

中山大学硕士学位论文

基于消息传递机制的社会关系理解

**Social Relationship Understanding via Message
Passing Mechanism**

学 位 申 请 人: 李雷来

导 师 姓 名 及 职 称: 万海 副教授

专 业 名 称: 软件工程

答 辩 委 员 会 主 席(签 名): _____

答 辩 委 员 会 委 员(签 名): _____

二零一八年三月十四日

论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体，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。

学位论文作者签名: _____

日期: _____

学位论文使用授权声明

本人完全了解中山大学有关保留、使用学位论文的规定，即：学校有权保留学位论文并向国家主管部门或其指定机构递交论文的电子版和纸质版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆、院系资料室被查阅，有权将学位论文的内容编入有关数据库进行检索，可以采用复印、缩印或其他方法保存学位论文。

学位论文作者签名:

导师签名:

日期: 年 月 日

日期: 年 月 日

论文题目： 基于消息传递机制的社会关系理解

专 业： 软件工程

硕 士 生： 李雷来

指导老师： 万海 副教授

摘要

本文致力于研究社会关系理解问题，社会关系理解是为了推断出一个给定场景中人之间的社会关系。近来关系理解在计算机视觉领域受到极大的关注，任务的效果也随着深度学习方法的发展得到了快速的提高，但是现有的工作主要通过挖掘人对的图像基本特征，或者引入物体和关系共现频率的先验知识来提升结果。这些工作将图片中的每个人对的关系检测独立的看待，并没有考虑到这些人对之间的相互关系。因此，在社会关系理解人物中很自然的考虑到这样的交互信息。例如，如果一张图片两个人对是朋友，那么第三个人对的关系往往是朋友或至少是其他亲密的社会关系，而不是无关系。因此，为了捕捉到这样交互的线索，本文提出一个端到端的可训练的人对关系网络，采用RNNs来实现人对之间的消息传递达到推理的目的，进而提高关系的分类结果。

在PPRN中，本文提出了一个消息传递和消息池化模块来实现不同人对之间的信息传递，达到不同人对间关系互相约束的目的，之后再实现了一个基于注意力机制来结合周边物体特征的模块。在这个过程中，首先进行的是人对关系之间的消息传递和池化这两个模块不断迭代，再融合区域生成网络生成的物体区域的图像特征，最后进行分类优化。

在实验中，本文在两个大规模数据集中验证了PPRN模型的有效性，这两个公开数据集包括三个不同的关系粒度，接着进行了具体案例的分析。实现结果表明了PPRN模型在与其他基准模型的对比中取得了最优的结果，同时说明了在视觉关系理解任务中考虑不同人对间相互影响的重要性。

关键词： 社会关系理解，消息传递，注意力机制，神经网络

Title: Social Relationship Understanding via Message Passing Mechanism
Major: Computer Technology
Name: Leilai Li
Supervisor: Prof. Hai Wan

Abstract

This paper focuses on social relationships understanding which aims at inferring the social relations among people in a given scene. Relationship Understanding has attracted increasing attention in computer vision recently. Great progress has been made since the rise of deep learning. However, previous works mainly improves the results by mining the basic features of person pair or introducing prior knowledge of object and relationship co-occurrence frequencies, without taking into account the interaction of different pairs. It is natural to consider these interaction cues, *i.e.*, the mutual influence of multiple person pairs, in social relationship understanding. For instance, if two person pairs in an image are *Friends*, then the third pair is always *Friends* or at least other similar relations but not *No Relation*. Therefore, to capture these interaction cues, we propose a novel end-to-end trainable Person-Pair Relation Network (PPRN) using standard RNNs, a inference network that learns iteratively to improve its predictions via message passing among person pair nodes.

In PPRN model, we provide a message passing and message pooling module to implements the message passing between various person pairs, achieving the purpose of the mutual restraint between different person pair, and we also implements a attention module to combine the contextual object feature. In this process, the first step is to iterate the two modules of message passing and pooling between person pair's relationships, and then combine the image features of object bounding box generated by the region proposal network, and finally optimization.

In the experiments, we evaluate our model in two large-scale datasets, and the two datasets contain three relational granularity, and further analyze by case stud-

ies. Experimental results demonstrate that our model outperforms baselines, which justifies the significance of considering the interaction between various person pairs in social relationship understanding.

Keywords: social relationship understanding, message passing, attention mechanism, neural network

目 录

第 1 章 引言	1
1.1 研究背景和意义	1
1.2 研究现状	5
1.3 本文工作	7
1.4 论文结构	8
第 2 章 预备知识	10
2.1 图像的视觉信息抽取	10
2.2 物体检测与识别	18
2.3 社会关系检测	20
2.4 本章小结	24
第 3 章 基于消息传递的人对间关系网络	26
3.1 基本框架	26
3.2 本章小结	34
第 4 章 实验设计与分析	37
4.1 数据集	37
4.2 实验设置	39
4.3 PISC数据集实验结果	40
4.4 PIPA-relation数据集实验结果	42
4.5 实验结果分析	43
4.6 案例研究	44
4.7 本章小结	45
参考文献	47
致 谢	51

第1章 引言

1.1 研究背景和意义

每个人的社会关系从构成了我们日常生活中社会结构的基础。自然的，我们利用一个人所在场景的社会关系来理解和解释当前的场景。社会学研究表明，这种对人的社会理解允许对其特征和可能的行为进行推断。当前，我们的社交生活很大部分是在社交媒体上，例如Facebook、Twitter、微信和微博等包含多模态信息的App，人们会通过文字、视频和音频等媒介含蓄的留下一些痕迹，但是我们能明确的捕捉到他们的社会关系通过分析多模态的信息。随着科技的发展和未来的到来，智能和潜在的自主系统会成为我们的帮手和同事，我们希望它们不仅可以熟练的完成任务，还希望他们能够融入和在我们人类生活的不同情况下采取适当的行动。此外，通过更好地了解这些隐藏信息，我们希望告知用户潜在的隐私风险。理解社会关系也有助于避免潜在的隐私风险，通过自动分析可能在文本等许多媒体中揭示社会关系的信息并告知用户这一点。在这个模式中，任务要求社会关系的概念和模式需要在生活中的所有方面共同努力，以便从一种感觉到的输入。虽然已经开始努力解决这一具有挑战性的问题，但社会生活的巨大多样性和复杂性阻碍了进展。最常见的，识别社会关系的计算模型仅仅限定于少数特定的类别。

在图像理解任务上，视觉概念识别获得了越来越多的研究者的关注，包括视觉属性和视觉关系^[1]。视觉关系和视觉属性检测的主要目的是构建场景图谱，场景图谱（scene graph）^[2]是对图片进行描述的一种半结构化的形式，场景图谱是由视觉三元组构成，并且包括关系三元组和属性三元组。场景图谱已经成为计算机视觉和人工智能领域的重要基础资源，因此如何自动的构建场景图谱成为了重要关注点，以利用自然语言信息的^[1]为代表的工作，代表场景图谱自动生成领域取得了极大的进展。同样，社会属性和社会关系^[3]对于场景理解同样重要。因此在当前工作，主要聚焦在解决社会关系检测问题上，并且可以从场景图谱

的生成借鉴有用的思想。给定一张图片，社会关系理解的目的是推断在当前图片这个场景下人之间的社会关系是社会关系检测的准确描述。除了前面提的用处，理解图像场景中这样的关系能帮助现有的算法产生更好的场景描述。例如在图1-1中的第一个样例，用正常的文字来描述的话，“一个妇人和女孩正在吃饭”。但是对于社会关系的这个问题下，可以认为是“一个母亲和女儿正在吃饭”。



图 1-1 PISC数据集中的一些图片例子

既然社会关系理解对于提升上述任务的关键资源，那么自然而然的，如何准确的理解社会关系成为需要研究者需要攻克的课题。一方面，一张图片的社会关系可以通过众包的方式，人工标注得到，比如现有的数据集PISC^[4]和PISC-relation^[5]。当然，自动端到端的方法包括基于人脸特征、年龄、人的头部特征等特征信息的^[5,6]。还有利用周边环境的信息的模型^[4,7]，这些模型通常需要一个物体检测器或者检测器中RPN（region proposal network），这都是需要引入额外的标注框或者预训练模型。也有通过对周边物体和社会关系共现的统计，例如“computer”和“professional”共同出现的概率较大，如果识别出存在“computer”，那么当前的关系很大概率是“professional”，通过神经网络引入这些先验知识来提升预测的准确率。这些自动识别社会关系的模型虽然不断在进步，但是从实验结

果来看，他们与人工标注的准确率还是存在很大鸿沟，离实际的应用还存在很大的距离。

然后，现有的学习模型大都倾向于利用外部的知识来辅助理解图片的社会关系场景，但是得到这些外部知识需要额外的人工标注，这是一件耗时耗力的工作，或者一些统计得到的先验知识同样包含一些噪音，这也直接引出了到底是否应该引入外部知识，例如是否利用周边物体的信息，以及如何在缺乏这些信息的情况下取得好的实验效果。受到场景图谱生成的启发，场景图谱的概念最初是在2015年由Johnson 等人^[2]提出的，是用于描述图片的一种新的半结构化的方式，基本组成单位是视觉三元组，形式为（头实体，关系，尾实体）。受到该领域下xu (2017)^[8]的工作首先将整张图片输入，考虑到图片中不同视觉三元组之间的相互影响。例如，当知道“马在草地上”倾向于提高检测到“人骑着马”这条视觉三元组。对于社会关系检测的场景，如果图片中包含三个人对，其中两个人对的社会关系是“朋友”，那么第三条关系的社会关系会倾向也是“朋友”或者其他亲密关系，而不是“无关系”。直观上来说，这个是成立的，因此我们可以利用这当前场景下的其他的关系的来推理出当前的关系。

本论文主要研究如何将前文提到的关系场景的上下文信息引入社会关系理解的框架中。本文的出发点完全区别于Li (2017)^[4]和Wang (2018)^[7]的工作，对于关系特征的提取方法采用和Li^[4]相同的策略。论文的切入点如图例1-2，图上六个人对的关系有五对是“朋友”关系，其中只有一对“奶奶-孙女”的关系，因此当前的图片应该是一个朋友聚会的场景拍下的。如果我们想推理出其中一个人对的关系并且已经知道其他部分人对的关系，那么直观的，我们会通过对已经知道的关系进行一个场景的判断，从而推理当前人对的关系到底是什么。在当前例子中，如果已经知道了2对或者3对都是“朋友”关系，那么当前的人对大概率也是“朋友”关系。因此，类似于前文提到的场景图谱的生成，以及现有的社会关系理解的研究现状，将当前的工作成为社会关系图谱生成 (social graph generation)。

在上述的介绍中，我们分别提到了两方面的相关内容，一方面社会关系理解的意义和作用，另外提到了与社会关系同属视觉理解领域的场景图谱的生成，收到这些工作的影响，可以列出他们的共性和特性如下：

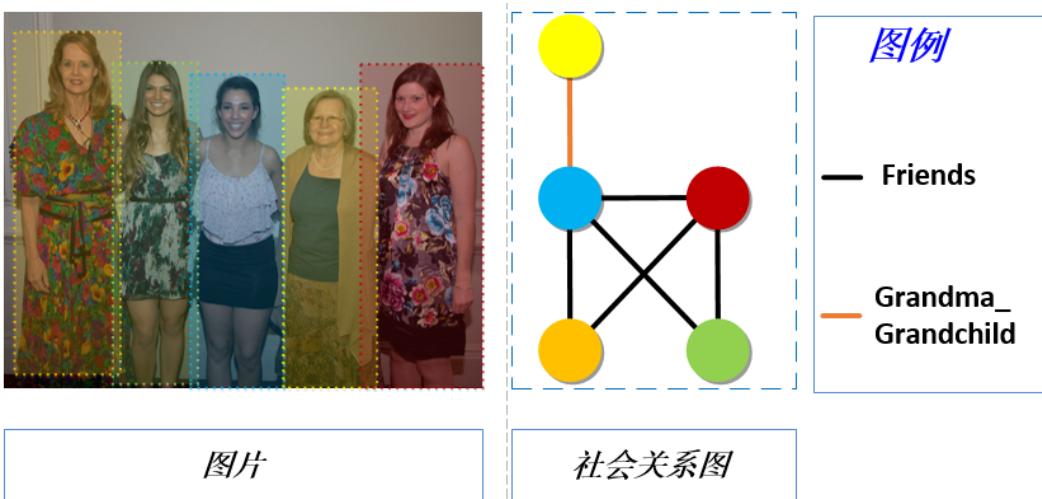


图 1-2 本论文动机的示例图，该图片来自PIPA-relation数据集，其中图片中对应阴影颜色的人对应社会关系图的部分，图上节点间的边表示他们之间的社会关系

- (1) 场景图谱的基本组成是视觉三元组，社会关系图中是人对和人对间的关系，但是场景图谱中并没有人的类别的概念，社会关系图中节点间的关系与人的类别无关。
- (2) 场景图谱中的关系类别较多，有80-100个类别，但是在社会关系中，现有数据集不同粒度的关系类别分别为3、6、16，数量上远远不一样。并且在场景图谱中，关系的类型主要以空间关系为主，少量含有语义的关系，但是在社会关系图中，除了“无关系”和空间存在较大关联，其他的均为语义的关系。
- (3) 与现有研究工作的区别是，之前的方法均将同一张图片上的不同人之间的关系割裂来看，但是他们间的关系互相影响，现有的研究工作忽略了这一点。

要想解决社会关系理解问题，一种可行的方法是借助场景中除了人以外其他的信息，由于现有的数据集并没有标注其中的物体信息，所以需要借助额外的检测模型，但是由于模型的准确率的原因，会引入相当一部分的噪音，我们不能简单的加入这些信息，或者说我们是否需要加入这部分信息。其次是借助场景图谱生成的思想，认为一张图片中所有人的社会关系不是割裂开的，是一同生成的，并且它们之间是相互影响的，但是由于场景图谱和社会关系图的区别，我们需要设计一个在社会关系理解人物下人对关系之间的交互机制。

1.2 研究现状

1.2.1 视觉关系理解的应用

在计算机视觉领域，社会关系信息被探索来提升几个常见的任务，例如人的轨迹预测^[9,1]、多目标追踪^[10,11]和群体活动识别^[12-14]。例如，在Deng等人（^[15]）群体活动识别的任务中，群体活动识别需要推理出图上人之间的结构信息，当前的方法是判断每个个体的动作，并且判断图上人之间的关系。但是由于图片特征的复杂和不确定性，这两个任务都是很有挑战的，推断出图上的结构信息能帮助排除一些未参与群体活动的人，得到更好的预测结果。因此预测这些人之间的社会关系能有助于群体活动识别。如图例1-3所示，利用深度学习模型得到的人的表征和场景的表征后，如果知道了图例中第三个人和另外两个人之间不存在关系的情况下，排除第三个人对任务判断的影响。能更为容易的得出当前的群体活动场景是“waiting”。如Alahi等人（2016）^[16]隐含的引入了社会性的约束来预测符合社会常识规则的人类轨迹。

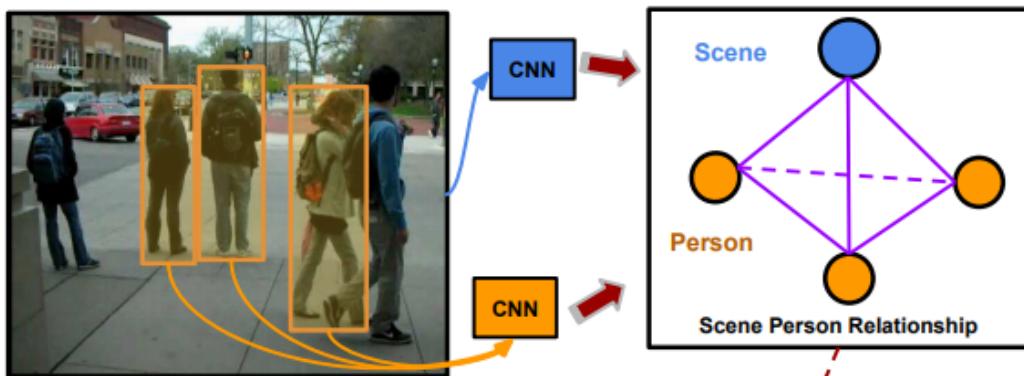


图 1-3 来自Deng等人^[15]的示例图

前面的工作大多数都是利用向量化的社会关系来帮助推理，与社会关系理解同属同一个视觉关系理解任务类别，即场景图谱的生成，又或者称为视觉关系检测。该任务的最早是Johnson等人于2015年提出的^[2]，是一种用于描述图片场景的新的方式，与本文工作不同的是，场景图谱中的主要组成部分不仅包括人，还包括很多日常物体，如图??所示。两个工作最核心的挑战在于推理出物体之间的关系。由于场景图谱常用于图片检索^[2]，Marino（2017）^[17]利用得到场景

图谱结合图神经网络提高图片的分类效果。在以上关于视觉信息理解的两个方向上，都有大量的工作提出，分别用于解决不同场景的问题，因此如何更好的实现视觉理解这个难题自然而然的成为许多研究者想要攻克的难题。对于场景图谱来说，本文的工作主要受到了以下几个工作的启发。Xu等人（2017）对于场景图谱进行建模，分别包含物体节点和关系节点，然后采用Message Passing的方式进行迭代。通过相邻的节点或边对目标节点或边进行约束，从而对这些特征进行微调，达到提升的效果。Yang等人（2018）利用图卷积网络进行不同节点之间的Message passing来达到约束的效果。融合自然语言知识提出的VRD模型^[18]，本文初始步骤中的特征提取工作中包含空间信息的提取便收到该工作的启发。

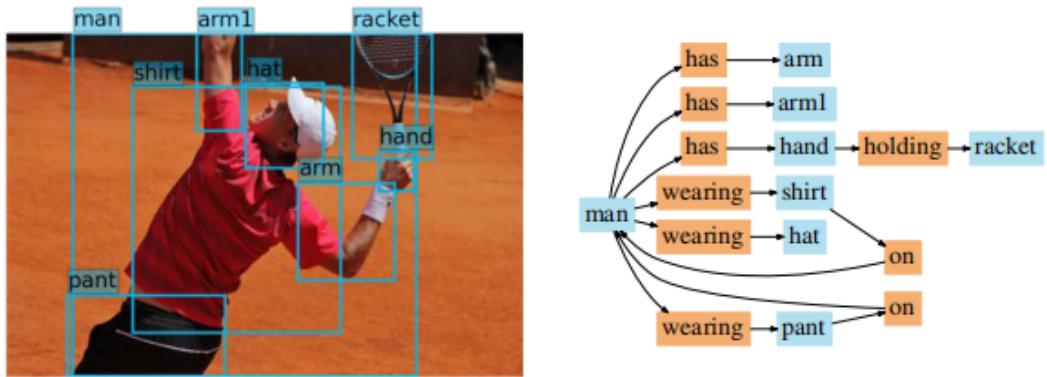


图 1-4 场景图谱^[8]的示例图

1.2.2 社会关系理解的方法

而对于社会关系理解，作为最早的工作，wang等人（2010）开始引入家庭关系当前背景来识别别人之间亲属关系。在后来的工作中^[10,19,20]，为了捕获这些社会关系展现出来的一些规律，探索了面部表情和属性等用于亲属关系识别和验证。并且为了促进社会关系理解领域的研究和发展，Li^[4]和Sun^[5]构建了大规模的数据集，并且利用深度学习的模型直接从图片中学习来识别关系。对于Sun构建的数据集PIPA-Relation，该数据集的关系包括5个领域，然后基于这5个领域又划分为16条关系。同样，Li基于关系模型理论，定义了一系列的关系类别，包含两个不同层次关系的划分，粗粒度的3类关系和细粒度的6类关系。在Sun（2017）等人的工作中，不仅提出了两个关系粒度的数据集，而且提出了Dual-glance模型。该模型是一个流水线的模型，并且包括两个模块，主要的

的创新点在于利用人们对周边的物体信息来提升关系检测的效果。Wang (2018) 等人提出了基于知识图谱的深度推理模型,通过图门控神经网络引入物体的社会关系的共现的常识知识来实现物体和关系之间的推理,本质上还是利用周边的物体信息,但是在Dual-glance中,用到只是周边的物体区域,并不需要准确的识别出物体区域的物体类别。

在关系理解领域,除了社会关系理解,还有视觉关系检测任务。Lu等人^[1]提出首先提出VRD模型,利用自然语言得到的词向量作为先验知识来帮助单条三元组的检测。后来, Zhang等人^[2]提出VTransE 模型,模拟知识图谱中关系翻译,即 $h + r = t$,本质上还是一个关系的分类器。但是并没有考虑到视觉关系的复杂度,并且不同于自然语言的状态翻译。Li (2017) 等人^[3]进一步提出了多任务混合模型,任务包括场景图谱、区域描述生成、物体识别。和VRD类似,通过多种任务联合训练,引入额外的信息。Zellers等人 (2018)^[4]提出motif-net, 通过分析数据集得出关视觉关系严重依赖头尾实体的类别,利用双向循环神经网络对物体类别编码信息进行编码处理,提升了物体识别的准确度,进而提升了任务的效果。Yang 等人^[5]提出采用图卷积神经网络的方法来实现物体和关系的相互约束,一次生成一张图片的所有视觉三元组。

1.2.3 研究现状小结

当前社会关系理解领域的工作主要方式是引入额外的信息,例如引入面部表情和属性、年龄等。以及通过物体检测的方法识别出当前场景中的物体,来优化单纯通过提取人的区域的表征。最新的工作尝试引常识知识,对物体和关系间的常识知识进行表征来提高模型的效果。无论是采用额外的物体检测模型还是通过引入常识知识,都是外部信息,不可避免的需要额外的消耗或引入一些噪声。同时,对于人的轨迹预测、多目标追踪和群体活动识别任务,引入社会关系的信息能有效的提高这些任务的效果。因此,如何利用更少的信息,来提升预测效果是当前的一大挑战。同时以消息传递机制为技术栈的微调机制的方法在场景图谱,同属视觉理解领域得到了运用,在一定程度上解决了场景图谱的生成问题。然而,由于场景图谱和社会关系任务理解两个不同人物上的区别,如何在社会关

系理解中引入消息传递机制仍然面临很多的挑战

1.3 本文工作

本文首先通过介绍现有的社会关系理解最新研究工作，分析它们的模型设计的出发点，模型的结构，分析这些工作的忽略的信息，即没考虑到整张图片不同的人对的关系之间的互相影响。因此，本文提出了一个考虑到同一张图片不同关系之间存在交互的模型：包含人对消息传递机制的模型，人对关系网络（Person-pair Relation Network），简称PPRN，最后把该模型应用到社会关系理解的任务中，在两个公开数据集上做了对比实验。本文提出的PPRN模型包括以下三个模块：

- (1) 特征抽取模块，对于特征抽取模块，本文采用一个ResNet101^[21]抽取两个人的单独区域的特征，同时另外一个ResNet101抽取人对区域的特征。同时，对于这个模块，人对的位置信息也加入了最后的表征向量中。
- (2) 消息传递模块，本文首次尝试在社会关系理解任务中引入多个人对之间的关系的想法。对于消息传递模块，利用以GRU单元为组件的RNN来实现消息传递，并且设置多个迭代步，实现消息传递和池化模块的交互。采用RNN 最后隐藏层的输出作为图片中人对的关系表征。
- (3) 消息池化模块，本文设计了一个交互机制来处理人对之前的关系的信息，同时利用注意力机制来提高模型的表现效果。

在实验中，本文在两个公开数据集上验证了**PPRN**模型的有效性，它们分别是PIPA-Relation和PISC，其中PISC包括两个粒度的子数据集PISC-coarse和PISC-fine。接着，本文进一步通过案例研究的方法分析了**PPRN**模型的具体效果。从实验结果可以看出，**PPRN**模型在与其他基准模型的比较中取得了在两个数据集上取得了最优的结果，说明了引入不同关系之间交互的消息传递机制在社会关系理解为基础场景中的重要性。

1.4 论文结构

全文的组织结构描述如下：

第1章：介绍了社会关系理解的相关背景，点明了当前社会关系理解存在的缺点，由此引入了同属视觉关系理解人物的场景图谱，分析并总结了场景图谱的研究现状，并针对两者的共同点引出社会关系理解的提升方案。

第2章：首先介绍了图像领域的视觉信息抽取的方法，从简单的神经网络、卷积神经网络、以及残差网络等介绍。之后基于前面的神经网络的知识，介绍了现有工作中常用的物体检测与识别的方法，并进行了讨论与对比。然后详细介绍了当前社会关系理解模型的各个部件，最后对本章内容进行了总结。

第3章：针对社会关系理解的特点，提出了基于消息传递机制的社会关系理解方法，设计了图像中不同人对的社会关系交互的PPRN模型，并且具体介绍了模型的细节。同时实现了结合周边物体信息的模块，分析了模型的设计原则和具体细节。

第4章：首先介绍了两个在社会关系理解领域常用的数据集，并对PPRN模型在不同关系粒度的数据集上进行了训练和测试，接着针对实验结果进行了分析。然后还进一步通过案例研究的方式分析了PPRN在社会关系理解的变现。

第5章：总结了本文的研究工作，并且提出了进一步的研究展望。

第2章 预备知识

2.1 图像的视觉信息抽取

2.1.1 一般神经网络

神经网络 (neural network) 的方面的研究就出现了，早起的神经网络主要是指生物学中的“生物神经网络”，在当前计算机领域特指“神经网络学习”。神经网络最基本的结构是神经元模型，神经元模型如图2-1，在这个模型中包括输入端、神经元权重 (weight) 、偏差 (bias) 、激活函数 (activation function) 、阀值、输出。在这个模型中，当前神经元接收其它 n 个神经元的输入，与连接权重相乘之后加上偏差，然后激活函数的处理得到激活值输出。理想中的激活函数

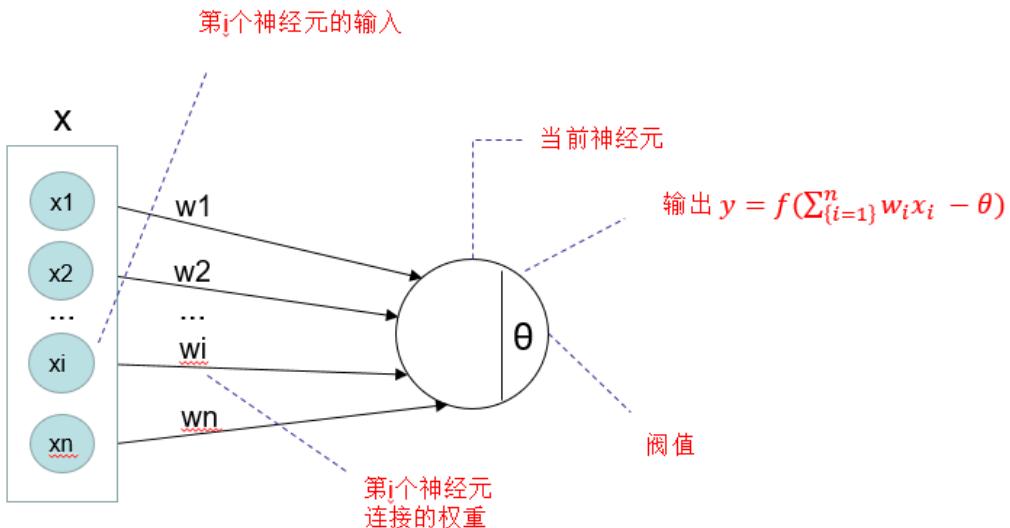


图 2-1 神经元模型示意图

是如公式2-1，将输入值映射为输出值“1”或者“0”，其中“1”对应神经元兴奋，“0”对应于神经元抑制，但是因为该阶跃函数具有不连续、不光滑等性质。因此常采用Sigmoid函数作为激活函数，一般来说激活函数是非线性的、可微的。如果不使用线性激活函数，采用线性激活函数（恒等激活函数，图例中 $f(x) = x$ ）的话，

那么神经网络只是把输入线性组合再输出，和没有采用神经网络是一样的。

$$sgn(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (2-1)$$

例如sigmoid函数可以把 $\Phi - \infty, +\infty$ 输入值映射到(0,1)区间内。常见的激活函数还有tanh(hyperbolic tangent, tanh)函数、修正线性单元(rectified linear units, ReLU)函数等等。表2-1详细的列出了3个常用激活函数的原函数、一阶导数、以及函数的值域。其中tanh函数只是sigmoid函数向下平移再拉升的结果。并且在实际应用中，tanh的效果是好于sigmoid，因为tanh的函数值域是属于(-1,1)的，使用tanh代替sigmoid，会使得神经元输出的均值趋近于0而不是0.5，这样的结果会使得下一层的学习变得更加简单。但是对于多层的神经网络，sigmoid和tanh在极大或极小时梯度会趋近于0，会造成梯度弥散问题。但是对于ReLU来说，当小于0时，梯度是小于0的，当大于0时，梯度是常数。ReLU激活函数在实际训练中取得了良好的效果。但是对于小于0的部分，此时的梯度为0，神经元不会训练，因此研究者们提出了LeakyReLU等激活函数解决这一问题。

表 2-1 常见激活函数的介绍

函数名称	原函数	一阶导数 $f'(x)$	原函数值域
sigmoid函数	$\sigma(x) = \frac{1}{1+e^{-x}}$	$f'(x) = \sigma(x)(1-\sigma(x))$	(0,1)
tanh函数	$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	$f'(x) = 1 - \tanh^2(x)$	(-1,1)
ReLU函数	$relu(x) = max(0, x)$	$sgn(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$	[0, $+\infty$)

前面介绍了神经元模型和各类激活函数，这里需要提到是深度神经网络、目标函数、优化方法。相比前面的单层神经网络，更常见的如图2-2所示的包含多个层级结构的神经网络，又称为“多层前馈神经网络”（multi-layer feed forward neural networks），其中神经元同一层之间不存在连接，跨层的神经元之间也不存在连接。就如图2-2所示，其中输入层的神经元接收外接的输入，隐层与输出层的神经元对输入的数据进行处理，这里的网络包含两个隐藏层（hidden

layer) 和一个输出层 (output layer)，假设输入为 x ，那么该网络可以形似化为 $H_\theta(x) = f_3(w_3f_2(w_2(f_1(w_1x + b_1)) + b_2) + b_3)$ ，其中 f_1, f_2 分别是隐藏层的激活函数， f_3 是输出层的激活函数。假设对于当前任务的目标函数 (object function) 如2-2，需要最小化目标函数 $L_{(\mathbf{X}, \mathbf{Y})}$ 的值，对于训练的过程来说，就是不断接受输入层的 $\Phi x, y\Psi$ ， y 是样本的标签，随着 x 的不断输入，不断调整网络的连接权重 w_1, w_2, w_3 。常用的优化算法包括随机梯度下降法 (stochastic gradient descent, SGD) 来迭代优化连接权重，对于一条样本 x_i, y_i ，其连接权重的更新方法如。

$$L(\mathbf{X}, \mathbf{Y}) = \sum_{(x_i, y_i) \in (\mathbf{X}, \mathbf{Y})} (y_i - H_x(x_i)) \quad (2-2)$$

不同层之间的梯度通过误差逆传递 (error Back Propagation 简称BP) [22]算法进

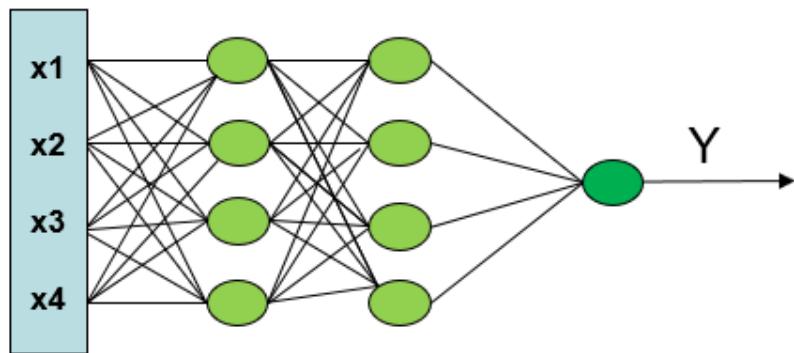


图 2-2 两个隐藏层的神经网络示意图

行整个网络的学习。SGD因为更新比较频繁会造成损失函数动荡，最终停留在局部最小值或鞍点。之后又新衍生出的包括Monmentum、Adagrad^[23]和Adam^[24]，这些算法能减少迭代的轮数，训练速度更快的收敛到最优值。SGD的梯度更新方法如公式所示，找出参数的梯度，然后往梯度的方法去更新参数。

$$W_{t+1} \leftarrow W_t - \alpha \Delta g_t \quad (2-3)$$

对于SGD with Monmentum来说，在梯度下降的过程中加入了惯性，使得梯度方向不变的维度上速度加快，梯度方向有改变的维度上速度减慢。其更新公式

如2-4, β_1 经验值为0.9, g_t 为在t时刻的梯度。这样便可以加快收敛减少震荡。

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t \\ W_{t+1} &\leftarrow W_t - \alpha m_t \end{aligned} \quad (2-4)$$

对于Adagrad来说，其梯度更新规则为公式2-5。对于优化器来说，学习率 α 非常重要，太小会花费太多时间学习，太大容易过拟合，无法正确学习。Adagrad便是根据梯度来动态的调整学习率，Ada即Adaptive的意思。前期梯度较大的时候能够约束学习率，梯度比较小的时候能够方法学习率。

$$\begin{aligned} m_t &= \sum_{\tau=1}^t g_{\tau}^2 \\ W_{t+1} &\leftarrow W_t - \alpha \frac{1}{\sqrt{m_t + \epsilon}} g_t \end{aligned} \quad (2-5)$$

而Adam可以说是前面Monmentum和Adagrad的结合，如公式2-6所示，Adam既保留了Momentum对梯度方向的调整策略，同时也加上了对过去梯度的平方值做学习率的调整。

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \\ W_{t+1} &\leftarrow W_t - \alpha \frac{1}{\sqrt{v_t + \epsilon}} m_t \end{aligned} \quad (2-6)$$

2.1.2 卷积神经网络的介绍

卷积神经网络（Convolutional Neural Network,CNN）是深度学习的代表算法之一，由LeCun首次实现并且应用。卷积神经网络的主要作用是提取特征，该网络受到生物学的影响，相比较与全连接神经网络，卷积神经网络的主要特性包括局部感知和参数共享。局部感知指对于具有空间特征的输入来说，每个神经元没必要知道全局的信息，只需要感知局部的信息，然后在更高层将局部的信息合并起来得到更高层的信息。对于权值共享来说，每个卷积核与位置无关，因为假设对于图像来说，其中某一部分的统计特性和其它的部分是一样的，所以对于其中的一个卷积核来说，可以应用到图像上的任何地方去。所以，局部感知和参数共享不仅能提取到更多的特征，并且能大幅度减少参数的数量。因此，

卷积神经网络广泛的应用在图像、视频、音频和文本等各种模态的数据上，并且都取得了巨大的成功。

卷积神经网络的特征提取层主要包括两个模块，分别是卷积层（convolutional layer）和池化层（pooling layer），两者的顺序，一般是先通过卷积层，然后是池化层。对于卷积层，主要的作用是提取特征，卷积层的核心是卷积核（kernel），其本质还是神经元。但是卷积核的感受野和全连接的神经元是不同的，这里的感受野是局部的，并且感受野的大小由卷积核的大小控制。如图??所示，当前卷积核的大小是 4×4 的，对于输入的图片 $6 \times 6 \times 3$ ，其中图片输入的3为图片的通道数、 6×6 为高宽，假设滑动的步长为1，卷积核通过在输入图片上按照步长进行滑动并且进行对应位置的点乘运算，最后形成一个 4×4 的特征图。以上综合起来就是卷积操作，其中 3×3 就是网络的参数。按照惯例，输入的图片可以有固定的高宽和通道数时，卷积核可以有不同的高宽，但是必须是固定的通道数，这里一般和输入的通道数一致。有多少个卷积核，最后就能得到多少个特征图（feature map）。

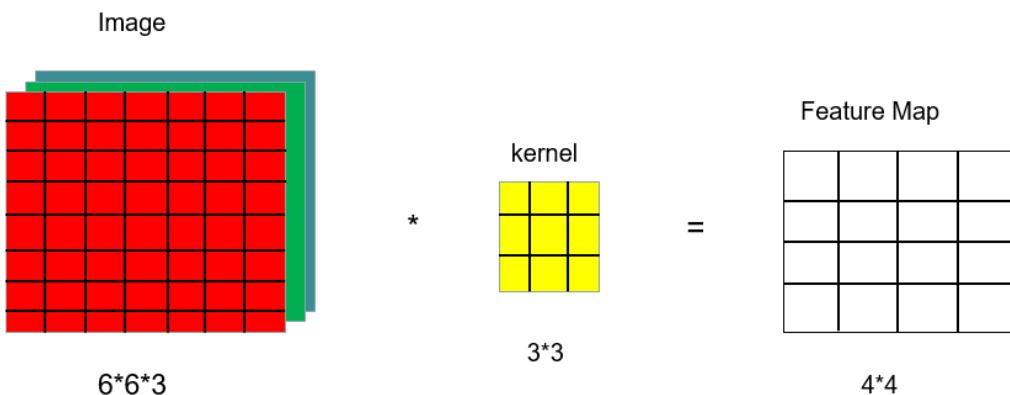


图 2-3 卷积神经网络的卷积层示意图

对于池化层来说，主要的作用是对于卷积层输出的特征图提取主要特征，降低网络的参数，且有防止过拟合的作用。常见的池化包括平均池化(Average pooling)和最大池化(Max pooling)。具体细节如图2-4所示，池化也是通过类似卷积的操作实现的，在图例中，池化也是以 2×2 在特征图上进行滑动，滑动的步长为2，而最大池化是选着窗口中的最大值作为输出，平均池化是选择窗口中所有值的平均值进行输出，假设输入的特征图为 $C \times W \times H$ ，那么经过如图例所示

的操作后得到的特征图为 $C \times \frac{W}{2} \times \frac{H}{2}$ 。



图 2-4 卷积神经网络的池化层示意图

综合上述对于卷积神经网络的卷积和池化的介绍，因为每层的输入和输出都表现为特征图的形式，因此卷积神经网络可以和全连接的网络一样可以有多层，并且取得更好的效果。LeNet-5^[25]是Yang LeCun等人在1988年提出的，它是第一个成功应用于数字识别问题的卷积神经网络，在著名的MINIST数据集上，LeNet-5可以取得大约99.2%的准确率。LeNet-5是一个经典的卷积神经网络，前5层分别是卷积层和池化层，后2层全连接层。之后于2012年提出的AlexNet^[26]，其网络结构如图2-5首次使用Relu激活函数替代Sigmoid，并且验证了其在较深网络上的作用，成功解决了Sigmoid在较深网络的梯度弥散问题，虽然Relu很早就提出了。其次，AlexNet首次在训练中使用dropout层抑制一部分激活的神经元，以避免过拟合，并且通过实践证明了效果。与此同时，模型还采用了数据增强等trick来防止过拟合，使用cuda提高训练速度。而之后提出的VGG^[27]，相比较与之前的LeNet和AlexNet，最大的特点是网络更深，具有16-19层，不包含池化和最后的softmax层。ResNet^[21]是何凯明等人(2016)提

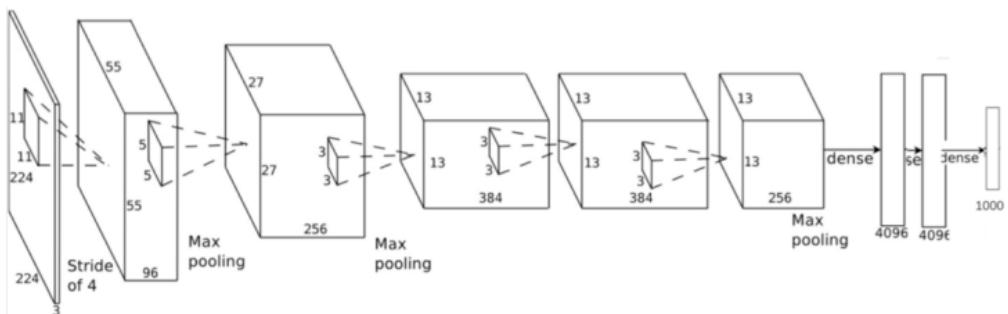


图 2-5 AlexNet网络结构示意图

出的，针对前面网络并不能随着层数的叠加而性能的提高，ResNet首次提出了

残差学习单元。如图2-6所示，假设模块的输入为 x ， $F(x)$ 指的是网络中的一系列的张量运算，假设神经网络最优的拟合结果为 $H(x) = F(x) + x$ ，那么神经网络的最优的映射函数 $F(x)$ 为 $H(x)$ 和 x 之间的残差。通过不断的叠加这个模块，可以不断堆叠加深网络的深度但是不降低网络性能。

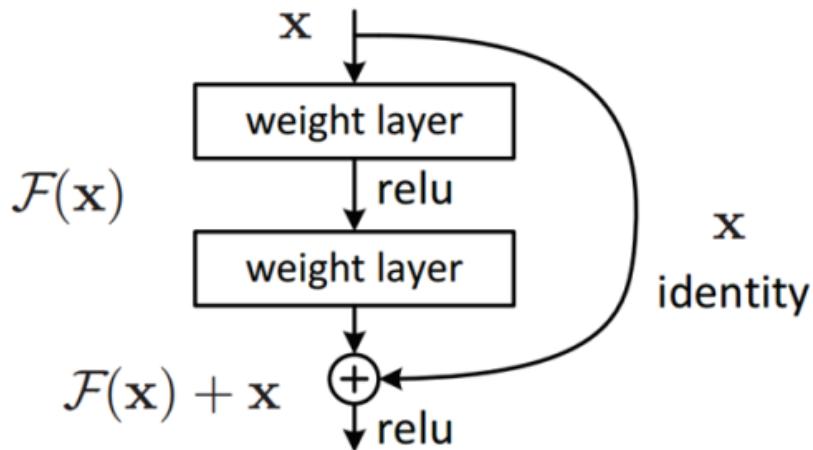


图 2-6 ResNet残差学习单元示意图

综上，卷积神经网络自提出以来，得到了极大的发展，表现为新的激活函数，降低过拟合的dropout层，残差学习模块，加上cuda硬件加速的发展，我们能训练更复杂更深的神经网络，取得更好的性能。

2.1.3 消息传递

图推理是消息传递的一种形式，条件随机场（Conditional Random Field）广泛的用在了图推理问题上。Johnson^[2]将CRF用于图像检索领域的场景图谱与对应图片的绑定的推断。本文用的方法类似于CRFasRNN^[28]和Graph-LSTM^[29]。本文的方法是收到^[8]的启发，Xu在场景图谱的生成任务中，设计了原始图和对偶图，原始图用于图片中物体节点的预测，对偶图用于关系节点的预测。和之前不同的是，我们只是对其中的关系节点构建图，通过关系的消息传递机制，我们的模型是迭代的提纯社会关系理解的预测。而不是像标准的循环神经网络（Recurrent Neural Network）只是单次预测，并不能达到提纯的效果。下面的部分我们介绍关于CRF和本文中用到了循环神经网络（RNN）以及门控循环网络（Gated Recurrent Unit,GRU）。

2.1.3.1 条件随机场

CRF是一种判别式无向图模型，用于对条件分布建模，在给定随机变量 \mathbf{X} 的条件下，构建条件概率模型 $\mathbf{P}(\mathbf{Y}-\mathbf{X})$ 。条件随机场的定义如下：设 \mathbf{X} 与 \mathbf{Y} 是随机变量， $\mathbf{P}(\mathbf{Y}-\mathbf{X})$ 是在给定 \mathbf{X} 的条件下 \mathbf{Y} 的条件概率分布。令 $G = (V, E)$ 表示一个无向图， V 是节点的集合， E 是无向边， Y_v 表示与节点 v 对应的标记向量， $w \neq v$ 表示在图中所有与节点 v 有边连接的所有节点 w ， $w \neq v$ 表示节点 v 以外的所有节点，如果所有的节点 Y_v 都满足公式2-7的马尔科夫性质，那么称 (Y, X) 为一个条件随机场。

$$P(Y_v | X, Y_w, w \neq v) = P(Y_v | X, Y_w, w \neq v) \quad (2-7)$$

CRF引入了特征函数（指数函数的形式），对于其中的特例线性链条件随机场，线性链条件随机场如图2-7。其中概率分布如下所示，其中 t_k 和 s_l 都是特征函数，

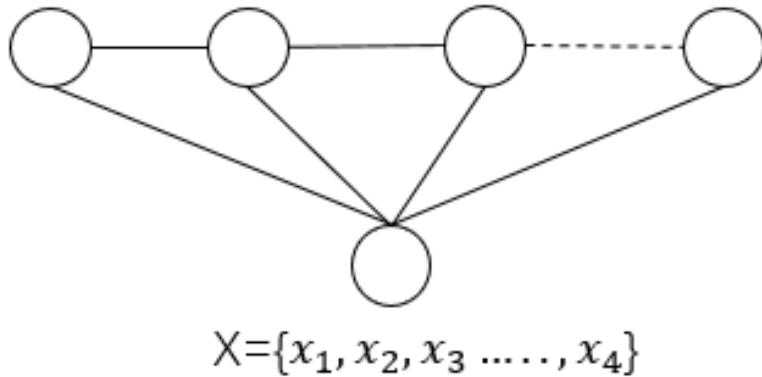


图 2-7 线性链条件随机场

λ_k 和 μ_l 对应的权值。

$$P(y | x) = \frac{1}{Z} \exp((\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i))) \quad (2-8)$$

2.1.3.2 循环神经网路

循环神经网络（RNN）是神经网络的一种，主要用于处理序列序列建模问题。其基本结构如图2-8所示，相比其他类型的网络，其就是对神经网络展开 k 个步骤，所有的输入共享一个网络模块 \mathbf{s} 。假设其中的参数为 \mathbf{W}_s 和 \mathbf{W}_x ，那么所有模块的这两个参数是共享的。对于标准的RNN结构，第 t 步，RNN的输出向量 \mathbf{s}_t 如公式2-9所示计算。由于标准的RNN网络会由于结构的问题，无法处理

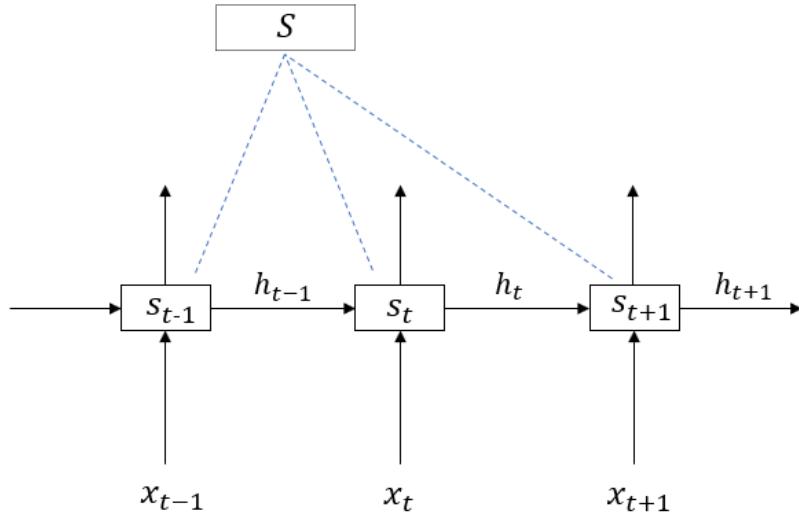


图 2-8

梯度消失和梯度爆炸的问题。后续的研究人员对于其提出的梯度消失的问题，有LSTM^[30]和GRU^[31]等改良的RNN网络。这里着重介绍本文中用到的GRU结构。基本结构如图2-9，GRU解决梯度消失主要是引入了一个更新门（update gate），该机制可用于控制在BP算法时，计算梯度的时候，通过更新门记住之前的信息，使得求导时不会陷入极小的情况，从而达到解决梯度消失问题。

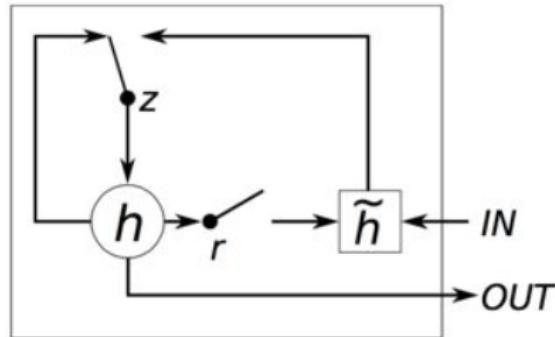


图 2-9

$$s_t = \tanh(\mathbf{W}_s * h_{t-1} + \mathbf{W}_x * x_t + b) \quad (2-9)$$

2.2 物体检测与识别

在社会关系识别的任务中，有一个重要的模块是利用物体识别模型得到人的场景信息，即采用物体识别模型识别去当前图片中包含了哪些物体，得到该物体

在图片的区域。物体识别的模型包括Ross等人提出的RCNN^[32], fast-RCNN^[33], 以及Ren等人(2016)^[34]提出的faster-RCNN。前面提到的三个模型都是基于区域的物体检测模型, RCNN^[32]首次提出在目标图像中有多个目标框, 然后判断目标框是否包含物体, 具体的检测步骤如下: (1) 其中采用选择性搜索的方法得到图片中的所需要的目标框区域, 将得到的区域调整为卷积神经网络输入的消息。 (2) 利用一个预训练好的卷积神经网络, 提取第一步得到的区域中的特征。 (3) 将第二步中得到的特征当作一个线性SVM的输入, 得到物体的类别, 另外训练一个线性回归模型得到物体的目标框。 RCNN的主要缺点是对于一张图片中的每个感兴趣区域, 需要遍历提取其中的特征, 然后依次执行物体的分类和物体框的回归, 需要耗费较多时间。由于全卷积和池化层不改变某个区域在特征图和原图的位置, 因此fast-RCNN 在RCNN 的基础上提出了ROI(region of interest) 池化层, 将图片输入到卷积神经网络中, 对于特征图上的区域, 经过ROI池化层进行调整, 然后再继续之后的全连接层和一个线性回归层进行分类和目标框的确定。综上, fast-RCNN 较大程度上提高了物体检测的性能。由于fast-RCNN 在大数据集上的表现依然不能满足实际的需求, 因为RCNN和fast-RCNN均采用选择性搜索的方法得到所需要的区域, 这个步骤是比较耗费时间。因此faster-RCNN提出RPN (region proposal network), RPN主要网络包括两部分, 一部分主要是对生成的anchors进行判断是foreground还是background, 其中foreground代表目标, 另外一部分主要是对检测框的位置进行调整。经过RPN网络后得到候选区域, 再利用ROI池化得到特征向量进行物体类别的判断和物体框的进一步精确判断。

综上, 以上的篇幅主要是回顾了在社会关系检测的工作中, 有用的物体检测方法和一些相关工作。结论是得益于GPU等硬件设备的发展, 物体识别领域的算法也得到了快速的发展, 尤其是随着特征提取模块的发展, 卷积网络越来越深, 能学习到更多更丰富的特征。对于一幅图片, 我们能在得到更多的、更准确的物体框和类别。

2.3 社会关系检测

本章将回顾社会关系检测领域的一些相关工作，并且对于消息传递机制的介绍，以及消息传递机制的相关工作的一些介绍。

2.3.1 已有工作的简单介绍

社会关系检测是社交网络的一个基础^[?]，社会关系检测作为一个重要的多学科问题，在计算机视觉领域受到越来越多的关注。随着这个问题被提出以来，有大量的工作用于从图片中抽取两个人之间的社会关系。主要有wang 等人（2010）^[3]，以及Dibeklioglu等人（2013）^[19]和Zhang 等人（2015）^[6]提出的利用面部表情、年龄、性别、姿势等多种特征的联合模型。Li等人（2017）^[4]提出的多次观察的**dual-glance**模型。以及Wang 等人（2018）^[7]提出的基于常识知识的深度推理模型**GRM**。以及从视频中抽取社会关系的工作ding等人（2010）^[35]、Ramanathan等人（2013）^[36]。Sun等人（2017）^[5]基于范围的理论，将社会关系划分为5个范围，同时接着这五个范围又划分为16种社会关系并且扩充了**PIPA**（people in photo album）数据集^[37]，得到了**PIPA-relation**。

Zhang 等^[6]提出的模型认为从心理学的角度出发，认为人的关系主要由人的面部表情的一些特点决定的。首先，模型设计了一个基准模型用于提取图片中两个人对的特征，对于两个人对，基准模型采用共享参数的深度卷积网络（DCN），利用DCN提取得到的特征分别记为 $\mathbf{x}^r, \mathbf{x}^l$ ，并且 $\forall \mathbf{x}^r, \mathbf{x}^l \in R^{2048 \times 1}$ ，经过一个权重矩阵 $\mathbf{W} \in R^{4096 \times 256}$ 得到特征向量 x_o 。对于已经标注好的人脸图片两个人脸分别为 $\mathbf{x}^l, \mathbf{x}^r$ ，其中 x_l, x_r 是左上角的坐标， w, h 分别是两个人脸包围盒的宽度和高度。2) 人脸的相对位置 $\frac{x^l - x^r}{w^l}, \frac{y^l - y^r}{h^l}$ 。3) 人脸之间的比例 $\frac{w^l}{w^r}$ 。以上的三项空间特征会和DCN得到的 \mathbf{x}_t 拼接来学习得到关系类别。除此之外， \mathbf{w}_{gi} , \mathbf{W} , \mathbf{K}^l , and \mathbf{K}^r 可以采用标准的正太分布初始化。结合之前符号的定义，该模型损

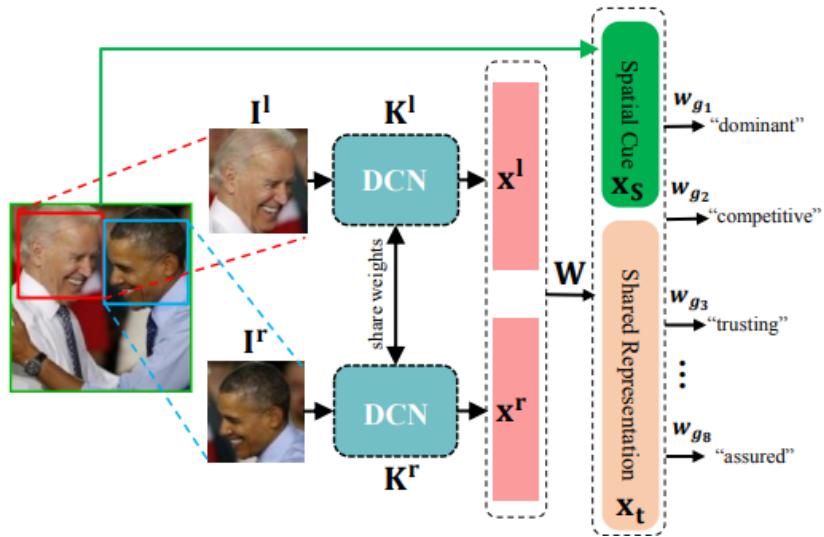


图 2-10 zhang 的模型

失函数定义如下：

$$\begin{aligned} \arg \max_{\Omega} p(\{\mathbf{w}_{g_i}\}_{i=1}^8, \mathbf{W}, \mathbf{K}^r | \mathbf{g}, \mathbf{x}_t, \mathbf{x}_s, \mathbf{I}^r, \mathbf{I}^l) \propto \\ \left(\sum_{i=1}^8 p(g_i | x_t, x_s) p(w_{g_i}) \right) \left(\sum_{j=1}^K p(k_j^l) p(k_j^r) p((W)) \right), \quad (2-10) \end{aligned}$$

s.t. $\mathbf{K}^r = \mathbf{K}^l$

基于以上的工作，该模型同样认为人的面部属性对最终的关系预测可以起到关键的作用。

Li等人 (2017)^[4] 基于之前的工作，针对社会关系理解的任务提出了包含两个关系粒度的数据集，PISC-coarse和PISC-fine (2017)^[4]。该工作首次提出利用图片中的场景来协助预测两个人之间的关系，场景具体表示为该图片中的物体。直观来说，如果一幅图片中包含电脑桌子等物体，那么大概率是“同事”关系。**dual-glance**模型分为两个模块，first glance和second glance，first glance 的输入为一张图片 \mathbf{I} 和两个人身体的包围盒。针对图片 \mathbf{I} ，首先修剪出3个小块，前两个小块分别覆盖住两个人， p_1 和 p_2 ，第三个小块覆盖两个人，表示为 p_{uo} 。这三个小块的像素被修正为 224×224 大小，作为后续三个CNNs 网络的输入，其中 p_1 和 p_2 的特征抽取网络是共享网络参数的。此外，包围盒的位置信息对于视觉信息是一种补充，例如亲密关系往往的离得比较近的，无关系的两个人对的包围盒离的较远。位置信息 \mathbf{g} ，经过预训练的ResNet^[21] 抽取得到表示人对关系的特征向量

为 \mathbf{v} , 经过拼接和全连接网络后得到这个人对的特征向量 \mathbf{v}_{top}

对于second glance模块, 模型利用faster-RCNN^[34]中的RPN产生一系列的区域候选框 P_I , 这些候选框包含物体的概率大于超参数 m 。对于一个人对, 我们从集合 P_I 中选择部分候选框 $R(b1, b2; I)$, 选择方式如2-11, 其中函数 $G(b1, b2)$ 表示两个包围盒的IOU, τ_u 是阀值。

$$R(b1, b2; I) = \{c \in P_I : \max(G(c, b1), G(c, b2)) < \tau_u\} \quad (2-11)$$

利用faster-RCNN得到的特征图, 采用ROI pooling抽取固定长度的向量物体特征向量 $v \in R^k$ 。同时用 $\{v_i | i = 1, 2, \dots, N\}$ 作为 $R(v1, b2; I)$ 中的物体向量集合。然后依次采用公式2-12方式将 \mathbf{v}_{top} 和物体的向量集合得到 h_i

$$h_i = v_i + w_{top} \otimes v_{top} \quad (2-12)$$

之后采用attention的方式将 h_i 得到最终的得分 s_i , 具体细节如公式2-13所示, 其中attention的权重 $a_i \in [0, 1]$

$$\begin{aligned} a_i &= \frac{1}{1 + \exp(-(W_{h,a}h_i + b_a))} \\ v_i^{att} &= a_i v_i \\ s_i &= W_s v_i^{att} + b_s \end{aligned} \quad (2-13)$$

GRM (2018) ^[7]同样认为引入当前人对的周边物体的信息对判断人对之间的社会关系是有帮助的, 但是现有前面的模型忽略了周边物体的语义和这些物体以及社会关系的先验知识。除此之外, 周边物体和社会关系的交互太过简化了。因此, GRM采用深度学习结合先验知识制定了一个图推理模型 (Graph Reasoning Model) 来实现社会关系检测任务。首先, GRM基于训练集中的样本构建了一个描述物体和社会关系共现的图谱。形式化来说, 先验知识图谱表示为 $G = \{V, A\}$, 其中 V 表示图上的节点集合, A 表示节点间的邻接矩阵。当前的图 G 包含两种节点类型, 一种节点表示社会关系, 一种表示物体类别, 针对当前图片场景构建的图的节点来说, 采用相应图片对应区域抽取的特征向量初始化。对于社会关系节点, 这里采用Li等人的方式^[4]的方式修剪出三部分包含人的区域, 利用预训练好的ResNet提取出特征向量, 与空间信息等特征向

量进行拼接得到一个 d 维的特征向量 $f_h, f_h \in \mathbb{R}^d$ 。 f_h 作为所有社会关系节点的初始化向量。对于物体节点来说，我们需要用到在大规模训练集上预先训练好的faster-RCNN^[34]，由于PISC-coarse和PISC-fine等数据集并没有标注好的物体类别。这里的大规模数据集指的是COCO^[38]，COCO是专门为了物体识别标注的数据集，包含我们日常生活中常见的80类的物体。利用物体检测模型提取出高于置信度 ϕ （ ϕ 是一个超参数），对于这里未检测到的物体，采用全0的特征向量。之后利用GGNN(Gated Graph Neural Network)网络^[39]来执行图上的消息传递。通过GGNN，能探索人对和图片场景中物体的交互。物体的类别是一个关键的因素用于区分不同的社会关系，但是由于有的物体的信息对于判定社会关系时不重要的、甚至起到了干扰的作用，因此GRM提出了图注意力的机制，有选择的采用能起到区分不同社会关系的物体节点，按照区分能力的大小给予不同的权重。综上所述，GRM模型提供了一个可解释的方法来提高社会关系检测的能力，从周边场景中推理得到有效的信息。

前文提到了图片上人对的特征和物体的信特征抽取，GRM接下来需要执行不同节点间的信息传递，对于其中GGNN的执行过程，对于图G中的节点 v ，其对应的隐藏状态为 h_v ，GGNN模型融合邻接节点的信息来更新 v 节点隐藏状态，GGNN采用类似于Gated Recurrent Unit (GRU)^[31]机制的方式实现节点间的信息融合。所以，对于第 t 步GRU unit的输入为 a_v^t 和 h_v^t ， a_v^t 是融合邻接物体节点的信息向量，融合方式如公式2-14，其中 A_v^t 代表前面提到的物体和社会关系的邻接矩阵，矩阵的值为它们在训练集中共现的概率。

$$a_v^t = A_v^t [h_1^{t-1} \dots h_{|V|}^{t-1}]^\top + b \quad (2-14)$$

经过 T 此GRU的迭代后，分别得到物体节点、关系节点的隐藏层表示。但是由于周边的物体在区分不同的关系起到不同的作用，GRM采用attention机制来结合物体的信息。attention的如公式2-15，其中得到邻接物体节点权重为 α_{ij} 。进过GGNN、attention两个模块后，得到表示社会关系的特征向量 \mathbf{f}_i ，用做最后的

分类，取概率最大的作为当前人对的关系。

$$\begin{aligned}\mathbf{h}_{ij} &= \tanh(\mathbf{U}^a h_{r_i}) \odot \tanh(\mathbf{V}^a h_{o_j}) \\ e_{ij} &= \text{Atten}(\mathbf{h}_{ij}) \\ \alpha_{ij} &= \sigma(e_{ij})\end{aligned}\tag{2-15}$$

综上所述，现有的工作均在例如只是提取人在图像上区域的特征进行分类的基准模型的基础上，进一步的引入了更多的信息，更多的约束。直观来看，单纯从两个人的图像区域信息直接推断出关系类别确实是很大的挑战，所以一种解决方案是如何引入上下文的信息来辅助判别，这也是本文工作的切入点。

2.4 本章小结

首先，由于在图像领域的实际任务越来越复杂，传统的特征提取方法依靠大量人工的先验知识来设计特征提取器，这并不能应对当前的需求，如何提取图像的特征向量变得越来越重，同时这也是深度网络得到发展的原因之一。本章首先给出了神经网络的发展过程，从最简单的逻辑神经元模型出发，解释了连接权重、激活函数和优化方法等概念。到图像领域常用的卷积神经网络，并且简要的阐述了卷积神经网络的特点局部连接与权重共享。同时举例说明了卷积神经网络的卷积层和池化层的细节。之后卷积神经网络的架构的发展，从第一个开始应用的卷积神经网络**LeNet**，以及之后网络层数不断加深的**AlexNet**和**VGG**等架构，不断刷新了深度神经网络的表现能力，与此同时在些网络结构中也成功的运用了避免梯度消失的激活函数**ReLU**、以及防止过拟合的**dropout**层等。这些方法极大的促进了深度神经网络的发展，即卷积核越来越小，层数越来越深。同时也介绍了和本文相关消息传递机制，我们也简要的介绍了消息传递的两种图推理方式**CRF**和**RNNs**。

然后，回顾了物体检测与识别的三个经典方法**RCNN**、**fast-RCNN**和**faster-RCNN**，按照时间轴的顺序分析每篇文章的贡献和缺点，**RCNN**首次提出了基于感兴趣区域的物体检测，但是由于对每张图片的每个区域的特征提取是独立的，所以需要耗费大量的时间。之后的**fast-RCNN**对于每张图片的所有感兴趣区域只运行一次，然后运用**ROI**池化的方法，虽然这样极大的降低了检测时间，但对于

大规模的真实数据集还是不够理想。而faster-RCNN 从提取感兴趣区域出发，提出RPN网络替代选择性搜索，极大的提高了检测的速度。

最后，本文对社会关系理解的近年工作进行了回顾，Li提出的**dual-glance**包含两个部分，其主要思想是利用物体检测模块的RPN网络生成的物体候选区域来当作上下文，用该特征来提存由人的区域提取得到的特征向量，该方法相比基准模型有较大的提升。之后Wang提出的**GRM**模型，首次在社会关系理解任务上引入先验知识，论文的先验知识指的是物体和社会关系的共现频率，通过GGNN对先验知识的建模，提升了模型的效果。这些工作均说明了，单纯从图像特征上来判别社会关系，其能达到的效果是有限的，因此引入更多的约束，这些约束包括外部信息和内部信息，并对这些信息进行建模是一个解决方案。

第3章 基于消息传递的人对间关系网络

本章将提出一个消息传递的人对间关系网络，简称PPRN。本章节将说明PPRN网络的特征提取模块、消息传递模块、消息池化模块，最后定义整个模型的优化方法和具体实现细节。

3.1 基本框架

首先，需要提到的是PPRN的输入和现有的模型是存在一些差别。对于**dual-glance**和**GRM**来说，输入是一张图片，两个人的包围盒坐标。在第一章的介绍中提到，本文的出发点是希望对人对间的关系进行建模。对于图像的社会关系理解任务，需要识别出图片上每个人对之间的社会关系，但是图像特征的模糊性和不确定性，这是一个很难的任务。利用上图上人对之间的关系信息，即结构信息，图像中人对间关系的结构信息能抑制对不准确检测结果。因此，PPRN网络的输入是以图片为单位而不是人对，每次同时输出一张图片上所有对间的关系。PPRN主要是由3个模块组成，并且模型的主要预算模块是向量，以下提到的所有向量编码维度均为n维。如果将人之间的社会关系视作节点，本文对于图片上所有这样关系节点的全连接图成为**社会关系图谱**，节点间的连边不含语义，采用类似于**GRM**的方式，利用特征提取模块得到的编码来初始化社会关系图谱上的关系节点。人对关系采用**GRU**来探索一个节点和图上其它节点间的交互，一张图片往往是一个场景，使用场景中其它节点的社会关系来对当前节点的社会关系消歧。此外，引入注意力机制对探索不同节点交互时的影响力。

如图3-1所示，PPRN模型是一个端到端的架构。PPRN模型接受一张图片和图片上所有人的包围盒的作为输入，经过前文提到的3个模块，并且每个模块的作用各有分工。模型首先提取图片中关于人和人对包围盒的特征，图3-1中的 p_{uij} 表示图上第i-th个人和第j-th个人的联合区域。 b_i 表示第i-th个人的包围和的坐标和包围盒的面积编码。在消息池化模块中，消息池化函数计算关系间的消息，然后作为下一轮GRU神经元的输入。其中 \oplus 表示一个可学习的权重参数。

而对于消息传递模块，通过循环神经网络来实现通过反复传递消息来提醒当前节点目前场景中其它节点的行为，这个是图推理的一种实现方式，可以优化对社会关系的估计。在最后一次迭代时，利用GRU神经元最后的隐藏层的输出连接一个全连接层得到最后的预测结果。

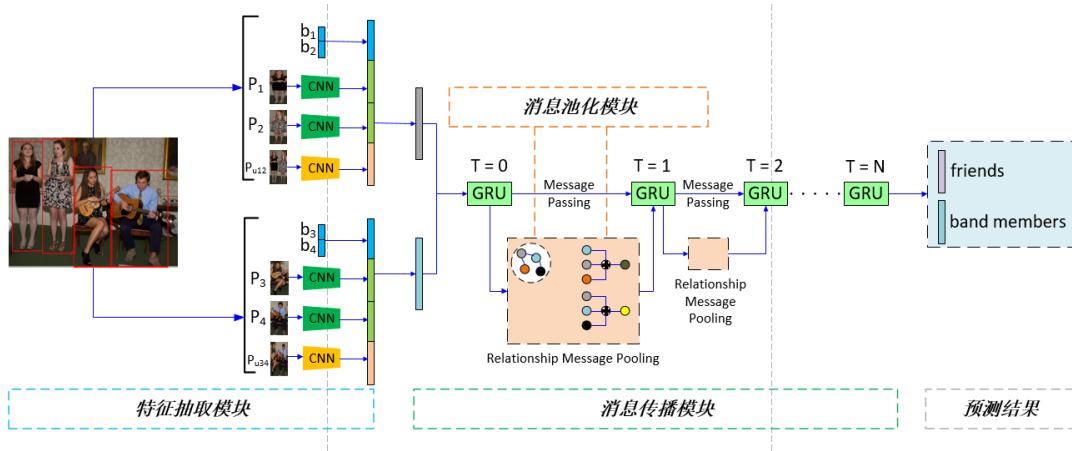


图 3-1 PPRN模型的结构示意图

3.1.1 特征抽取模块

模型微调是指给定一个预训练模型，这里的预训练网络通常是在大规模标注数据集上训练好的模型，例如常用的预训练模型有Vgg和ResNet。此时的模型已经能很好的抽取大部分图片的信息，比如卷积神经网络的浅层往往是提取基础特征，例如边缘、轮廓等基础特征。深层卷积层提取抽象特征，例如脸型，而最后的全连接层根据特征组合进行连接分类。对于在大规模数据集上已经训练过的模型来说，已经具备了提取浅层特征和深层次抽象特征的能力。基于预训练的模型来新的数据集能节省很多的计算时间和计算资源，而且还能提升模型的效果。常见的模型微调方法是将我们现有的分类层替代预训练模型的最后一层，然后以较小的学习率微调前面的所有层，原因是不想过快的扭曲前面已经学习好的特征，然后着重训练最后的分类层。或者冻住预训练模型前面的若干层，微调后面的网络层。

特征提取模块包括物体的特征和人对的特征，对于人对的特征，采用的预训练模型是ResNet-101^[21]，本文的做法微调ResNet-101模型的后2层和最后分类层。特征抽取模块的主要目标是提取出输入的图片和人的包围盒坐标区域的图像

特征。因此，该模块首先修剪出三个小块， p_1 和 p_2 为个体的区域， p_u 为两个人区域的并集区域，这些区域的特征包含了用于识别关系的基本组成部分。这些区域首先被调整为大小 224×224 ，然后输入三个ResNet，其中 p_1 和 p_2 的区域的网络是共享参数的。三个模型最后一个卷积层的输出拼接在一起，形成视觉特征向量 \mathbf{v}_1 。其次，位置信息同样很重要，对于其中的关系类别中的“无关系”，这单纯从视觉特征来判断是很难的，此时位置信息往往能起到很好的区分作用。我们表示一个包围盒 i 的特征为 $b_i^{pos} = \{x_i^{min}, y_i^{min}, x_i^{max}, y_i^{max}, area_i\} \in R^5$ 。其中这些值都是相对位置。 x_i^{min} 和 y_i^{min} 是包围盒的左上角坐标， x_i^{max} 和 y_i^{max} 是右下角坐标。最后这些特征拼接在一起后通过全连接层形成关系特征向量 \mathbf{v}_h

3.1.2 消息传递模块

对于一张图片中 n 个人对的关系，通过特征抽取模块，得到了图片中所有人的关系特征编码， $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ ，并且对于一张图片构成的社会关系图谱，是一张全连接图。对于特征抽取模块得到的关系向量编码通过如公式3-1所示，转换到一个低维空间，其中 φ 可视为全连接层。

$$h_i^0 = \varphi(\mathbf{v}_i) \quad (3-1)$$

为了学习到图片中的场景信息信息对关系编码的约束，我们采用了循环神经网络来执行社会关系检测理解的推理工作。与Zheng^[28]的模型不同的是，我们采用通用的RNN 神经元来计算隐藏层状态。相比较LSTM，GRU少了一个门，因此更加简单并且参数较少，学习起来更快，因此本文采用的是GRU模块。我们采用第 t 步的隐藏层状态编码表示社会关系图谱中所有关系节点信息，关系节点的向量编码会随着RNN 序列的长度每一步都更新。而关键的是每一步GRU的输入来自消息池化模块的输出，消息池化模块承担着关系节点间交互的任务。具体细节如公式3-2所示，其中 \mathbf{v}_h 会当作GRU 第一步的输入，并且此时的隐藏层的初始化为值为0的向量。这里的 σ 和 $tanh$ 表示逻辑斯谛回归和双曲正切函数。 \odot 操作表示点乘操作， \mathbf{r}_t 表示重置门， \mathbf{z}_t 表示更新门，下标 t 表示迭代步骤。 \mathbf{W}_r 和 \mathbf{W}_z 表示

两个门需要的参数，这些参数可训练的。

$$\begin{aligned}
 \mathbf{r}_t &= \sigma(\mathbf{W}_r[\mathbf{h}_{t-1}, \mathbf{x}_t]), \\
 \mathbf{z}_t &= \sigma(\mathbf{W}_z[\mathbf{h}_{t-1}, \mathbf{x}_t]), \\
 \hat{\mathbf{h}}_t &= \tanh(\mathbf{W}[\mathbf{r}_t \odot \mathbf{h}_{t-1}, \mathbf{x}_t]) \\
 \mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \hat{\mathbf{h}}_t
 \end{aligned} \tag{3-2}$$

3.1.3 消息池化模块

消息传递模块利用RNNs解决推理问题，但是在每个迭代步的时候，GRU单元会接受多个来自社会关系图上其他节点的消息，需要有一个聚合模块来将这些信息结合成一个有意义的编码向量。直观来看，常见的池化能实现这个功能，例如常用的最大池化和平均池化。在理解当前图片的社会关系图谱时，只利用上下文的结构信息中最相关的那部分是最合理的方式。而本文的消息池化模块也是出于这个目的提出的。下面将会说明简要说明注意力机制的基本情况。如公式3-3所示，其中第 t 步的节点 i 的前一步隐藏层状态为 $\mathbf{h}_{i,t-1}$ ， $\mathbf{m}_{i,t}$ 表示来自其他节点消息的聚合，而 $\mathbf{m}_{i,t}$ 将会作为第 t 步中公式3-2中输入，即 \mathbf{x}_t 。其中符号 $[.]$ 表示两个向量编码的拼接， σ 表示激活函数， \mathbf{w} 是需要学习的参数向量， $\mathbf{h}_{j \rightarrow i,t-1}$ 是节点 j 在第 $t-1$ 步时的隐藏层编码，并且等同于公式3-2中节点 j 的 \mathbf{h}_{t-1}

$$\mathbf{m}_{i,t} = \sum_{j \neq i} \sigma(\mathbf{w}^T [\mathbf{h}_{i,t-1}, \mathbf{h}_{j \rightarrow i,t-1}]) \mathbf{h}_{j \rightarrow i,t-1} \tag{3-3}$$

3.1.4 结合周边物体信息模块

本文在模型3-1的基础上，本文实现了类似**dual-glance**的方法结合周边物体的信息，如图3-2。当前模块主要包括两个步骤，利用faster-rcnn中的RPN(region proposal network)生成物体置信度高的区域，其次利用注意力机制得到关于周边物体区域的得分。

在一张图片中，往往会存在一些日常见到的物体，例如桌子、电脑、杯子，

如果在一张图片中检测到了被杯子、桌子或者床等物体，那么当前的社会关系往往是“家庭”。因此本文在利用关系上下文的基础上，进一步加入周边物体信息来提升效果。但是在PISC 和PIPA-relation数据集中，不包含物体包围盒的标注以及物体类别的标注，这里采用的方法是利用在大规模训练集上，例如COCO^[38]预训练好faster-rcnn模型，利用其中的区域生成网络。区域生成网络的输入可以是任何大小的图片，输出是一些长方形的检测框，其中每个检测框都带有是否包含物体的得分。由于期望RPN和fast-rcnn 共享有vgg^[27]（本文采用vgg，也可采用其他的预训练模型）生成的特征图，这里的vgg 网络有13个可训练的卷积层。

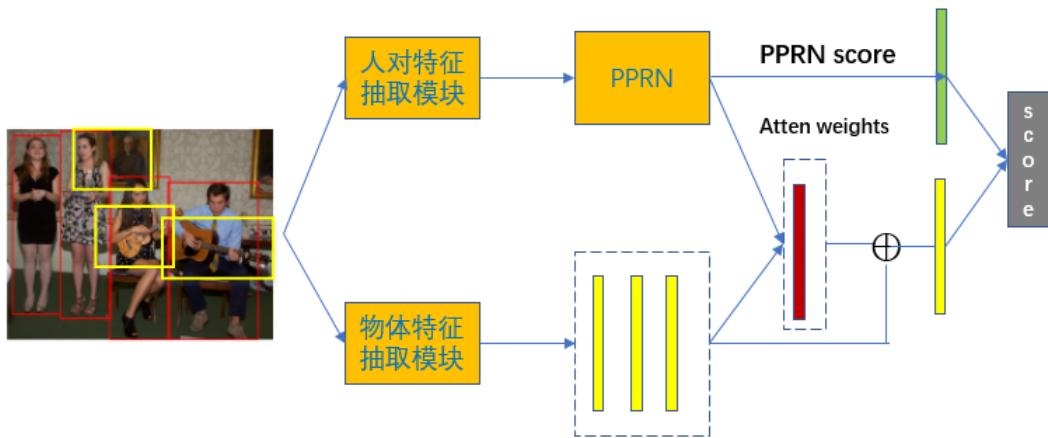


图 3-2 PPRN模型的结构示意图

经典的方法生成检测框都非常耗费时间，例如RCNN和fasr-rcnn采用的选择性搜索，而faster-rcnn 抛弃传统的方法，直接从特征图生成检测框，能极大的提升检测的速度。如图3-3，可以看到RPN网络由两个模块组成，上面的模块为物体的前后景分类，检测的目标为前景。下面的模块为检测框区域坐标的回归，用于获得更精准的检测框， proposal层结合检测框的坐标和前景类别获取在图上的区域，这两个模块结合起来相当于完成了目标定位的功能。具体执行步骤如下：

- (1) 对于， vgg 特征提取模块得到 d 张特征图，相当于每个点都是 d -dimensions。
- (2) 对于 d 张特征图，首先经过 3×3 的卷积核的处理，用于结合周边的空间信息，同时每个点对应的 d -dimension不变。
- (3) 假设在 d 张特征图中每个点上有 k 个锚点，每个锚点分为前景和后景，所以

在前景分类中， $d\text{-dimension}$ 的特征向量转化为一个 $2k$ 的得分向量，而每个锚点都有四个坐标的偏移量，对于检测框回归模块，特征向量转化为 $4k$ 的得分向量。其实RPN就是在和原图相同的尺寸上，设置密密麻麻的候选锚点。然后用cnn网络来判断哪些是包含物体目标的前景，那些是不包含物体目标的后景。

- (4) 全部锚点框都拿去训练太多了，因此会在锚点框中随机选取128个positive的锚点框和128个negative锚点框用于训练

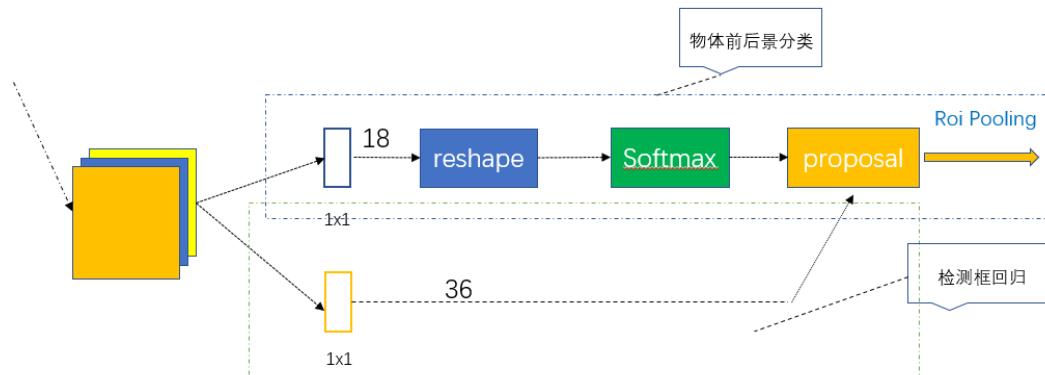


图 3-3 RPN网络结构示意图

上文提到的锚点框，其示意图如图3-4其实就是对于特征图的每个点生成一组检测框，检测框的尺寸由输入的图片的尺寸决定的，在本文中，对于每个检测点生成3种尺寸，每种尺寸3种形状的锚点框。这也是常见的多尺度方法，这9个初始的锚点框的准确度有RPN的区域回归模块处理。对于生成的检测框，因为尺度是原图的尺度大小，需要映射为特征图的尺度大小，Roi pooling层划分出非均匀的网格，对网格采用最大池化然后得到锚点框的特征向量。

利用前文提到的RPN网络，对于图片 I ，生成是前景的检测框，并且过滤掉IOU (Intersection-over-Union) 低于超参数 τ_u 的检测框，得到一组检测框 P_I ，采用预训练的vgg来得到特征图 $conv(I)$ ，对这里的每个检测框，经过一层Roi pooling 层得到固定长度的物体区域特征 v ，对于图片所有的检测框，从 $conv(I)$ 生成 $\{v_i|i = 1, 2, \dots, N\}$ 特征向量，对于关系来说，和它关联的物体是比重是不同的，例如对于关系“家庭”，“碗”的权重比“桌子”的权重大的多，所以我们期望对于不同的物体区域引入不同的权重系数。因此，在当前模块中，同样引入注意力机制 (attention mechanism)，经过消息传递和池化的关系的特征编码 v_{top} 和物体特征

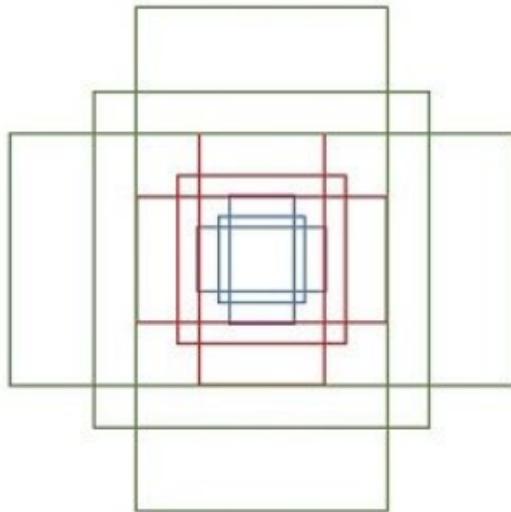


图 3-4 anchor示意图

编码 \mathbf{v}_i 共同决定权重系数的值：

$$\mu_i = \mathbf{u}^T \cdot (\mathbf{W}_r v_{top} + \mathbf{W}_o v_i) \quad (3-4)$$

$$\mathbf{a}_i = softmax(\mu_i) = \frac{exp(\mu_i)}{\sum_j^N exp(\mu_j)} \quad (3-5)$$

根据关系得到每个物体区域特征的权重后，各个区域权重分别与物体区域的特征编码相乘，如公式3-6所示：

$$\mathbf{v}_{att} = \sum_i \mu_i \mathbf{v}_i \quad (3-6)$$

最终用于分类的的编码由是将消息传递、池化模块得到的关系编码向量 v_{top} ，和带权物体特征编码特征向量 \mathbf{v}_{att} ，最终的分类层函数如公式3-7：

$$\mathbf{s} = \mathbf{W}_c [\mathbf{v}_{top}, \mathbf{v}_{att}] + b \quad (3-7)$$

3.1.5 优化和实现细节

对于图片I的第k个人对，模型预测出的得分为 $\mathbf{s}_{I,k} \in \mathcal{R}^{|\mathcal{C}|}$ ，我们采用softmax函数对得分进行归一化来得到每个类别的概率 $\mathbf{p}^{I,k} \in \mathcal{R}^{|\mathcal{C}|}$ ，如公式3-

8所示：

$$p_i^{I,k} = \frac{\exp s_i^{I,k}}{\sum_{j=1}^{|C|} \exp s_j^{I,k}}, i = 1, 2, \dots, |C| \quad (3-8)$$

这里的 C 表示社会关系的类别， $|C|$ 表示类别的数量，由于是分类任务，模型最终的损失函数如公式3-9，其中 $N(I)$ 表示图片 I 的人对数量， $L(\cdot)$ 表示交叉熵损失函数， I 表示整个图片集合。

$$\mathcal{L} = -\frac{1}{\sum_{I \in I} N(I)} \sum_{I \in I} \sum_{k=1}^{N(I)} \sum_{i=1}^{|C|} L(y_i^{I,k}, p_i^{I,k}) \quad (3-9)$$

关于本文中提到的预训练模型ResNet-101和vgg-16，使用的均是由深度学习框架pytorch提供的。并且在数据集PISC-fine和PIPA-relation中存在类别不均衡问题，PISC-fine类别分布如表3-1，在实际训练过程中，我们采用常见的过采样和降采样的方法来构造最终的训练集，例如对于PISC-fine中类别为“Commercial”的关系，我们增加3倍的样本，并且这些样本的人对会互换位置（例如p1和p2是Commercial，那么新增加的样本为p2和p1也是Commercial）。

在本文中，最开始微调ResNet-101提取人对信息时候采用的是优化方法是SGD(Stochastic Gradient Descent)，不同数据集在微调所需要的迭代次数是不一样的，同时由于本文模型的输入是以图片为单位，但是不同图片的人对数量是不一样的，所以batch_size也是不一样的。首先，微调不同数据集所需要迭代次数如表3-2，上文同样提到了，综合考虑到效率和模型效果，本文采用的方法是微调网络最后两个卷积层，其他层的参数冻结，不参与学习更新。其次，利用前面微调的模型参数，固定住，然后只更新消息传递和池化模块的参数，这个步骤采用的优化算法是Adma^[24]，不断迭代直到模型收敛。PPRN模型用pytorch实现，训练时每个mini-batch会包含24张图片，每个epoch需要的时间是200ms。

表 3-1 PISC-fine类别分布表

关系类别	Friends	Family	Couple	Professional	Commercial	No Relation
样本数量	12686	788	1552	20842	523	11979

1 2 3

表 3-2 fine-tune ResNet-101需要的迭代轮数

数据集	PISC-coarse	PISC-fine	PIPA-relation
迭代轮数	17	24	2

3.2 本章小结

本章提出了一个全新的基于消息传递机制的关系网络，简称**PPRN**。**PPRN**模型是一个端到端的架构，主要由三个部分构成，特征抽取模块、消息传递和池化模块和结合物体信息模块，两个模块是串联的，并且它们的作用和分工不同。

特征抽取模块的作用是输入一张图片，图片中所有人身体的检测框坐标，提取出其中表示人对关系的基本特征编码，包括单个人区域、两个人的联合区域和位置信息，特点是利用两个不同的预训练模型来处理单人区域和个人的联合区域，我们同时结合了图像特征和位置特征，特征提取采用的是微调后的ResNet-101。基于特征提取模块得到的向量编码已经能够表达人对间的社会关系了，而消息传递和池化模块的作用是结合当前图片的场景信息进行关系类别推理，学习到图片中不同人对关系之间约束。学习这个约束的原因是，在一张图片中，场景往往是固定的，所以图片中不同人对的社会关系之间是相互影响的。本文采用以GRU为基础的RNN 作为推理模型，每次迭代时，利用注意力机制结合图片中其它人对关系的编码，实现不同人对关系间的消息传递，RNN最后一步的隐藏层编码表示所有的人对关系。

然后，本章在**dual-glance**模型的基础上，实现了类似的结合物体信息模块。物体信息指的是利用基于大规模训练集的预训练faster-rcnn模型，RPN网络生成图片的物体检测框，包含物体的概率大于一定的阀值。本章详细的介绍了RPN网络的细节，如何生这些检测框坐标。最后提取出这部分上下文物体区域的信息。因为不同物体区域对和不同的社会关系的相关度不一样，同样利用注意力机制为

¹<https://download.pytorch.org/models/vgg16-397923af.pth>

²<https://download.pytorch.org/models/resnet101-5d3b4d8f.pth>

³<https://pytorch.org/>

每个物体区域的特征编码分配权重。

最后，本章介绍了优化和实现细节。优化主要是采用常见的交叉熵损失函数，并且针对PISC-fine和PIPA-relation数据集的数据不均衡问题，进行数据上的过采样和降采样。模型的训练过程是分段式的，包括微调预训练模型和本文提出的模块训练两部分，结合不同的学习率和优化方法模型得意训练。

第4章 实验设计与分析

本章将利用常见的社会关系理解的数据集对上一章提到的PPRN模型进行检测任务的验证，具体来说即对图片中两个标定的坐标人的关系类别。本章首先介绍当前所用到的两个数据集训练/验证/测试的数据分布情况，数据集的特点。再介绍若干对比模型，介绍实验的参数设置，然后分别对实验结果进行说明和分析，并且对其中消息池化进行了不同实现方法的对比。最后通过案例研究的方法来分析PPRN模型在关系补全任务中发挥的效果。

4.1 数据集

4.1.1 数据集简介

现有的大规模社会关系理解的数据集主要有两个：分别是PIPA-relation^[5]数据集和PISC^[4]数据集，下面简单介绍这两个数据集。

PISC数据集全称是People in social Context，它是Sun等人在2017年通过人工标注平台得到的数据集，这些图片主要来自Visual Genome^[40]、COCO^[38]、YFCC100M^[2]、instagram 和twitter等社交网站，Google和Bing商业搜索引擎。从这数据的来源可以保证数据集的图片是足够高的方差的，人的面部表情，以及场景类型。PISC数据集包含22670张图片以及对应的社会关系标注，在PISC数据集上，又包含两个粒度的识别任务，coarse-level和fine-level。如图4-1所示的划分方式，先是粗粒度的，再到细粒度的关系类别。

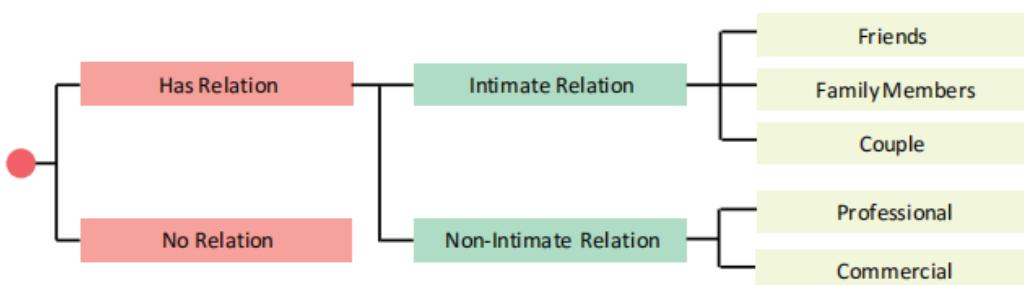


图 4-1 PISC的关系划分

PIPA-relation数据集的全称是People in Photo Album Relation，总共包括37107张图片。同样是人工标注的数据集，基于社会关系理论^[41]划分的，Sun^[5]详细的给出了每个关系域的特征。然后所有的社会关系划分为5个关系域，在构建数据集的过程中，这5个关系域划分为16种社会关系，五个关系域分别是Attachment domain、Reciprocity domain、Mating domain、Hierarchical power domain和Coalitional groups domain。Attach domain划分为*father-child*, *mother-child*, *Grandpa-grandchild*和*grandma-grandchild*, Reciprocity domain划分为friends, siblings和classmates。Mating domain只包含单条关系*lovers/spouses*。Hierarchical power domain划分为presenter-audience, teacher-student, trainer-trainee and leader-subordinate。Coalitional groups domain划分为band members, dance team members, sport team members and colleagues。

数据集的情况如表4-1，“Train”表示训练集图片的数量，“Valid”和“Test”分别表示验证和测试集的图片数量。“#train”表示训练集人对的数量，“#valid”和“#test”分别表示验证和测试集的人对数量。

表 4-1 PISC、PIPA-relation数据集的统计表1

数据集	Train	Valid	Test	#train	#valid	#test
PISC-coarse	13142	4000	4000	14536	25636	15497
PISC-fine	16828	500	1250	55400	1505	3691
PIPA-relation	5857	261	2452	13729	709	5106

4.1.2 数据集分析

对于数据集PISC-coarse、PISC-fine和PIPA-relation，本文做了基本的数据集分析，如表4-2，其中“Sui”表示一张图片有多个对的比例，“Unsui”则表示一张图片只有一个人对的百分比。“Single Rel”一张图片包含的关系类别只有一种，“Multi Rel”表示一张图片包含多种关系。从表统计数据可以得到两个结论，从“Sui”和“Unsui”来看，绝大多数的图片是包含多个人对的，从“Multi Rel”来看，一张图片的关系种类往往是相同的，因为直观来看，给定场景下的社会关系时稳定的。例如，假如张图片是一个会议的场景，那么其中往往会有多个对，并且这些人对间的关系往往是“colleagues”或“presenter audience”。

表 4-2 PISC、PIPA-relation 分析表，单位为(%)

数据集	Sui.	Unsui.	Single Rel	Multi Rel
PISC-coarse	87.1	12.9	79.9	20.1
PISC-fine	83.9	16.1	86.4	13.6
PIPA-relation	71.9	28.1	94.9	5.1

此外，对于关系类别这项统计，本文做了进一步工作，并且发现一张图片的关系种类大多只是一种，少部分是两种。如图4-2所示，在PISC-coarse中，大约79.9%的图片是只有一种关系类别，20.0%的图片有含有两个关系。

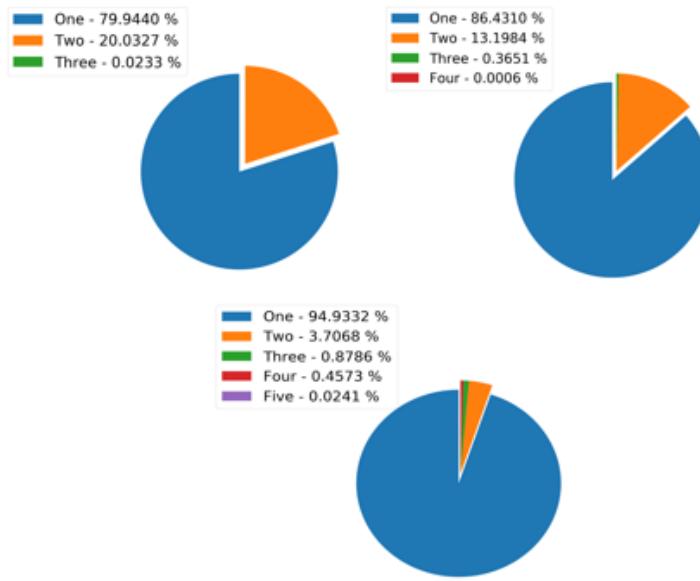


图 4-2 数据集的关系类别统计

4.2 实验设置

4.2.1 实验方法

类似于GRM模型，我们采用的是每个类别的召回率和mAP（mean average precision）。mAP常作为物体检测的衡量标准，它是不同召回值的下最大精度的平均值。举例如下，假设当前图片有五个苹果，首先对物体检测模型所有检测框分类的类别为苹果按得分从大到小排序。依次计算准确率和召回率，这里的准确率指的是TP和FP。假如总共有10个检测框被识别为苹果，其计算结果如表格??所示，例如其中的对于排名第三的样本，**Precision = 2/3 = 0.67**，**Recall = 2/5 = 0.4**。根据数据，以Recall为横坐标，Precision为纵坐标，可以画出如

表 4-3 mAP计算示例数据

得分排名	correct?	Precision	Recall	得分排名	correct?	Precision	Recall
1	True	1.0	0.2	6	True	0.5	0.6
2	True	1.0	0.4	7	True	0.57	0.8
3	False	0.67	0.4	8	False	0.5	0.8
4	False	0.5	0.4	9	False	0.44	0.8
5	False	0.4	0.4	10	True	0.5	1.0

图4-3所示的曲线c。将recall值划分为11个值，我们将 $Recall \geq \bar{r}$ 的Precision替换为最大的，由此可得到曲线c。计算方法如公式4-1所示。我们给定的数据，对于苹果这一类别的 $AP = (5 \times 1.0 + 4 \times 0.57 + 2 \times 0.5)/11$

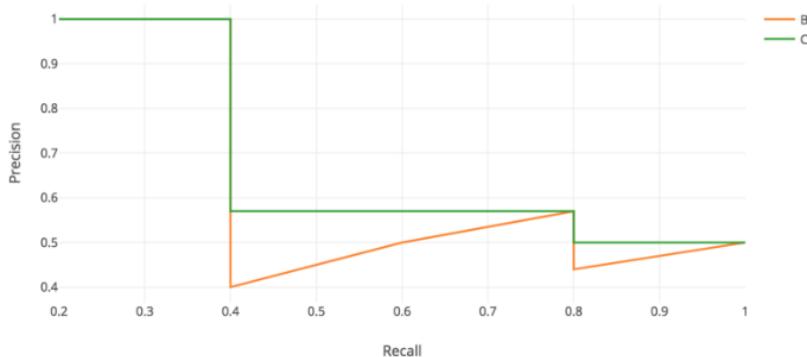


图 4-3 mAP计算示例图

$$AP = \frac{1}{11} \sum_{r \in \{0.0, 0.1, \dots, 1.0\}} P_{interp}(r) \quad (4-1)$$

$$P_{interp}(r) = \max_{\bar{r} \geq r} p(\bar{r})$$

在预训练模型，包括ResNet-101和Vgg-16最后一层的维度都是是4096，经过全连接层，在消息传递和池化模块的中的编码长度为512。结合周边物体信息模块中注意力机制的attention_size=30，在微调ResNet-101和Vgg-16，卷积层学习率设置为 1×10^{-4} ，新增的分类层学习率为 1×10^{-3} 。训练时，学习率设置为 1×10^{-4} ，权重衰减为 5×10^{-4} 。消息传递和池化模块的迭代数 $T = 4$ 。并且微调时的优化方法是SGD，训练迭代模块的优化方法是Adam。

4.3 PISC数据集实验结果

在实验中，本文主要与以下几个模型进行对照，在我们的整体框架中，主

要的有特征提取模块 (m0)、消息传递和池化模块 (m1)，周边物体信息模块 (m2)，对于m0和m1组成的网络称为PPRN-(obj region)，而m0、m1和m2三个模块一起的模型称为PPRN-(obj region)：

- **Union CNN**^[1]学习Lu等人^[1]利用一个CNN网络来预测关系方法，同样利用一个CNN网络来提取人对的联合区域的特征来进行分类任务。
- **Pair CNN**^[4]由两个共享参数的CNNs提取两个修剪出来的图像特征进行分类。
- **Pair CNN + BBox + Union**^[4]在前面两个特征提取模块的基础上，进一步结合两个个体包围盒的位置信息。
- **Dual-glance**^[4]实现两个粒度的分类任务，分别是PISC-coarse 和PISC-fine，分别包含三种和6中类别的关系。Dual-glance利用了PairCNN+BBox+Union，以及物体区域的特征来提存预测结果。
- **GRM**^[7]提出了一个图推理模型来提升社会关系理解任务，该模型集成物体和社会关系共现概率的先验知识，GRM采用的是人对特征和上下文物体特征之间的消息传递。

在表4-4中，我们展示了在PISC数据集上，按照每个类别的召回率和“mAP”的设定得到的实验结果。

表 4-4 在PISC-coarse上的实验结果，单位为百分比(%)

评价标准	Intimate	Non-Intimate	No Relation	mAP
Union CNN ^[?]	72.1	81.8	19.2	58.4
Pair CNN ^[?]	70.3	80.5	38.8	65.1
Pair CNN + BBox + Union ^[?]	71.1	81.2	57.9	72.2
Pair CNN + BBox + Global ^[?]	70.5	80.0	53.7	70.5
Dual-glance ^[?]	73.1	84.2	59.6	79.7
GRM ^[?]	81.7	73.4	65.5	82.8
PPRN-(obj region)	81.9	67.3	74.7	81.8

与此同时，在表4-5中，我们也展示了在PISC-fine数据集上，同样按照mAP的设定之后得到的实验结果。根据4-4和4-5所示的实验结果，我们可以得出以下结论：

表 4-5 在PISC-fine上的实验结果，单位为百分比(%)

评价标准	Friends	Family	Couple	Professional	Commercial	No Relation	mAP
Union CNN ^[?]	29.9	58.5	70.7	55.4	43.0	19.6	43.5
Pair CNN ^[?]	30.2	59.1	69.4	57.5	41.9	34.2	48.2
Pair CNN + BBox + Union ^[?]	32.5	62.1	73.9	61.4	46.0	52.1	56.9
Pair CNN + BBox + Global ^[?]	32.2	61.7	72.6	60.8	44.3	51.0	54.6
Dual-glance ^[?]	35.4	68.1	76.3	70.3	57.6	60.9	63.2
GRM ^[?]	59.6	64.4	58.6	76.6	39.5	67.7	68.7
PPRN-(obj region)	61.0	67.1	56.2	76.9	46.0	68.1	69.7

- (1) 首先，Pair CNN + BBox + Global, Dual-glance和GRM都引入了额外的Faster-RCNN^[34]来抽取当前图片的上下文信息(物体区域)。GRM进一步利用识别出的物体类别构建了物体类别和社会关系类别的语义共现的知识图谱，通过神经网络融入关系类别和上下文线索的先验常识知识，进而解决社会关系理解问题。需要注意到的是这些模型均引入了额外的检测标注，而这些步骤是会带来额外的噪声和性能消耗的。
- (2) 其次，结合表4-4的数据，对于coarse-level的识别，PPRN-(obj region)取得了75.1% 的准确率，81.8%的mAP。对于fine-level的识别，PPRN-(obj region)取得了65.6%的准确率，69.7% 的mAP。本文提出的模型在fine-level的识别任务上超过了所有的基准模型，但是在coarse-level的识别任务上略微低于GRM，但是仍然超过了其它引入了周边物体区域的模型。

4.4 PIPA-relation数据集实验结果

在PIPA-relation数据集上，本文将和现有的方法进行对比，例如:Two Stream CNN^[5]，Dual-glance^[4]和GRM^[7]，这些模型在之前均取得了最好额度效果。同时，我们直接从文献中获取这模型的数据，如表4-6所示，由于有的关系的样本数量较少，所有的基准模型均只采用准确率衡量模型的效果，因此本文同样如此。值得提到的是，PPRN-(obj region)明显超过现有的基准模型，比GRM超

过2.4%，比Dual-glance超过5.1%。但是PPRN+(obj region)的效果次于PPRN-(obj region)，轻微好于最好的基准模型GRM 0.4%。

表 4-6 Accuracy (in %) evaluating our PPRN model and previous methods on PIPA-Relation.

Methods	accuracy
Two stream CNN ^[?]	57.2
Dual-Glance ^[?]	59.6
GRM ^[?]	62.3
PPRN-(obj region)	64.7

4.5 实验结果分析

在本文中，最核心的部分即消息传递机制的引入，其中起到核心作用的包括采用可学习的参数来聚合社会关系图谱中其它节点的隐藏层编码。同时为了进一步证明当前方法的作用，我们分别采用了其它标准的池化方法，包括average pooling、max pooling。实验结果如表4-7所示，并且在表中给出了PPRN+(obj region)在PISC数据集上的实验结果。

表 4-7 RCNN的mAP和准确率的实验结果，消息池化模块采用不同的方式的结果，所有实验数据的单位为百分比(%)

评价标准	PISC coarse		PISC fine	
	accuracy	mAP	accuracy	mAP
RCNN	-	63.5	-	48.4
PPRN(max pooling)	74.3	80.8	64.1	68.3
PPRN(avg pooling)	74.6	80.1	63.8	68.3
PPRN(attention)	75.1	81.8	65.7	69.7
PPRN+(obj region)	74.9	81.2	65.3	69.1

根据表4-6和表4-7所展示的实验结果，我们可以得到以下的结论：

- (1) 首先，对于前面提到的PPRN-(obj region)模型的主要工作是处理关系上下文信息，没有考虑周边的物体信息，考虑这部分的信息是需要引入额外的物体检测模型。PPRN+(obj region)在消息传递机制的基础上进一步加入周边物体的信息，但是在多个实验中并没有超过未引入这部分信息的基准模型。从周边的物体信息的角度来看，周边的物体信息可以认为是形成了一个场景，物体上下文特征起作用的前提是物体检测和识别模型性能效果，

同时这部分的特征形成的场景信息往往是单一的，对最后的关系编码的约束也是单一的，同时考虑关系的上下文类似于考虑上下文物体的信息。

- (2) 其次，在GRM等引入关系类别和物体类别共现的常识知识，很大程度依赖物体的类别是否正确，才能确定加入的常识知识是否正确。第一，例如在当前图片中，检测到了*laptop*，容易倾向于推理出当前场景下所有人对关系为*professional*，但是faster-rcnn并不是总能识别正确，所以会带来很大的噪声。第二，检测出的*laptop*会帮助*professional*关系的正确分类，但是对*friends*来说并不存在这样的辅助作用，并且这两类关系在现实数据集和场景中大量存在。
- (3) 如表4-7所示，RCNN是只用到了图片中物体信息，是最差的效果。Dual-glance训练了一个基于注意力机制的人对模块根据不同的人对关系来结合不同比重的物体特征。对于模型PPRN+(obj region)来说，模型效果并没有提升，依据前面的分析，关系上下文覆盖了周边物体区域上下文的部分，并且实验证明了我们的假设。

4.6 案例研究

除此之外，为了观察PPRN-(obj region)模型的在社会关系理解任务上的具体表现，我们在PISC-fine的测试集中随机抽取了部分图片，并且列出了他们的具体表现。在图4-4中，我们分别展示了4个样例的结果，左边代表原图片，右边是社会关系图谱，其中(A)表示标注，(B)表示我们模型分类的结果，(C)表示基准模型中GRM的结果。其中图片中人的包围盒的颜色对应社会关系图谱中节点的颜色，节点间不同颜色的边表示不同的社会关系，在图4-4中列出了颜色和关系的对应表。从这四个例子来看，相比于GRM生成的社会关系图谱(C)，(B)的准确率和mAP更高。并且这四个例子都有各共同点，(B)中每张图超过一半的人对关系都预测为是一样的，如4.1.2提到的，每张图片的场景是稳定的，每张图片包含的社会关系几乎是相似或者是一样的。因此，本文提出的模型恰好能利用上这个线索，对于第样例(a)来说，有两类关系，*Commercial*和*No Relation*，橘色包围盒的人和绿色包围盒的人之间的关系在GRM中错误的预测为*couple*，但是PPRN能

正确的分类。就这张图片来说，PPRN的准确率是100%，但是GRM是33.3%。

4.7 本章小结

本章节主要介绍和分析了PPRN模型在PISC和PIPA-relation两个社会关系理解数据集上的实验结果。

首先，本章介绍了两个大型数据集的图片来源、构造方式、关系类别的划分，其中PISC数据集包含两种关系粒度，PIPA-relation包含5个关系域和这5个关系域对应的15种关系类别与此同时，本文统计了各检测个数据集划分的图片数量，人对数量。本章进一步分析数据集的特点，具体来说包括两个指标：包含多个人对的图片比例，以及包含多种关系类别的比例。同时，本章给出了实验结果的评价方法mAP，取了一个例子来说明在物体领域怎么计算mAP。

接着，本章给出了模型的参数设置、不同训练步骤的优化方法。

然后，我们分别在这两个数据集上进行了社会关系检测这项社会关系理解任务，给出了对比实验组的设定，详细说明了对比工作引入了哪些信息、有哪些设定。在实验结果中，我们分别给出了每个关系类别的召回率和测试集上的mAP。从实验结果来看，融合关系上下文的模型，结合同一图像中其它关系的约束能提高关系的表达能力，在PISC-fine和PIPA-relation的实验结果超过了设置的对照组。接着，本章对消息池化模块采用了不同的池化方法进行对比，发现现有的池化方法超过了常见的平均池化和最大池化。另外，在消息传递机制之后，本章实现了一个结合周边物体信息模块，发现这部分的信息加入后并没有进一步提高实验结果。通过之后的分析，本文发现不管是引入周边物体的信息还是先验知识，均属于场景信息，但是这部分的信息可以由当前图片中其它人对的关系提供，并且不存在额外的检测标注、噪声引入。

最后，本章进行了案例分析，给出了一些真实测例的结果。

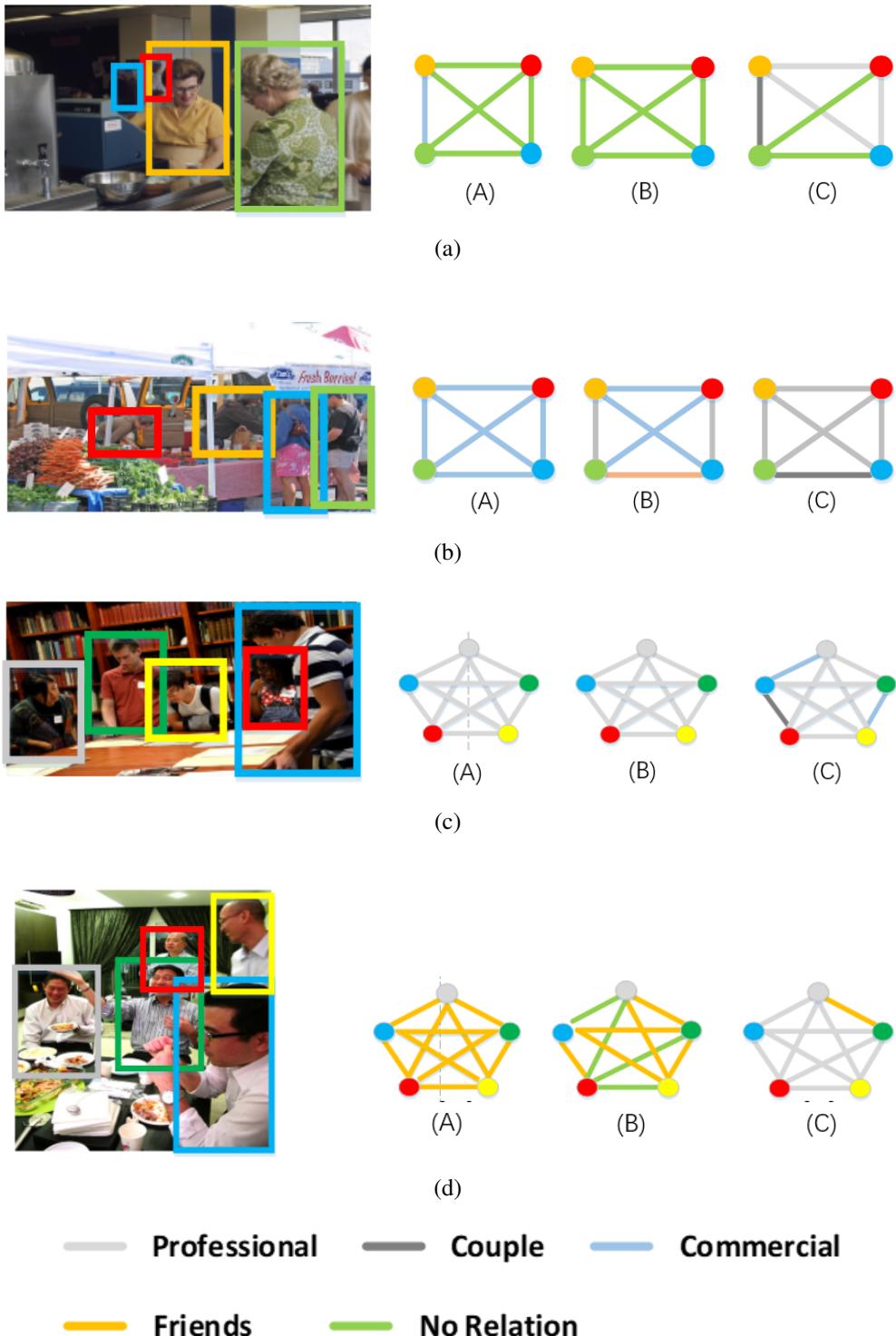


图 4-4 PISC-fine 数据集的部分测试样本

参考文献

- [1] Lu C, Krishna R, Bernstein M S, et al. Visual Relationship Detection with Language Priors [J]. *European Conference on Computer Vision*, 2016: 852–869.
- [2] Johnson J, Krishna R, Stark M, et al. Image retrieval using scene graphs [J], 2015: 3668–3678.
- [3] Wang G, Gallagher A C, Luo J, et al. Seeing people in social context: recognizing people and social relationships [J], 2010: 169–182.
- [4] Li J, Wong Y, Zhao Q, et al. Dual-Glance Model for Deciphering Social Relationships [J]. *International Conference on Computer Vision*, 2017: 2669–2678.
- [5] Sun Q, Schiele B, Fritz M. A Domain Based Approach to Social Relation Recognition [J]. *Computer Vision and Pattern Recognition*, 2017: 435–444.
- [6] Zhang Z, Luo P, Loy C C, et al. Learning Social Relation Traits from Face Images [J]. *International Conference on Computer Vision*, 2015: 3631–3639.
- [7] Wang Z, Chen T, Ren J S J, et al. Deep Reasoning with Knowledge Graph for Social Relationship Understanding [J]. *International Joint Conference on Artificial Intelligence*, 2018: 1021–1028.
- [8] Xu D, Zhu Y, Choy C B, et al. Scene Graph Generation by Iterative Message Passing [J]. *Computer Vision and Pattern Recognition*, 2017: 3097–3106.
- [9] Robicquet A, Sadeghian A, Alahi A, et al. Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes [J], 2016, 9912: 549–565.
- [10] Chen Y, Hsu W H, Liao H M. Discovering informative social subgraphs and predicting pairwise relationships from group photos [J], 2012: 669–678.
- [11] Qin Z, Shelton C R. Improving multi-target tracking via social grouping [J], 2012: 1972–1978.
- [12] Direkolu C, Connor N E O. Team activity recognition in sports [J], 2012: 69–83.
- [13] Lan T, Sigal L, Mori G. Social roles in hierarchical models for human activity recognition [J], 2012: 1354–1361.

- [14] Lan T, Wang Y, Yang W, et al. Discriminative Latent Models for Recognizing Contextual Group Activities [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34 (8): 1549–1562.
- [15] Deng Z, Vahdat A, Hu H, et al. Structure Inference Machines: Recurrent Neural Networks for Analyzing Relations in Group Activity Recognition [J]. computer vision and pattern recognition, 2016: 4772–4781.
- [16] Alahi A, Goel K, Ramanathan V, et al. Social LSTM: Human Trajectory Prediction in Crowded Spaces [J], 2016: 961–971.
- [17] Marino K, Salakhutdinov R, Gupta A. The More You Know: Using Knowledge Graphs for Image Classification [J]. computer vision and pattern recognition, 2017: 20–28.
- [18] Liang K, Guo Y, Chang H, et al. Visual relationship detection with deep structural ranking [C]. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
- [19] Dibeklioglu H, Salah A A, Gevers T. Like Father, Like Son: Facial Expression Dynamics for Kinship Verification [J], 2013: 1497–1504.
- [20] Xia S, Shao M, Luo J, et al. Understanding kin relationships in a photo [J]. IEEE Transactions on Multimedia, 2012, 14 (4): 1046–1056.
- [21] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [J]. computer vision and pattern recognition, 2016: 770–778.
- [22] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors [J]. Nature, 1988, 323 (6088): 696–699.
- [23] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization [J]. Journal of Machine Learning Research, 2011, 12 (Jul): 2121–2159.
- [24] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv:1412.6980, 2014.
- [25] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86 (11): 2278–2324.
- [26] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks [J], 2012: 1097–1105.
- [27] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. international conference on learning representations, 2015.

- [28] Zheng S, Jayasumana S, Romeraparedes B, et al. Conditional Random Fields as Recurrent Neural Networks [J]. international conference on computer vision, 2015: 1529–1537.
- [29] Liang X, Shen X, Feng J, et al. Semantic Object Parsing with Graph LSTM [J]. european conference on computer vision, 2016: 125–143.
- [30] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9 (8): 1735–1780.
- [31] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning Phrase Representations using RNN-Encoder–Decoder for Statistical Machine Translation [J]. empirical methods in natural language processing, 2014: 1724–1734.
- [32] Girshick R B, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [J]. computer vision and pattern recognition, 2014: 580–587.
- [33] Girshick R B. Fast R-CNN [J]. international conference on computer vision, 2015: 1440–1448.
- [34] Ren S, He K, Girshick R B, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J], 2015, 2015: 91–99.
- [35] Ding L, Yilmaz A. Learning relations among movie characters: a social network perspective [J], 2010: 410–423.
- [36] Ramanathan V, Yao B, Feifei L. Social Role Discovery in Human Events [J], 2013: 2475–2482.
- [37] Zhang N, Paluri M, Taigman Y, et al. Beyond frontal faces: Improving Person Recognition using multiple cues [J]. computer vision and pattern recognition, 2015: 4804–4813.
- [38] Lin T, Maire M, Belongie S J, et al. Microsoft COCO: Common Objects in Context [J]. european conference on computer vision, 2014: 740–755.
- [39] Li Y, Tarlow D, Brockschmidt M, et al. Gated Graph Sequence Neural Networks [J]. arXiv: Learning, 2016.
- [40] Krishna R, Zhu Y, Groth O, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations [J]. International Journal of Computer Vision, 2017, 123 (1): 32–73.
- [41] Bugental D B. Acquisition of the algorithms of social life: a domain-based approach. [J]. Psychological Bulletin, 2000, 126 (2): 187–219.

致 谢

谨此向我的导师张三教授致以衷心的感谢和崇高的敬意！本论文的工作是在张老师的悉心指导下完成的。在传授予我专业知识和宝贵经验的同时，张老师以其严谨的治学态度和精益求精的工作作风不断促进论文相关工作的进行，使我受益匪浅。

在攻读硕士的这三年里，导师和实验室的同学们不仅为我创造了优越的科研和学习环境，使我得以在计算机科学领域中自由翱翔，同时在思想上、人生态度和意志品质方面给予了谆谆教诲，这些教益必将激励着我在今后的人生道路上奋勇向前。特别感谢实验室的甲师兄、乙同学以及其他师弟师妹，他们不仅在学术上给了我许多指引和建议，而且在生活上予以帮助，从他们身上我学到了很多知识。

感谢王五老师及其实验室的同学在领域一、领域二方面的学习给予我的帮助。他们开创性的研究拓展了我的学术视野，无数次的争论和探讨使我的研究工作有了长足的进展。

由衷感谢我的室友A、B和C同学，以及其他经常到我们宿舍进行学习交流的D、E、F和G同学，是他们令我的学习生活都更加充满动力。衷心的感谢我的父母和其他亲朋好友对我的关心、支持和理解，没有他们对我的关心、鼓励和支持，我无法完成现在的硕士学业。

最后，感谢所有曾经教育和帮助过我的所有老师。衷心地感谢为评阅本论文而付出宝贵时间和辛勤劳动的专家和教授们！

李四

二零一四年三月二十四日