中山大学硕士学位论文

基于消息传递机制的社会关系检测的算法研究

Multiobjective Optimization Algorithms Based on Hybrid Local Search for Multiobjective Pickup and Delivery Problem with Time Windows

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送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检测的算法研究
专 业: 软件工程
硕 士 生: 李雷来
指导老师: 万海 副教授

摘 要

摘要内容。

关键词: 关键词

Title: Multiobjective Optimization Algorithms Based on Hybrid Local Search for Multiobjective Pickup and Delivery Problem with Time Windows

Major: Computer Technology

Name: Xiaoxin Li

Supervisor: Prof. Hai Wan

Abstract

英文摘要

Keywords: 英文关键词

目 录

第 1 章　引言

根据天津大学模板修改的符合中山大学毕业论文（至少是硕士论文）要求的Latex模板。

1.1　使用方法

本模板只包括内容方面的设计预定义，编译自行解决。作者使用的是Windows环境下MikTex+TeXstudio的组合。

1.2　使用建议

1.2.1　普适问题

普遍适用的论文排版问题：

- 图片标题在下，表格在上；一定要有标题，不能只是图1-1；与文字内容的间隔自行把握。
- 参考文献建议使用.bib文件；也有使用Google Scholar的引用的，但有指出当中的“//”不符合规范。
- 部分评审反馈，目录不包含摘要及目录本身，请根据情况自行斟酌。
- 打印时需要右边翻页的问题（每章开始在右边页），可以在生成pdf后通过插入空白页解决（这样插入不会改变页码）；或者尝试设置openright（未测试，有待探讨）。

1.2.2　细节问题

一些细节的问题建议：

- 每个章节都有label，key使用ch:intro形式，以下使用sec:background等。图片key可以参考fig:scenes，表格参考tab:exp。
- 图片、表格尽量在页的顶部，即float优先选择t。

图 1-1 图例

表 1-1 示例表

表头	栏1	栏2	栏3	栏4
内容1	b	—	768×576	19
内容1	a	240/7	768×576	—

- 另外，为了打印时彩打方便，可以把需要彩打的图片尽量排版在一页，不过比较难调。
- 虽然每个body的tex文件中包含了!Mode:: “TeX:UTF-8”在文件开头，但仍有必要在IDE中将新建的tex文件设为UTF-8 编码，否则可能无法正常显示中文。

1.2.3 其他说明

参考文献[1]目前采用上标表示。使用cite命令。

目前页眉设置：每章第一页页眉只有中间的“中山大学硕士毕业论文”，后续页左边显示“中山大学硕士毕业论文”，右边显示“第n章”。

目前页脚设置：仅包含页码，居中，无横线。

参考文献和附录计算页数，包含在目录，页眉设置同每章第一页。正文前的部分无页眉。

1.3 例子

图例子。label要在caption后。多图或子图方法上网查吧。

表例子。推荐使用这种三行表。缺省值使用三个“-”产生长横线“—”。

公式例子，与普通Latex数学公式无异。

$$1 + 1 = 2 \tag{1-1}$$

1.3.1 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，但是我们能明确的扑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和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[2?]、多目标追踪[3,4]和群体活动识别[5–7]。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[8]。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）[9]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[8]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[10] 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1-2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 1-2 PISC数据集中的一些图片例子

既然社会关系理解对于理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[11]和PISC-relation[12]。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[12,13]。还有利用周边环境的信息的模型[11,14]，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[9]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）[15]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”，那么第三条关系的的社会关系会倾向也是“朋友”或者其他的亲密关系，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。本论文完全区别于Li（2017）[11]和Wang（2018）[14]的工作，没有引入额外的检测标注，但是采用和Li等特征提取方法相同的策略。论文的切入点如图例1-3，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）。

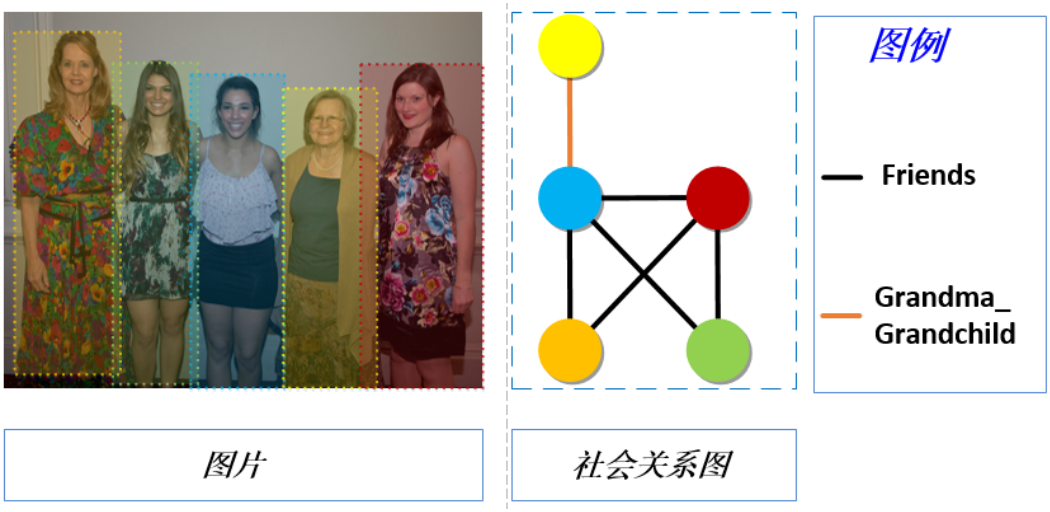

图 1-3 本论文动机的示例图，该图片来自PIPA-relation数据集，其中图片中对应阴影颜色的人对应社会关系图的部分，图上节点间的边表示他们之间的社会关系

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：

(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社会关系，但是场景图谱中并没有人的的类别的概念，社会关系图中节点间的社会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对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。

第 2 章　预备知识

2.1　图像的视觉信息抽取

2.1.1　神经网络

2.1.1.1　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。 理想中的激活函数是如

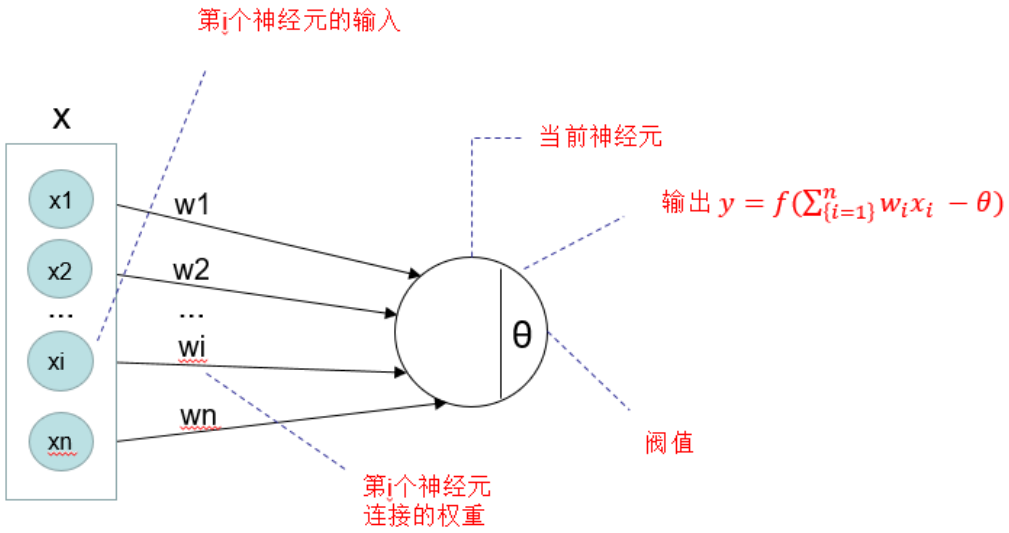

图 2-1　神经元模型示意图

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。

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

例如sigmoid函数可以把Φ − ∞, +∞Ψ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是，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	$\sigma(x) = \frac{1}{1+e^{-x}}$	$f^x(x) = \sigma(x)(1-\sigma(x))$	(0,1)
tanh函数	$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	$f'(x) = 1 - tanh(x)^2$	(-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$H_\theta(x) = f_3(w_3 f_2(w_2(f_1(w_1 x + b_1)) + b_2) + b_3)$, 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2-3，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。

$$w_i \leftarrow w_i + \triangle w \tag{2-2}$$

不同层之间的梯度通过误差逆传播（error Back Propagation 简称BP）[16]算法进行整个网络的学习。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[17]和Adam[18]，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。

$$L(\mathbf{X}, \mathbf{Y}) = \sum_{(x_i, y_i) \in (\mathbf{X}, \mathbf{Y})} (y_i - H_x(x_i)) \tag{2-3}$$

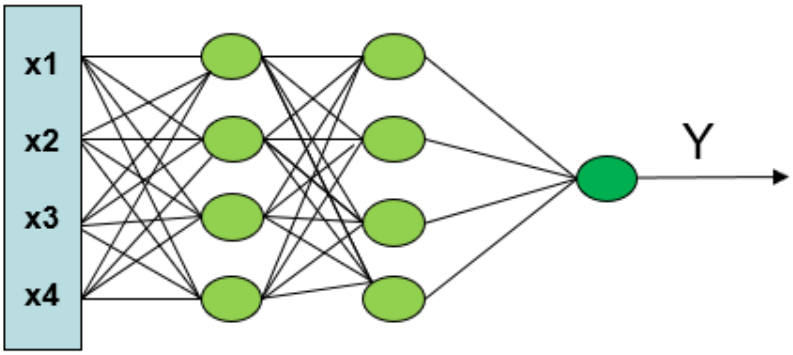

图 2-2　两个隐藏层的神经网络示意图

2.1.1.2　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$6\times6\times3$，其中图片输入的3为图片的通道数、6×6为高宽，假设滑动的步骤为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）。

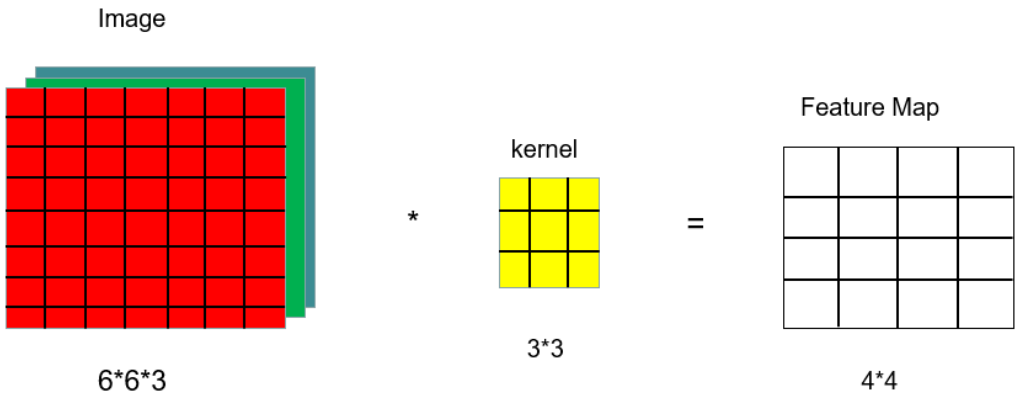

图 2-3 卷积神经网络的卷积层示意图

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，滑动的步骤为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[19]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[20]，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[21]，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[22]是何凯明等人(2016)提出的，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

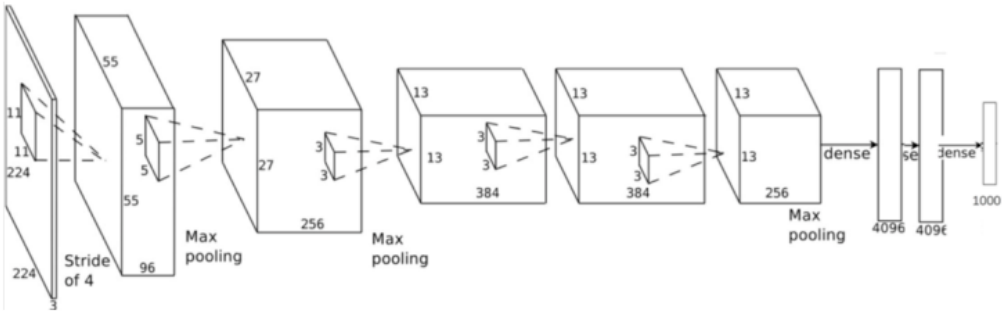

图 2-5 AlexNet网络结构示意图

叠加深网络的深度但是不降低网络性能。

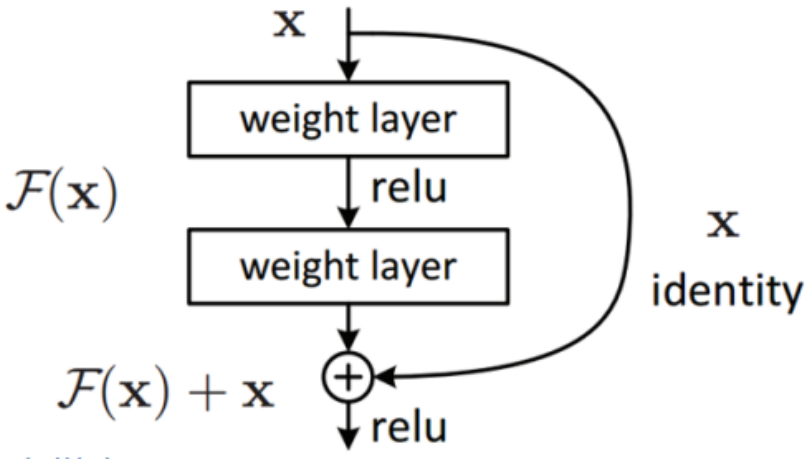

图 2-6 ResNet残差学习单元示意图

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。

2.1.2 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[23]，fast-RCNN[24]，以及Ren等人（2016）[25]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[23]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。

2.2 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。

2.2.1 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）[10]，以及Dibeklioglu等人（2013）[26]和Zhang 等人（2015）[13]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）[11]提出的多次观察的**dual-glance**模型。以及Wang 等人（2018）[14]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ding2010learning等人（2010）、Ramanathan等人（2013）[27]。Sun等人（2017）[12]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[28]。

Zhang 等[13]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x^r, x^l，并且$\forall x^r, x^l \in R^{2048\times1}$，经过一个权重矩阵$\mathbf{W} \in R^{4096\times256}$得到特征向量$x_t$。对于已经标注好的人脸图片两个人脸分别为和，利用DCN 提取得到的特征分别记为和，并且。除了图片中本来的特征，模型利用了两张人脸在图片的空间信息。1）两张人脸的位置分别表示为$x^l, y^l, w^l, h^l, x^r, y^r, w^r, h^r$，其中$x-, y-$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$x_t$拼接来学习得到关系类别。 除此之外，$\mathbf{w_{gi}}$，$\mathbf{W}$，

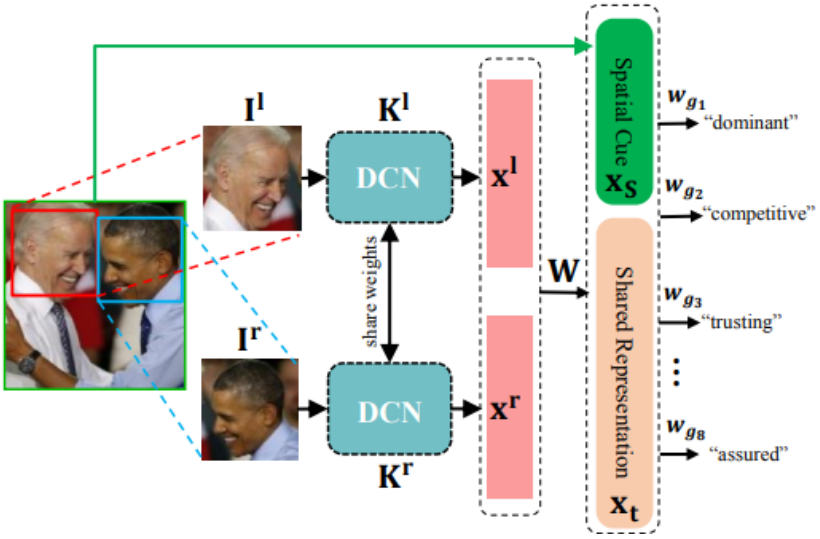

图 2-7　zhang的模型

$\mathbf{K^l}$，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，该模型损失函数定义如下：

$$arg\max_{\Omega} p(\mathbf{w}_{g_i}{}_{i=1}^{8}\mathbf{W}\mathbf{K}^r|\mathbf{g}\mathbf{x}_t\mathbf{x}_s\mathbf{I}^r\mathbf{I}^l) \propto$$
$$(\sum_{i=1}^{8} p(g_i|x_t,x_s)p(w_(g_i)))(\sum_{j=1} Kp(k_j^l)p(k_j^r))p((W)), \tag{2-4}$$
$$s.t.\mathbf{K}^r = \mathbf{K}^l$$

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）[11]基于之前的工作，针对社会检测的任务提出了包含两个关系粒度的数据集，PISC-coarse和PISC-fine（2017）[11]。该工作首次提出了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**I** 和两个人身体的包围盒。针对图片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_u。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**g**，经过预训练的ResNet[22] 抽取得到表示人对关系的特征向量为v，经过拼接和全连接网络后得到这个人对的特征向量v_{top}。

对于second glance模块，模型利用faster-RCNN[25]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2-5，其中函数$G(b1,b2)$表示两个包围盒的IOU，τ_u是阀值。

$$R(b1,b2;I) = \{c \in P_I : max(G(c,b1),G(c,b2)) < \tau_u\} \tag{2-5}$$

利用faster-RCNN得到的特征图，采用ROI pooling抽取出固定长度的的向量物体特征向量$v \in R^k$。同时用$\{v_i|i = 1,2,....N\}$作为$R_(v1,b2;I)$中的物体向量集合。然后

依次采用公式2-6方式将v_{top}和物体的向量集合得到h_i。

$$h_i = v_i + w_{top} \otimes v_{top} \tag{2-6}$$

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2-7所示，其中attention的权重$a_i \in [0,1]$

$$\begin{aligned} a_i &= \frac{1}{1+exp(-(W_{h,a}h_i + b_a))} \\ v_i^{att} &= a_i v_i \\ s_i &= W_s v_i^{att} + b_s \end{aligned} \tag{2-7}$$

GRM（2018）[14]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出的特征向量初始化。对于社会关系节点，这里采用Li等人的方式[11]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$f_h, f_h \in \mathrm{R}^d$。$f_h$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[25]，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[29]，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[30]来执行图上的消息传播。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播，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2-8，其中A_v^τ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。

$$a_v^t = A_v^\tau [\ h_1^{t-1} \ ... \ h_{|V|}^{t-1}\ ]^\tau + b \tag{2-8}$$

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2-9，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。

$$\begin{aligned} \mathbf{h}_{ij} &= tanh(\mathbf{U}^a h_{r_i}) \odot tanh(\mathbf{V}^a h_{o_j}) \\ e_{ij} &= \mathbf{Atten}(\mathbf{h}_{ij}) \\ \alpha_{ij} &= \sigma(e_{ij}) \end{aligned} \tag{2-9}$$

2.2.1.1

参考文献

[1] Wu Y, Lim J, Yang M-H. Online object tracking: A benchmark [C]. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013: 2411–2418.

[2] Robicquet A, Sadeghian A, Alahi A, *et al*. Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes [J], 2016, 9912: 549–565.

[3] Chen Y, Hsu W H, Liao H M. Discovering informative social subgraphs and predicting pairwise relationships from group photos [J], 2012: 669–678.

[4] Qin Z, Shelton C R. Improving multi-target tracking via social grouping [J], 2012: 1972–1978.

[5] Direkolu C, Connor N E O. Team activity recognition in sports [J], 2012: 69–83.

[6] Lan T, Sigal L, Mori G. Social roles in hierarchical models for human activity recognition [J], 2012: 1354–1361.

[7] 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.

[8] Lu C, Krishna R, Bernstein M S, *et al*. Visual Relationship Detection with Language Priors [J]. european conference on computer vision, 2016: 852–869.

[9] Johnson J, Krishna R, Stark M, *et al*. Image retrieval using scene graphs [J], 2015: 3668–3678.

[10] Wang G, Gallagher A C, Luo J, *et al*. Seeing people in social context: recognizing people and social relationships [J], 2010: 169–182.

[11] Li J, Wong Y, Zhao Q, *et al*. Dual-Glance Model for Deciphering Social Relationships [J]. international conference on computer vision, 2017: 2669–2678.

[12] Sun Q, Schiele B, Fritz M. A Domain Based Approach to Social Relation Recognition [J]. computer vision and pattern recognition, 2017: 435–444.

[13] 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.

[14] 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.

[15] 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.

[16] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors [J]. Nature, 1988, 323 (6088): 696–699.

[17] 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.

[18] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv:1412.6980, 2014.

[19] 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.

[20] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks [J], 2012: 1097–1105.

[21] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. international conference on learning representations, 2015.

[22] He K, Zhang X, Ren S, *et al*. Deep Residual Learning for Image Recognition [J]. computer vision and pattern recognition, 2016: 770–778.

[23] 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.

[24] Girshick R B. Fast R-CNN [J]. international conference on computer vision, 2015: 1440–1448.

[25] 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.

[26] Dibeklioglu H, Salah A A, Gevers T. Like Father, Like Son: Facial Expression Dynamics for Kinship Verification [J], 2013: 1497–1504.

[27] Ramanathan V, Yao B, Feifei L. Social Role Discovery in Human Events [J], 2013: 2475–2482.

[28] 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.

[29] Lin T, Maire M, Belongie S J, *et al*. Microsoft COCO: Common Objects in Context [J]. european conference on computer vision, 2014: 740–755.

[30] Li Y, Tarlow D, Brockschmidt M, *et al*. Gated Graph Sequence Neural Networks [J]. arXiv: Learning, 2016.

[31] Cho K, Van Merrienboer 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.

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

二零一四年三月八日