积被首次提出，一方面它可以实现跨通道的融合和信息交互，另一方面它可以对通道数进行升维和降维。随后在GoogLeNet[6] 和ResNet[7]中，1×1卷积也被应用到了网络中，GoogLeNet使用1×1卷积得到了更加紧凑的网络，使参数量大大减少；而ResNet则使用1×1卷积设计出了BottleNeck结构，如图3.7所示。

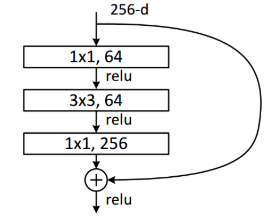


图 3.7 ResNet的BottleNeck结构

1×1卷积先将256通道数的输入降到64通道，使通道数减少为原来的四分之一，再进行普通的3×3卷积操作，卷积后的输出通道数等于输入通道数64，最后再把通道数恢复到256。这种BottleNeck结构使用两个卷积有效的减少了参数个数和计算量，同时又加强了通道之间的交汇融合。

因此，本文也将引入1×1卷积的通道融合作用，在网络中实现上节所推导的公式。

将上文中推导的谓语*p*的后验概率向量进行进一步改写

公式（3.10）

其中，, , ，则如果把它们看作特征图中的三个通道，即可以通过1×1卷积对它们进行融合。如图3.8所示。

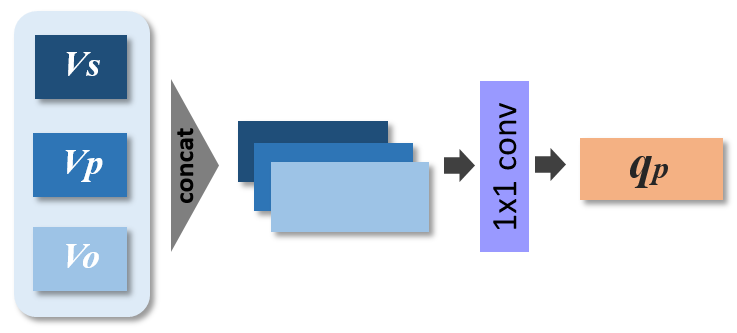


图 3.8 利用1×1卷积进行通道融合

每一次迭代的网络可以看作一个单元（Unit），假设单元数量为*t*个，则将迭代*t*次，因此整个联合统计预测模块的结构如图3.9所示。

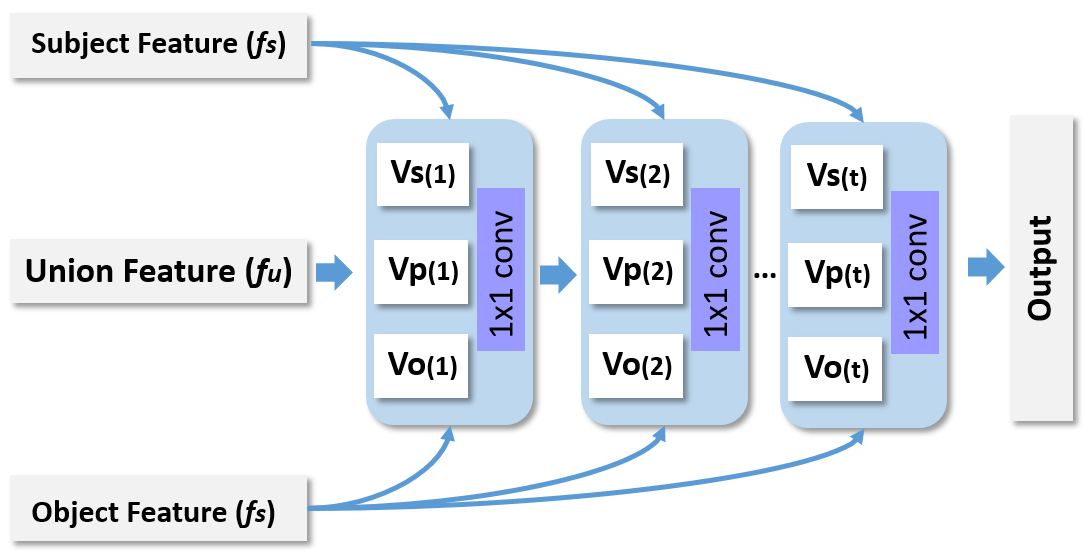


图 3.9 联合统计预测模块结构图

通过联合统计预测，即可得到最后的输出，即谓语的分类概率向量，向量中概率最大的类即可作为谓语的分类。

* + 1. 损失函数

谓语识别任务的本质上是一个分类问题，因此本文将用到分类问题常用的损失函数——交叉熵损失函数。但是由于同种谓语在不同关系中可能存在较大的特征差异，而本文希望可以尽可能的缩小这种差距，即希望使同种谓语的联合特征在关系空间中更加相近，因此本文将同时引入中心损失函数。

本文使用交叉熵损失结合中心损失函数的方法来训练网络，一方面利用交叉熵损失来使谓语关系接近真实分类，使不同分类之间的特征差异最大化；另一方面利用中心损失函数来使同一类谓语的联合特征更加相近，以此使同类特征之间的差异最小化。

* + - 1. 交叉熵损失函数

交叉熵损失函数（Cross Entropy Loss）因为其求导简单、易于计算等特点，被广泛的用于深度学习的分类问题中。交叉熵损失的数学定义如公式（3.11）。

公式（3.11）

其中表示期望的输出值（即真实标签），表示模型的预测值。由于One-Hot标签的特殊性，只有相应分类的部分为1，其他全为0，因此上述公式可以简化为

公式（3.12）

而在神经网络中，一般来说最后一层在输出前都会先经过Softmax计算，它会对网络前一层的输出计算指数并做归一化，使所有分类对应输出值之和为1。如公式（3.13）

公式（3.13）

其中，*i*和*j*都表示类别的索引，表示第*i*类对应的输出值。

因此，经过Softmax的交叉熵损失（下文简称Softmax损失）为

公式（3.14）

* + - 1. 中心损失函数

中心损失函数[30]在2016年提出被用于人脸识别任务，在人脸识别任务中取得了不错的效果。交叉熵损失试图使不同的分类之间的特征差异尽量大，而中心损失函数则尽量使每个分类的内部特征更加靠近类的中心。

如图3.10所示，在左边的图中，两个分类的特征点离分类超平面很近，区分不够明显，而通过中心损失训练后，在右边的图中，两个分类的特征点则分别都更向类的中心聚拢，离分类超平面也更远，这可以使分类的准确度更高（因为其可以容忍一定的分类超平面偏差，鲁棒性更好）。

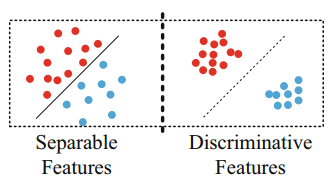


图 3.10 中心损失函数效果示意图

中心损失函数定义为一个mini-batch中，所有样本的特征与其所属分类中心的距离的平方和，其描述了类的内部差异。如公式（3.15）所示。

公式（3.15）

其中代表当前mini-batch中的第*i*个样本的特征，且其属于第个分类，表示第个分类的中心，*m*为mini-batch的大小。

想要缩小类的内部差异，则需要使中心损失尽量小，该过程可以使用迭代优化（如梯度下降）实现。而在迭代优化的过程中，分类的中心是会不断变化的，因此还需要对分类中心进行更新。事实上，中心损失的思想和过程很类似于K-Means聚类算法。中心损失函数的梯度公式和分类中心的更新公式如公式（3.16）和公式（3.17）所示。

公式（3.16）

公式（3.17）

其中，表示第*j*个分类的中心。条件函数表示当条件满足时，其值为1；当条件不满足时，其值为0.

本文使用交叉熵损失结合中心损失函数的方法，来训练网络，一方面利用交叉熵损失来使谓语关系接近真实分类，使不同分类之间的特征差异最大化；一方面使用中心损失函数来使同一类谓语的联合特征更加相近，以此使同类特征之间的差异最小化。

结合后的损失函数为

公式（3.18）

其中，为一个调节参数，用于调节交叉熵损失和中心损失的比重。

在本文的模型中，交叉熵损失用在模型最后的输出上，而中心损失则用在由三个特征模块得到的联合特征上。如上文中的图3.1所示。

* 1. 实验与分析

本节将对上文所提出的谓词检测算法进行实验，下面将先介绍实验使用的数据集及评价指标，接着介绍实验的设置，最后对实验结果进行分析和比较。

* + 1. 数据集

本文使用的数据集是Lu等人在2016年公开的视觉关系数据集VRD（Visual Relationship Dataset）[10]。该数据集中包括5000张图片，100种目标类别和70种谓语类别。在这些图片中，总共有37993个关系样本，包括6672种<主语，谓语，宾语>三元组的关系种类。数据集被默认划分为两部分，一部分的4000张图片作为训练集，另一部分的1000张图片作为测试集。

数据集中的每张图片都会包含若干组关系，用<主语-谓语-宾语>三元组的形式表示出来，并给出主语和宾语的位置。数据集中的包括的数据如表3.3所示。

表 3.3 VRD数据集数据描述

|  |  |
| --- | --- |
| 数据名称 | 描述 |
| Object\_List | 包含100种目标类别的列表 |
| Relation\_List | 包含70种谓语类别的列表 |
| Image\_Path | 所有图片的名称 |
| Tuple\_Label | 所有关系对应的<主语-谓语-宾语>三元组的标签 |
| Sub\_Bboxes | 所有关系中主语的位置（用Bounding Box表示） |
| Obj\_Bboxes | 所有关系中宾语的位置（用Bounding Box表示） |

在数据集包含的100种目标里，涵盖了人及动物、自然景观、建筑、交通、家具家电、服饰、日常用品和食物等几大类目标，每一大类中的具体目标举例如表3.4所示。

表 3.4 VRD数据集中的目标样例

|  |  |
| --- | --- |
| 目标类型 | 目标样例 |
| 人及动物 | person(人), horse(马), cat(猫), elephant(大象) |
| 自然景观 | sky(天空), tree(树), grass(草地) |
| 建筑 | building(楼房), tower(塔), roof(屋顶) |
| 交通 | truck(卡车), bus(巴士), train(火车), motorcycle(摩托车) |
| 家具家电 | table(桌子), chair(椅子), bed(床), lamp(灯) |
| 服饰 | hat(帽子), coat(外套), pants(裤子), tie(领带), shoes(鞋) |
| 日常物品 | phone(电话), laptop(笔记本电脑), bottle(瓶子), kite(风筝) |
| 食物 | pizza(匹萨) |

VRD数据集中的70种谓语里，包括表示方位的介词、动作词、动作+介词和比较词等。具体的谓语举例由表3.5给出。

表 3.5 VRD数据集谓语举例

|  |  |
| --- | --- |
| 谓语类型 | 谓语举例 |
| 介词 | on(上面), above(上方), below(下方), inside(内部), behind (后面) |
| 动作词 | wear(穿), look(看), ride(骑), eat(吃), feed(喂) |
| 动作+介词 | sit next to(坐在旁边), park behind(停在后面), stand under(站在下面) |
| 比较词 | taller than(比…高) |

本文将使用数据集默认的训练集、测试集划分，训练集为4000张图片，测试集为1000张图片。

* + 1. 评价指标

当训练好谓语识别模型后，就可以根据学习出来的模型对谓语进行预测，并得到其相应的置信度。

普通的分类任务常用的评价指标有准确率（Accuracy, Acc）或平均均值精度（Mean Average Precision, mAP）等，然而这些评价指标在视觉关系检测的任务中却不太适用。

由于视觉关系检测任务与普通的分类任务不同，不但涉及分类、还涉及对检测出的关系的置信度的排序，且数据集中的标注可能存在不够完整的情况，比如

1）目标检测标注不完整；

2）有些目标对之间存在关系，但是没有标注出来；

3）一对主语和宾主之间可能存在多种合理的谓语，但是没有标注出来（比如“next to”和“near”都可以表示“旁边”的关系等）。

因此，需要寻找除了准确率和平均均值精度外，更加适用于视觉关系检测的评价指标。本文使用的评价指标延用了Lu等人提出的视觉关系检测评价指标，即Recall@50和Recall@100这两个评价指标。

Recall@K的定义为，检测出的所有<主语-谓语-宾语>三元组关系中，置信度排序最靠前的*K*个关系中检测正确的个数，与真实关系数量的比值。

Recall@K的具体计算方式可以用如公式（3.19）表示

公式（3.19）

其中，*K*为定义的阈值（将从检测出来的关系中取置信度最高的*K*个关系），*I*表示测试集中图片的数量，表示从第*i*张图片中检测出的Top第*j*个关系，表示第*i*张图的真实关系空间，表示第*i*张图片中真实关系的数量。条件函数表示当条件满足时，其值为1；当条件不满足时，其值为0。

* + 1. 实验设置

下面介绍模型训练过程中的一些实验设置。

在模块的设置上，视觉模块使用的基础模型为VGG16网络结构，并使用其在ImageNet上的预训练参数作为初始参数。语义模块则使用了word2vec的预训练参数作为语义嵌入层的初始参数。联合统计预测中的单元数设为8。

在训练的过程中，参数的优化方法使用的是随机梯度下降（Stochastic Gradient Descent, SGD），网络整体的学习率设置为0.001，语义嵌入层的学习率单独设置为0.0001.

训练框架使用的是Pytorch深度学习框架，并在单张NVIDIA GeForce GTX 1080Ti显卡上进行训练。

另外，图片数据在输入网络前，先要做一定的预处理操作，包括：

1. 调整尺寸，所有的图片将首先被统一调整为的大小；
2. 张量变换，所有图片将被转换为统一的的组织格式（其中*C*为通道，*H*为高度，*W*为宽度），且像素大小被映射到[0, 1]范围内；
3. 标准化，所有图片的像素值将按通道*C*分别被规范到标准正态分布，如公式（3.20）所示。

公式（3.20）

其中为通道的像素值，为通道的像素的均值，为通道的像素的标准差。

* + 1. 实验结果分析
       1. 与现有算法比较

为了更好的验证本章提出的谓语识别算法的效果，本文将把该谓语识别算法与现有的算法进行比较。参与比较的算法包括：

Visual Phrase [9]：把一组关系三元组看作一个整体，如“a person riding a horse”等，预测时不对主语、谓语、宾语进行切分。将每个关系三元组的整体作为不同分类来训练的代表做，并使用DPM检测器。

Joint-CNN[10]：一种具有*2N+K*种分类输出的神经网络，联动的判断主语、宾语和谓语的分类。

VR-L[10]：一种基于语言先验的视觉关系检测方法，通过外观模块把对象（包括主语和宾语）和谓语分别作为单独的模块进行训练，并把训练结果组合在一起用于预测最后的关系，再进一步通过语言模块把关系映射到向量空间中，并利用语言先验知识来进行约束和优化，最终得到预测出的关系。

VR-V[10]：与VR-L使用类似的方法，但是不加入语言先验，只使用外观模块进行视觉关系检测。

DR-Net[11]：一种深度关系网络，它基于谓语、主语和宾语之间的统计依赖关系，并把统计模型通过条件随机场理论和深度学习结合起来。

本文的谓语识别方法与以上已有算法的效果对比如表3.6所示。

从表中可以看出，直接对关系整体进行分类的Visual Phrase算法效果最差，说明了简单将不同的关系三元组作为不同的类别来进行训练和分类是一项很难的任务，因为存在巨大的可选关系空间，故模型需要识别的分类数为*O(N2K)*个（*N*表示目标种类数，*K*表示谓语种类数）。

基于主语、宾语和谓语联合分类的Joint-CNN仅仅略好于Visual Phrase算法，但效果依旧很差，说明在巨大的关系空间中同时对主语、宾语、谓语进行识别依旧是一项很难的任务。

表 3.6 本文谓语识别方法与现有算法在VRD数据集上效果对比

|  |  |  |
| --- | --- | --- |
| 算法 | Recall@50/% | Recall@100/% |
| Visual Phrase | 0.97 | 1.91 |
| Joint-CNN | 1.47 | 2.03 |
| VR-V | 7.11 | 7.11 |
| VR-L | 47.87 | 47.87 |
| DR-Net | 80.79 | 81.90 |
| 本文算法 | **83.31** | **85.82** |

对比只使用外观模块的VR-V和加入语言先验的VR-L，可以看出，在算法中加入语言先验可以为谓语识别带来巨大的效果提升。

DR-Net相较于前面几个算法有了巨大的效果提升，这说明在视觉关系中确实存在一定的统计依赖关系，也可以看出统计依赖关系在谓语识别中起到了显著的作用。

而本文提出的算法相较于DR-Net算法在Recall@50和Recall@100上分别有2.52%和3.92%的提升，说明将多个特征模块得到的联合特征中心化、并在联合统计预测时考虑到不同分量的权重可以一定程度上提高谓语识别的效果。

* + - 1. 不同组合比较

除了对比本文的谓语识别算法和其他现有算法的效果外，本文也对该算法中的不同组合进行了测试和比较，不同的组合选项分别表示为：

V (Visual)：使用视觉模块提取主语和宾语联合区域图片的视觉特征。

L (Location)：使用位置模块提取主语和宾语的位置特征。

E (Embedding)：使用语义嵌入模块提取主语和宾语的语义特征。

S (Statistical)：使用统计信息结合联合特征进行联合统计预测。

C(Center Loss)：使用中心损失函数进行训练。

不同模块的比较结果如表3.7所示。

表 3.7 本文谓语识别算法不同模块比较

|  |  |  |
| --- | --- | --- |
| 使用的模块 | Recall@50/% | Recall@100/% |
| V+S+C | 82.68 | 84.87 |
| L+S+C | 82.97 | 84.67 |
| E+S+C | 82.82 | 84.79 |
| V+L+E+C | 77.14 | 79.75 |
| V+L+E+S | 83.15 | 84.92 |
| V+L+E+S+C | **83.31** | **85.82** |

从表中可以看出，视觉模块、位置模块、语义模块分别于统计预测模块组合，呈现的效果较为类似，Recall@50和Recall@100的[平均指标分别为82.82](mailto:平均指标为Recall@50为82.82)和84.78，三种组合的差异在0.3%以内，说明这三个模块在算法中的贡献差不多，差别并不大。

对比只使用视觉模块、位置模块和语义模块与加上统计预测模块的结果，发现Recall@50和Recall@100分别相差6.17%和6.07%，说明统计预测模块在算法中扮演者十分重要的角色，可以为模型带来极大的效果提升。

对比不使用中心损失和使用中心损失进行训练的结果，发现加入中心损失进行训练可以在Recall@50和Recall@100上分别提升0.16%和0.9%，说明加入中心损失进行训练对模型有一定的提升作用。

* 1. 本章小结

本章首先介绍了谓语识别任务的问题描述和定义，并根据问题的特点提出了基于多特征联合统计预测的谓语识别算法。

接下来，本章详细介绍了该算法，先展示了该算法的整体框架，再详细介绍了该算法中的每个模块，包括视觉模块、位置模块、语义模块和联合统计预测模块，并对该算法使用的损失函数进行说明。

最后是实验部分，首先介绍了实验使用的数据集及评价指标，接着介绍实验的设置，最后从与现有算法的比较和不同模块的比较两个方面，对实验结果进行分析和比较。

# 基于目标对筛选和联合谓语识别的视觉关系检测

* 1. 研究问题

在第三章中，本文已经针对任务一提出了一种基于多特征联合统计预测的谓语识别算法。

接下来，本章将针对任务二和任务三，提出一种视觉关系检测算法。任务二和任务三的定义分别为：

任务二： 短语检测任务（Phrase Detection），输入一张图像，需在准确预测出三元组<subject – predicate - object>的同时，将关系三元组所在整体区域用一个Bounding Box标出来，并要求其与真实Bounding Box的重叠率（Intersection over Union, IoU）在50%以上。

任务三：关系检测任务（Relationship Detection），与任务二相似，输入一张图像，需要准确预测出三元组<subject – predicate - object>，并要求分别找出主语与宾语的Bounding Box位置，并保证与真实Bounding Box的重叠率在50%以上。

任务二、任务三与任务一的区别是，输入仅仅为一张图片，而不包括任何主语和宾语的信息，需要自己去找出合适的目标作为主语和宾语，最后需要给出完整的<主语-谓语-宾语>三元组，及整体区域的位置（任务二）或主语和宾语分别的位置（任务三）。可以看出，任务二和任务三是任务一的进一步推广。

因此，本章要研究的问题就是任务二和任务三中所定义的，如何在已知图片的情况下，找到图片中有效的视觉关系。

任务二的问题可以用以下数学表达式描述：

公式（4.1）

即在已知图片的条件下，求<主语-谓语-宾语>三元组及其相应区域位置的概率。其中表示主语的类别，表示谓语的类别，表示宾语的类别，表示谓语的类别空间，表示主语和宾语的类别空间，表示输入的图像，表示三元组整体区域的位置。

和任务二类似，任务三的问题可以用以下数学表达式描述：

公式（4.2）

即在已知图片的条件下，求<主语-谓语-宾语>三元组及主语和宾语区域位置的概率。分别表示主语的区域位置和宾语的区域位置，其他符号及变量与任务二相同。

* 1. 基于目标对筛选和联合谓语识别的视觉关系检测

在已知图片的情况下，找出图片中的视觉关系可以被抽象为一个检索加分类的问题，而不是一个简单的分类问题。

而在已知一幅图中所有目标的情况下，可以组成的目标对数量可能远大于实际存在关系的目标对数量，即大多数目标对之间是不存在关系的。

因此，基于这个动机，本文提出了一种基于目标对筛选和联合谓语识别的视觉关系检测算法，将视觉关系检测的过程划分为三个阶段，即目标检测阶段、目标对筛选阶段和谓语识别阶段。

下面本节将先介绍整个算法的流程框架，并重点对目标对筛选模型进行详细的介绍。

* + 1. 算法框架

本文提出的基于目标对筛选和联合谓语识别的视觉关系检测的整体算法框架如图4.1所示。

整体算法流程可以分解成三个阶段——目标检测阶段、目标对筛选阶段和谓语识别阶段。

1. 目标检测

首先，当已知一张图像时，需要把图像中属于目标空间的所有目标检测出来，并得到这些目标的类别和位置分别是什么。

在图4.1中，该图像中可以检测出的目标有2个“person（人）”（后面的成年男子和前面的男孩），1个“hat（帽子）”和1个“glasses（眼镜）”。每个目标的位置如图中彩色方框所示。

现有的目标检测模型有很多，如R-CNN[22]，Fast R-CNN[23]和Faster R-CNN[24]等，本文将使用Faster R-CNN作为目标检测器。

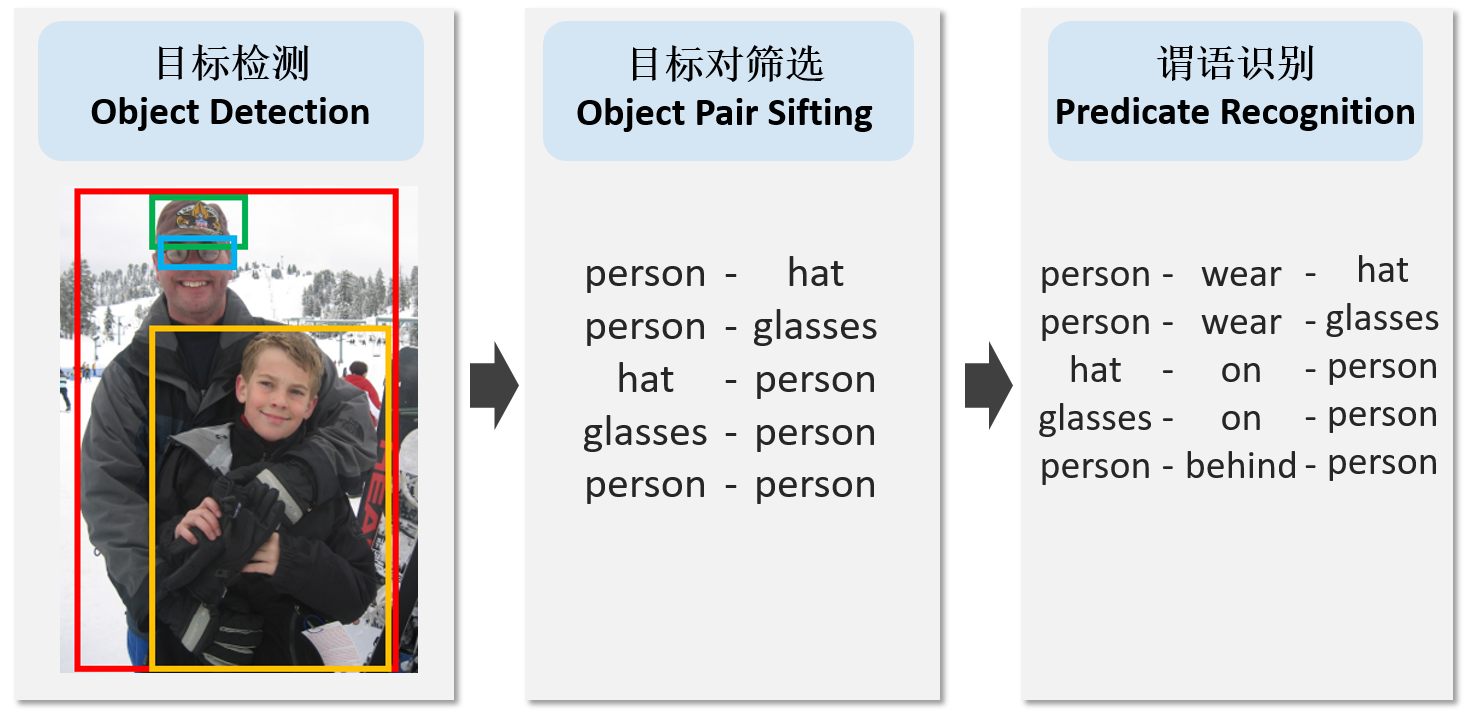


图 4.1 基于目标对筛选和联合谓语识别的视觉关系检测整体算法框架

1. 目标对筛选

得到所有检测出来目标后，假设目标的个数为*M*个，则可能的目标对数量为个，注意，目标不能和自己组合，且目标对中需要考虑目标的前后位置（因为分别对应主语和宾语）。

而事实上，这组目标对中，并不是所有的目标之间都存在关系，可能更一般的情况是，大多数目标对之间并不存在关系，而只有一小部分的目标之间存在关系。因此，在这个阶段，需要对目标对进行筛选，选出更可能存在关系的目标对。

在图4.1中，经过筛选的目标对为“person(man) - hat”、“person(man) - glasses”、“hat - person(man)”、“glasses - person(man)”、“person(man) – person(boy)”。而“hat - glasses”、“person(boy) - hat”、“person(boy) - glasses”等目标对之间并不存在明显的关系。

1. 谓语识别

得到了所有筛选后的目标对（<主语 - 宾语>）后，接下来要做的就是分别判断每对目标之间可以连接它们的谓语是什么。

该阶段的功能与第三章研究的问题相同，因此，把筛选后的主语、宾语、主语位置、宾语位置以及主语和宾语的联合区域图像，送入第三章提出的基于多特征联合统计预测的谓语识别算法中，即可以得到谓语的分类结果及置信度。

最后，只需要对所有得到的视觉关系的置信度进行从高到低的排序，即可以得到最终的视觉关系检测结果。

* + 1. 目标对筛选模型

有效的目标对筛选不但可以去除冗余的目标对，提升算法的效率，还可以减少谓语识别阶段的干扰信息，提高视觉关系检测的准确率。

目标对筛选可以抽象为一个二分类问题，一类代表存在关系，另一类代表不存在关系。而要判断一对目标之间是否可能存在关系，可以从两个角度出发。

一个是这两个目标之间的相对位置，如果两个目标之间离得很远，或者区域上没有什么交叠，那么它们之间则很可能不存在关系。

另一方面是两个目标的类别，比如“person（人）”和“coat（外套）”之间是很可能有关系的，但是“coat（外套）”和“cat（猫）”之间则更可能不存在关系。

同时，这两个方面也不是相互独立而存在的，很多时候需要综合考虑位置和目标类别两个方面，因为对于不同的目标对类别来说，存在关系时的位置特征可能是不一样的。比如对于“car（车）”和“wheel（轮子）”来说，它们之间存在关系的前提不仅仅是它们的位置要离得近，同时还需要所在位置的区域有所交叠才可以。而对于“car（车）”和“person（人）”来说，它们之间位置离得较近的时候就可能存在关系，即使位置的区域可能没有交叠，比如“car next to person（车在人旁边）”。

基于这个思路，本文的目标对筛选模型的结构示意图如图4.2所示。

模型的输入包括主语和宾语的位置信息，以及主语和宾语的类别信息。位置信息采用的是主语和宾语两个目标的图像掩模，原图中目标所在的区域值为1，其余区域值为0，由于有主语和宾语两个目标，故会提取两个图像掩模，然后把它们作为两个通道送入模型中。事实上，这里输入的位置信息与第三章谓语识别模型中位置模块的输入相同，详细信息可见3.2.2.2。

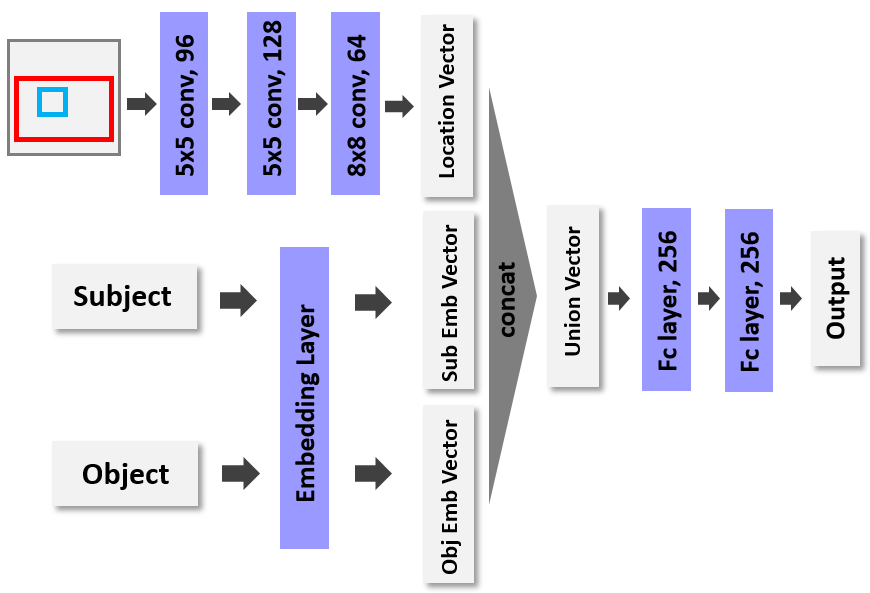


图 4.2 目标对筛选模型结构示意图

数据进入模型后，模型将分别提取位置向量及主语和宾语的语义向量。

位置向量由输入的位置信息通过三个卷积层后得到。这三个卷积层分别为96核的5×5卷积层、128核的5×5卷积层和64核的8×8卷积层，最终将得到一个64×1×1的Feature Map，将其展开即可得到一个长度为64的位置向量。

主语和宾语的语义向量由输入的主语宾语类别通过语义嵌入层后得到。语义嵌入层可以把主语和宾语的类别转换为一个300维的词向量来表示，这么做的好处是可以把较为稀疏的类别向量映射到更为稠密的实数向量空间中，并使其更能代表一定的语义信息。经过语义嵌入层后，最终将得到300维的主语语义向量和300维的宾语语义向量。

得到位置向量及主语和宾语的语义向量后，就可以把它们连接在一起，作为一个联合向量。再将其通过两个全连接层，即可以得到最终的二分类结果。

* 1. 实验与分析
     1. 数据集

本文使用的数据集为公开的视觉关系数据集VRD（Visual Relationship Dataset）。该数据集中包括5000张图片，100种目标类别和70种谓语类别。在这些图片中，总共有37993个关系样本，包括6672种<主语，谓语，宾语>三元组的关系种类。数据集被默认划分为两部分，一部分的4000张图片作为训练集，另一部分的1000张图片作为测试集。由于该数据集与第三章中使用的数据集相同，故数据集的具体介绍详见3.3.1节。

* + 1. 评价指标
       1. 目标对筛选评价指标

当训练好目标对筛选模型后，需要对该模型进行评测，因为该模型本质上为一个二分类模型，故使用的评价指标包括准确率（Accuracy, Acc）、精确度（Precision）和召回率（Recall）。

不妨把二分类中的两类设为正类和负类，正类即表示存在关系，负类表示不存在关系。那么根据实际的分类和预测的分类，就可以得到四种情况，如下表4.1所示

表 4.1 实际分类和预测分类的四种情况

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | 实际类别 | |
|  |  | *P* | *N* |
| 预测  类别 | *P’* | *TP*  (True Positive) | *FP*  (False Positive) |
| *N’* | *FN*  (False Negative) | *TN*  (True Negative) |

通过这张表，就可以定义三个评价指标。

Accuracy：所有样本中预测正确的比例，可以用如下公式表示

公式（4.3）

Precision：预测为正（即判断为有关系）的样本中，判断正确的比例，可以用如下公式表示

公式（4.4）

Recall：实际为正（即有关系）的样本中，被预测正确的比例，可以用如下公式表示

公式（4.5）

* + - 1. 视觉关系检测评价指标

本章的视觉关系检测评价指标延用了Lu等人提出的评价指标，即Recall@50和Recall@100这两个评价指标。Recall@K的定义为，检测出的所有<主语-谓语-宾语>三元组关系中，置信度排序最靠前的*K*个关系中检测正确的个数，与真实关系数量的比值。

该评价指标与第三章中使用的评价指标相同，故具体的计算公式详见3.3.2节。

* + 1. 实验设置

下面介绍模型训练过程中的一些实验设置。

模型参数上，使用了word2vec的预训练参数作为语义嵌入层的初始参数。

在训练的过程中，参数的优化方法使用的是随机梯度下降（Stochastic Gradient Descent, SGD），网络整体的学习率设置为0.001，语义嵌入层的学习率单独设置为0.0001.

训练框架使用的是Pytorch深度学习框架，并在单张NVIDIA GeForce GTX 1080Ti显卡上进行训练。

特别的，在构造数据集时对数据集中的负样本进行了采样。主要原因在于数据集中的正负样本比例十分不平衡，训练集中总共有199826组目标对，其中有关系的目标对仅有30355组，剩余的169471组都为负样本，故正负样本比仅为0.179，正负样本数量相差十分悬殊。

如果直接使用如此悬殊的样本数据进行训练，很可能造成模型学习了过多的负样本参数，以至于在模型预测的时候更多的偏向于负样本，而正样本的召回率偏低。因此，这里对负样本进行随机抽样，保证负样本和正样本数量之比在1.5倍以下，从而使正负样本更加平衡，以更好的训练模型。

* + 1. 结果分析
       1. 目标对筛选结果分析

对目标对筛选模型进行训练，并对最终选用的模型参数进行测试，该模型的结果如下表4.2所示

表 4.2 目标对筛选模型训练结果

|  |  |  |  |
| --- | --- | --- | --- |
| 评价指标 | Accuracy/% | Precision/% | Recall/% |
| 结果 | 81.79 | 39.51 | 70.31 |

如表4.2所示，目标对筛选的Accuracy尚可，为81.79%。但是Precision和Recall的差异却很大，相差了30%。

本文选取该模型结果的原因是，目标对筛选的作用旨在筛除掉很可能没有关系的目标对，以去除冗余的目标对、提升算法的效率，并且减少谓语识别阶段的干扰信息，提高视觉关系检测的准确率。

因此，为了不把可能存在关系的目标对误筛除，召回率Recall则显得比精确度Precision更加重要，故选用了Recall更高的模型结果来进行目标对筛选。

* + - 1. 视觉关系检测结果分析

为了更好的验证本章提出的视觉关系检测算法的效果，本文将把该视觉关系检测算法与现有的算法进行比较。参与比较的算法包括：

Visual Phrase[9]：把一组关系三元组看作一个整体，如“a person riding a horse”等，预测时不对主语、谓语、宾语进行切分。将每个关系三元组的整体作为不同分类来训练的代表做，并使用DPM检测器。

Joint-CNN[10]：一种具有*2N+K*种分类输出的神经网络，联动的判断主语、宾语和谓语的分类。

VR-L[10]：一种基于语言先验的视觉关系检测方法，通过外观模块把对象（包括主语和宾语）和谓语分别作为单独的模块进行训练，并把训练结果组合在一起用于预测最后的关系，再进一步通过语言模块把关系映射到向量空间中，并利用语言先验知识来进行约束和优化，最终得到预测出的关系。

VTransE[18]：一种视觉翻译嵌入网络，它通过将主语和宾语映射到共同的低维关系空间来对视觉关系建模，从而使谓语被认为是主语和宾语之间的翻译向量，即“主语+谓语≈宾语”。 并把知识表征学习的思想应用到其中，设计了一种全新的特征提取层，以全卷积的方式完成对象-关系的知识迁移。

VRL[19]：一种基于深度异结构强化学习框架的视觉关系及属性检测方法，它通过利用全局上下文线索来顺序检测关系和属性实例。

ViP-CNN[21]：一种视觉短语引导的卷积神经网络，它利用整个视觉短语来进行学习，具体做法是通过消息传递同时学习三个相互连接的识别问题（主语-谓语-并与），并利用该消息传递策略来减少参数数量。

DR-Net[11]：一种深度关系网络，它基于谓语、主语和宾语之间的统计依赖关系，并把统计模型通过条件随机场理论和深度学习结合起来。

为了更全面的进行比较，这里将视觉关系检测的三个子任务的效果一起进行比较，在表4.3中列出。

从表中可以看出，基于视觉短语的Visual Phrase算法效果最差，说明了在巨大的可选空间中，简单将不同的关系三元组作为不同的类别来进行训练和分类是一项很难的任务。从VR-L方法开始，把目标和谓语分别进行预测的算法，在各项指标上都首次有了极大幅度的提升。

在现有算法中，谓语识别任务效果最好的算法是基于统计依赖的DR-Net算法，它在谓语识别任务上相较于其他算法展现明显的优势，说明在判断谓语时加入统计依赖可以更好的帮助谓语的识别。

而短语检测任务中最好的算法是ViP-CNN，这是为数不多的基于视觉短语进行学习也能取得很好效果的算法，说明在同时识别主语、谓语和宾语时引入消息传递可以取得更好的效果。

在关系检测任务中，效果最好的是基于深度强化学习的VRL算法，说明强化学习可以更好的消除视觉关系检测中的歧义，以取得更加的效果。

表 4.3 本文视觉关系检测算法与现有算法在VRD数据集上对比

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 算法 | 谓语识别任务 | | 短语检测任务 | | 关系检测任务 | |
| R@50 | R@100 | R@50 | R@100 | R@50 | R@100 |
| Visual Phrase | 0.97 | 1.91 | 0.04 | 0.07 | - | - |
| Joint-CNN | 1.47 | 2.03 | 0.07 | 0.09 | 0.07 | 0.09 |
| VR-L | 47.87 | 47.87 | 16.17 | 17.03 | 13.86 | 14.70 |
| VTransE | - | - | 19.42 | 22.42 | 14.07 | 15.20 |
| VRL | - | - | 21.37 | 22.60 | 18.19 | 20.79 |
| ViP-CNN | - | - | 22.78 | 27.91 | 17.32 | 20.01 |
| DR-Net | 80.79 | 81.90 | 19.93 | 23.45 | 17.73 | 20.88 |
| 本文算法(不筛选) | **83.31** | **85.82** | 26.16 | 33.36 | 18.97 | 24.14 |
| 本文算法(筛选) | **-** | **-** | **26.59** | **34.37** | **19.43** | **24.93** |

而本文提出的算法，相较于以上所有算法在三个任务上都分别有不同程度的提升。在谓语识别任务中，相较于最好的DR-Net算法在Recall@50和Recall@100上分别有2.52%和3.92%的提升；在短语检测任务中，相较于最好的ViP-CNN算法在两个指标上分别有3.81%和6.46%的提升；在关系检测任务中，相较于最好的VRL算法分别有1.24%和4.14%的提升。

同时，对比本文算法不加目标对筛选和加入目标对筛选的结果，在各项指标上均有小幅提升，说明加入目标对筛选可以在提升算法的效率的同时，减少谓语识别阶段的干扰信息，一定程度上的提高视觉关系检测的准确率。

* 1. 本章小结

本章首先介绍了视觉关系检测任务的问题描述和定义，并根据问题的特点提出了基于目标对筛选和联合谓语识别的视觉关系检测。

接下来，本章详细介绍了该算法。先介绍了该算法的整体算法流程，包括算法中涉及的三个阶段——目标检测阶段、目标对筛选阶段和谓语识别阶段。再着重详细介绍了目标对筛选模型，对模型的输入输出和网络结构等进行了详细介绍。

最后是实验部分，首先介绍了实验使用的数据集及评价指标，接着介绍了实验的设置，最后分别对目标对筛选模型和视觉关系检测算法的结果进行了分析和比较。

# 总结与展望

* 1. 本文工作总结

计算机视觉的相关问题在最近几年成为了十分活跃的研究领域，在目标识别算法已经相对比较成熟的今天，进一步对图像内容进行理解和探索成了大势所趋，视觉关系检测应运而生。视觉关系检测的任务是从图片中找出用<主语-谓语-宾语>三元组表示的视觉关系，并标出其相应的位置。

由于可能组成的关系数量十分巨大，因此如何从有限的样本中学习出大量可能的关系成为了该任务的难点。近几年来，学者们围绕这个问题展开了许多研究，包括基于语言先验的方法、基于知识表征学习的方法和基于统计依赖的方法等。

本文针对视觉关系检测的第一个任务提出了一种基于多特征联合统计预测的谓语识别算法，并进一步针对视觉关系检测的任务二、三提出了一种基于目标对筛选和联合谓语识别的视觉关系检测算法。本文的主要工作包括：

1）本文首先分析视觉关系检测任务中存在的现象和现有方法的局限性。存在的现象包括同种谓语在不同场景下可能表现差异很大，且一幅图中可以组成的目标对数量可能远大于实际存在关系的目标对数量。另外，基于统计依赖的方法认为不同统计分量的贡献度是一样的，而忽略它们可能存在的重要程度差异。

2）提出一种基于多特征联合统计预测的谓语识别算法

视觉关系检测的第一个任务是在已知主语和宾语标签和位置的情况下，对谓语进行检测。针对同种谓语的表现差异可能较大的问题，本文提出一种基于多特征联合统计预测的谓语识别算法，同时考虑多种信息来共同进行预测，融合了视觉特征、位置特征和语义特征得到联合特征，再结合统计依赖关系进行联合预测。

在进行联合统计预测时，现有的基于统计的方法没有考虑到不同统计分量的重要程度，因此本文对不同分量的重要程度做出区分，把不同的分量作为不同的通道，并利用1×1卷积的通道融合功能将该思想融入到网络中。

本文还使用了交叉熵损失和中心损失联合训练的方法，旨在使不同谓语之间的特征差异更大，而同种谓语的特征更加相似。通过在公开数据集上的实验和对比，结果表明该方法的Recall@100指标可以达到85.82%，相较于现有的DR-Net算法有3.92%的提升，说明该方法可以得到更好的谓语识别结果。

3）提出了一种基于目标对筛选和联合谓语识别的视觉关系检测算法

视觉关系检测的任务二和任务三需要在一张图片中找出所有的视觉关系。由于检测出的目标之间很多可能都不存在关系，因此本文提出了一种基于目标对筛选和联合谓语识别的视觉关系检测算法的算法，把视觉关系检测分为三个阶段——目标检测、目标对筛选和谓语识别阶段。

先对输入的图片进行目标检测，再使用目标对筛选模型判断两个目标之间是否可能存在关系，目标对筛选模型主要从目标对的相对位置信息和语义信息两个方面出发来进行筛选。最后再利用之前提出的谓语识别算法对有关系的目标对进行谓语识别，以判断出它们之间是什么关系。

本文首先对目标对筛选模型进行实验和评估，再对整体的视觉关系检测算法进行实验，并在视觉关系检测公开数据集上进行测试，通过与现有方法的比较可以发现该算法可以取得一定的效果提升。

* 1. 未来工作展望

本文针对视觉关系检测现有的问题，从谓语识别和目标对筛选两个方面入手，对视觉关系检测方法做出了改进。但是在视觉关系检测中，仍然有很多可以改进的点，未来的研究工作可以从以下几个方面进行尝试：

（1）在本文的谓语识别模型中，融合各个不同模块得到的特征时，使用的是简单的把各个特征向量连接在一起再通过全连接层的方法。但事实上，在判断谓语的时候，不同的特征信息可能会起到不同的作用，可以考虑加入Attention机制对特征进行融合。

（2）在图像描述的任务中，以RNN为基础的网络得到了广泛的应用，而视觉关系也可以被看作一个有序序列，因此可以尝试把RNN等网络应用到视觉关系检测中。

# 参考文献

1. Ballard D H, Brown C M. Computer vision. 1982[J]. Prenice-Hall, Englewood Cliffs, NJ, 1982.
2. Klette R. Concise computer vision[M]. Springer, London, 2014.
3. Szeliski R. Computer vision: algorithms and applications[M]. Springer Science & Business Media, 2010.
4. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
5. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
6. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.
7. 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.
8. Farhadi A, Hejrati M, Sadeghi M A, et al. Every picture tells a story: Generating sentences from images[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2010: 15-29.
9. Farhadi A, Hejrati M, Sadeghi M A, et al. Every picture tells a story: Generating sentences from images[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2010: 15-29.
10. Lu C, Krishna R, Bernstein M, et al. Visual relationship detection with language priors[C]//European Conference on Computer Vision. Springer, Cham, 2016: 852-869.
11. Dai B, Zhang Y, Lin D. Detecting visual relationships with deep relational networks[C]//Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017: 3298-3308.
12. Galleguillos C, Rabinovich A, Belongie S. Object categorization using co-occurrence, location and appearance[C]//Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008: 1-8.
13. Gould S, Rodgers J, Cohen D, et al. Multi-class segmentation with relative location prior[J]. International Journal of Computer Vision, 2008, 80(3): 300-316.
14. Chao Y W, Wang Z, He Y, et al. Hico: A benchmark for recognizing human-object interactions in images[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1017-1025.
15. Yao B, Fei-Fei L. Modeling mutual context of object and human pose in human-object interaction activities[C]//Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 17-24.
16. Maji S, Bourdev L, Malik J. Action recognition from a distributed representation of pose and appearance[C]//Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2011: 3177-3184.
17. Sadeghi M A, Farhadi A. Recognition using visual phrases[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 1745-1752.
18. Zhang H, Kyaw Z, Chang S F, et al. Visual translation embedding network for visual relation detection[C]//CVPR. 2017, 1(2): 5.
19. Liang X, Lee L, Xing E P. Deep variation-structured reinforcement learning for visual relationship and attribute detection[C]//Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017: 4408-4417.
20. Liang K, Guo Y, Chang H, et al. Visual Relationship Detection with Deep Structural Ranking[J]. 2018.
21. Li Y, Ouyang W, Wang X, et al. Vip-cnn: Visual phrase guided convolutional neural network[C]//Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017: 7244-7253.
22. 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.
23. Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
24. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.
25. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
26. Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.
27. Le Q, Mikolov T. Distributed representations of sentences and documents[C]//International Conference on Machine Learning. 2014: 1188-1196.
28. Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method[J]. arXiv preprint arXiv:1402.3722, 2014.
29. Lin M, Chen Q, Yan S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.
30. Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition[C]//European Conference on Computer Vision. Springer, Cham, 2016: 499-515.