硕士学位论文

基于图表示和注意力机制的行人属性识别算法研究

Pedestrian Attribute Recognition Algorithm Based on Graph Representation and Attention Mechanism

作者：戚余航

导师：郭宇春

北京交通大学

2021 年 5 月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通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，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

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

学校代码：10004 密级：公开

北京交通大学

硕士学位论文

基于图表示和注意力机制的行人属性识别算法研究

Pedestrian Attribute Recognition Algorithm Based on Graph Representation and Attention Mechanism

作者姓名：戚余航 学　　号：18120118

导师姓名：郭宇春 职　　称：教授

学位类别：工学 学位级别：硕士

学科专业：通信与信息系统 研究方向：信息网络

北京交通大学

2021 年 5 月

致谢

本论文的研究工作是在我的导师郭宇春教授的悉心指导下完成的。郭宇春老师治学态度严谨，对待科研与教学工作一丝不苟，精益求精，对待学生认真负责。攻读硕士学位的三年当中，郭宇春老师始终都给予我悉心的指导和不懈的支持，鼓励我在科研工作中不断的探索创新。在此谨向郭老师致以诚挚的谢意和崇高的敬意。除了郭宇春老师之外，陈一帅副教授也给了我许多帮助，陈一帅老师对科研工作的热情深深的感染了我，为我树立了一个很好的榜样，在此也向陈老师表达我最诚挚的谢意。

另外，在实验室的科研工作与日常生活中，王珍珠同学，孙欢同学，李想同学，尹姜谊师姐，冯梦菲师姐，于滋灏师兄，周雪瑞师姐等人也帮助我解决了很多学习生活上的问题，在此一并向他们表达我的感谢之意。

最后，我特别要感谢我的家人，正是他们多年以来的默默付出，以及坚定不移的支持，才能够支撑我一直走到今天顺利完成学业。在此，我向我的家人们表达由衷的谢意。

摘要

行人属性识别是视频监控领域中的重要任务，因其在视频监控应用中的巨大潜力在近些年来受到了广泛关注，它可以应用于很多下游任务中，例如行人追踪，人脸验证，行人检索等。

行人属性识别任务属于计算机视觉领域中的多标签图像分类任务。多标签图像分类任务的研究重点在于如何利用标签之间的共现关系辅助模型提升分类性能。目前最先进的共现关系建模方法是图神经网络，但它存在两个严重影响性能的问题：1) 低频标签无法有效的从高频标签中迁移特征信息导致低频标签分类性能不佳。2) 大量图节点不参与图神经网络的信息传递过程，导致图神经网络发生严重的退化。对于图神经网络存在的问题，本论文通过分析行人属性出现的频率和条件概率，提出了一种新的定义共现关系图的方法，并与原有方法相结合，提出了一个基于非对称共现依赖关系图的行人属性识别模型，这个模型能够提升低频行人属性的识别性能并有效的缓解图神经网络的退化问题。

注意力机制也是提升行人属性识别任务性能的重要手段之一。然而，由于行人图片质量不佳以及缺乏有效的监督信号等问题，现有的基于注意力机制的模型在行人属性识别任务中通常无法准确的定位与特定行人属性高度相关的图像特征。对于注意力机制存在的问题，本论文使用一种基于词向量语义指导的空间注意力机制模块改进行人属性识别模型，词向量包含了行人属性的语义信息，可以作为一种先验知识对缺乏监督信号的问题进行弥补，并指导注意力机制模块准确的定位与特定行人属性高度相关的图像特征。

本文的主要贡献如下：

(1) 提出了一种基于非对称共现依赖关系图的模型，并在多个数据集上使用多个主干网络进行实验，实验结果表明该方法能够将 mAP 提升 0.6%～3.3%。

(2) 提出了一种基于词向量语义指导的空间注意力机制模块的行人属性识别模型，在行人属性识别数据集中相比于常规图像分类模型以及不使用词向量语义指导的注意力机制模型来说，mAP 提升 1.1%~1.9%。

(3) 将两种改进方法进行结合，相比于单独使用某一种改进方法能够将 mAP 进一步提升 0.4%~0.6%。

图 14 幅，表 7 个，参考文献 53 篇。

关键词：行人属性识别；图神经网络；共现关系；词向量；空间注意力

ABSTRACT

Pedestrian attribute recognition is an important task in the field of video surveillance. It has been widely concerned in recent years because of its great potential in video surveillance applications. It can be applied to many downstream tasks, such as pedestrian tracking, human face verification, pedestrian retrieval and so on.

Pedestrian attribute recognition is a multi-label image classification task in the field of computer vision. Today, the research focus of multi-label image classification is how to use the co-occurrence relationship between labels to improve the classification performance. The most advanced co-occurrence relationship modeling method is graph neural network. Graph neural network has two problems that seriously affect the performance: 1) low frequency labels can not effectively transfer feature information from high frequency labels, which leads to poor classification performance of low frequency labels. 2) A large number of graph nodes do not participate in the information transmission process of graph neural network, which leads to serious degradation of graph neural network and limits the performance improvement. For the problems of graph neural network, this paper proposes a new method of defining co-occurrence graph by analyzing the frequency and conditional probability of pedestrian attributes. Combined with the original method, a pedestrian attribute recognition model based on asymmetric co-occurrence dependency graph is proposed. This model can improve the recognition performance of low-frequency pedestrian attributes and effectively alleviate the degradation problem of graph neural network.

Attention mechanism is also one of the important methods to improve the performance of pedestrian attribute recognition. However, due to the poor quality of pedestrian images and the lack of effective supervision signals, the existing models based on attention mechanism can not accurately image features highly related to specific pedestrian attributes. For the problems of attention mechanism, this paper proposes a spatial attention mechanism module based on the semantic guidance of word embedding to improve the pedestrian attribute recognition model. Word embedding contains the semantic information of pedestrian attributes, which can be used as a priori knowledge to make up for the lack of supervision signals, guide the attention mechanism module to accurately locate the image features highly related to specific pedestrian attributes.

The main contributions of this paper are as follows:

(1) A model based on asymmetric co-occurrence dependency graph is proposed, and experiments are carried out on multiple datasets using multiple backbone networks. The experimental results show that this method can improve mAP by 0.6% ~ 3.3%.

(2) A pedestrian attribute recognition model based on spatial attention mechanism module guided by word embedding semantics is proposed. Compared with the conventional image classification model and the attention mechanism model without word embedding semantics, the mAP of pedestrian attribute recognition task is improved by 1.1% ~ 1.9%.

(3) The combination of the two improved methods can further improve the mAP by 0.4% ~ 0.6% compared with using only one improved method.

14 figures, 6 tables, and 53 references.

KEYWORDS：Pedestrian attribute recognition; Graph neural network; co-occurrence; Word embedding; Spatial attention mechanism

目录

1 引言

1.1 研究背景和意义

行人属性识别是计算机视觉领域中的一项重要任务。它指的是在某个包含行人的图片中提取行人的属性信息，属性信息主要包括性别，年龄，着装信息等。行人属性识别在视频监控领域中具有巨大的潜力，它可以在行人检索，行人重识别，人体动作识别等应用中发挥巨大的作用，因此，在近些年来，收到了广泛的关注。

在基于神经网络的深度学习技术流行起来之前，行人属性识别任务主要是通过传统的图像特征提取算法来完成，比如 HOG，LBP 特征提取算法等。然而，行人属性信息与传统的低维图像特征并不相同，它属于高维语义信息。在观察点或观察条件（视角，光照等）变化的情况下，高维语义特征具有更强的鲁棒性，而传统低维特征却会受到非常大的影响。因此，传统的图像特征提取算法并不能够满足行人属性识别任务的要求。然而，随着第三次人工智能浪潮的来临，行人属性识别领域乃至整个计算机视觉领域都迎来了一次彻底的革新。

近些年来，随着硬件计算能力的提升，基于神经网络的深度学习技术获得了蓬勃的发展。在计算机视觉领域，基于卷积神经网络的深度学习技术在各个研究方向中都展现出了巨大的潜力，行人属性识别识别也不例外。现如今的行人属性识别领域中，在性能上能够达到工业标准的方案，几乎无一例外都是依赖于基于卷积神经网络的深度学习技术而定制的。

目前，计算机视觉领域中多标签图像分类模型是解决行人属性识别问题的主流方法。通常，多标签图像分类模型使用卷积神经网络提取图像特征，并使用提取到的图像特征进行加工分类。然而，将现有的多标签图像分类模型直接应用于行人属性识别任务中存在两个问题。首先，在行人属性识别数据集上，各个属性之间具有很强的依赖关系，例如，上装类型和下装类型的关系，性别与发型的关系等，合理的利用这种关系可以提升属性识别的准确度。然而，现有的多标签图像分类模型并不能完全的，准确的建模这种依赖关系，导致属性识别准确度较低。另外，在行人属性识别数据集上，各个属性所处位置有一定的规律，将这种规律信息引入进行人属性识别中会提升属性识别的准确度。虽然现有的部分多标签图像分类模型考虑了空间位置信息，但是对于这种规律的利用仍然不够充分，这也影响了属性识别的准确度。综上所述，本课题以解决上述两个问题为目标，在现

有的多标签分类算法的基础上针对上述两点问题进行创新性改进。因此，本课题对于其下游具体任务具有重要的意义和价值。

1.2 国内外研究现状

本节将会介绍国内外研究现状。行人属性识别的研究历程可以从总体上概括为两个阶段，分别为基于传统计算机视觉方法的阶段和基于深度学习方法的阶段。本节首先介绍传统的行人属性识别方法，再介绍基于深度学习的行人属性识别方法，最后简述目前基于深度学习的方法的研究重点——共现依赖关系。

1.2.1 传统行人属性识别方法

近些年来，随着硬件计算能力的提升，基于神经网络的深度学习技术获得了蓬勃的发展。在基于人工神经网络的深度学习技术获得蓬勃发展之前，行人属性识别任务通常使用两大类方法来完成。第一大类方法主要是基于传统的计算机视觉特征提取算法而设计的。第二大类方法主要是基于统计学习理论而设计的。

传统的计算机视觉特征提取方法主要包括 HOG[1]，SIFT[2]等。这类方法主要致力于提取与行人属性相关的像素特征。传统计算机视觉特征提取算法的局限性在于只能提取低维像素特征。而行人属性属于高维语义特征，它的像素特征可能会随着观察角度，光照条件，障碍物遮挡，图像分辨率等一系列因素而发生变化，因此，行人属性在像素特征上不具有鲁棒性，进而，致力于提取图像像素特征的传统计算机视觉特征提取算法并不能够完美的处理行人属性识别任务。但是仍然不可以忽视传统计算机视觉特征提取算法在行人属性识别领域的发展历史中所起到的作用，它给后来的研究者们最大的启发就是为他们指明后续的研究方向——提取行人属性的高维语义特征。随着基于神经网络的深度学习技术的兴起，研究者们迅速的将神经网络应用于提取行人属性的高维语义特征上，并获得了良好的性能。之所以能够在深度学习时代迅速找到行人属性识别研究的切入点，得益于传统计算机视觉特征提取算法为研究者们指明的方向。

基于统计学习理论的方法主要包括 SVM[3]，CRF[4]。这类方法主要致力于根据统计信息为行人属性特征寻找具有鲁棒性的分类器。基于统计学系理论的方法的局限性在于过度依赖于行人属性识别数据集中的统计信息。即使是在现代，行人属性识别领域中的基准数据集的规模也往往很小，少则几千张图片，多则也仅仅能够达到数万张图片。然而，相比于行人属性识别应用中的庞大的数据规模，现有的行人属性识别基准数据集根本无法概括其统计特性。因此，基于统计学习理

论的方法也不能够很好的适应行人属性识别任务的需求。另外，在上文中曾经强调过，行人属性属于高维语义特征，高维语义特征包含很多信息，其中不仅包括对低维像素信息的概括，也包含了统计信息，以及其他的各种不同类型的信息。所以，基于统计学习理论的方法可以看作是行人属性识别领域的发展历史中对学习高维语义特征信息的一次尝试，具有重要的指导意义。所以，基于统计学习理论的方法在深度学习时代也为研究者们提供了可供参考的解决问题的思路，这类方法对于行人属性识别研究领域具有重大意义。

1.2.2 基于深度学习技术的行人属性识别方法

随着硬件的计算能力的不断提升，一种早已被提出，但在当时受限于硬件计算能力的技术又重新回到了研究者们的视线当中，这种技术就是人工神经网络。在 2012 年，随着 AlexNet[5]的出现，大量的基于人工神经网络的深度学习技术如雨后春笋一般涌现。人工神经网络尤其是卷积神经网络的出现彻彻底底的颠覆了计算机视觉领域，计算机视觉领域中诸如图像分类，目标检测，语义分割等更细化的研究领域以及从这些领域中延伸出来的应用问题都经历了从传统计算机视觉算法向基于人工神经网络的深度学习技术的过渡。

行人属性识别本质上属于计算机视觉领域中的图像分类问题。但是，一般的图像分类问题中，一张图片只有一个对应的类别。然而，在行人属性识别中，一张图片中的单个行人可能具有多个不同的类型的属性。所以行人属性识别任务需要解决的并不是一般的图像分类问题，而是多标签图像分类问题，多标签就对应着多个行人属性。

在基于人工神经网络的深度学习技术重新兴起的初期，深度卷积神经网络[6-11]在一般的图像分类问题中表现的非常出色，其性能大幅度的超越了传统的计算机视觉算法。得益于此，研究人员将多标签分类问题进行分解，将其分解成多个一般的图像分类问题，然后采用基于深度卷积神经网络的一般的图像分类问题的处理方法来处理多标签图像分类问题。具体来说，通常是训练一个基于卷积神经网络的深度学习模型，这个模型通常包含两个部分。一部分是深度卷积神经网络，这一部分用于提取图片的高维语义特征，另一部分为分类器组，这一组分类器中的每一个成员，都负责某一个特定标签的分类任务，各个分类器之间使完全独立的，不会发生相互影响。

这种深度学习模型有两个极为突出的优点，首先，在分类器组中，所有的分类器都共享同一个深度卷积神经网络主干所提取的图像高维语义特征，这样共享不仅仅能够避免模型参数量过大，同时也可以使模型具有可扩展性，原因是如果

标签的种类或数量发生变化，只需要对分类器组部分进行调整，由于分类器组中的成员之间完全独立，发生变化的类别所对应的分类器成员也不会影响到其他未发生变化的类别所对应的分类器成员。第二个优点是对于深度卷积神经网络主干，研究者们可以采用迁移学习的技术，首先在一些超大规模的数据集上进行预训练，比如 ImageNet[12]，然后再将深度卷积神经网络主干在多标签分类数据集上进行微调，从而将一些从超大规模数据集上学习到的知识迁移到多标签分类任务中。对于图像数据来说，纹理，颜色，轮廓等低维信息在各个不同的数据集之间相似，因此，这种迁移学习的方式相当于增加了先验知识，对多标签图像分类数据集中的信息做了补充，所以取得了良好的性能效果，如果多标签图像数据集本身的数据比较稀疏，或者数据集本身稀缺这种通用的底层低维特征，那么这种迁移学习的方式对性能的提升将会更加的明显。

然而，这种模型虽然在性能上能够远超传统的计算机视觉算法，并且具有诸多优点，但是随着深度卷积神经网络的发展速度趋于平缓，这种模型能够达到的性能上限也逐渐显现。因此，研究人员们开始探索这种模型中，除了深度卷积神经网络之外，限制性能的瓶颈究竟在哪里？在这种模型中去除深度卷积神经网络部分后，就只剩下了分类器组部分，虽然通过将分类器组中的分类器独立开来可以获得使模型具有良好的可扩展性的优点，但是这样做却是基于一个错误的前提假设，那就是：多标签图像中的各个标签是相互独立的。实际上，在多标签图像识别任务中，各个标签可能并不是独立存在的，而是具有相互依赖的关系。比如，在 MS-COCO 多标签分类数据集[13]中存在的标签有 bowl，cup，dining table，chair 等标签，它们显然不是独立存在的，可以想象这样一幅画面，一个餐桌上摆放着盘子，碗，杯子等餐具，餐桌旁还放有几把椅子，实际上，想象的这幅画面其实真真切切的出现在了数据集中。在行人属性识别数据集中，比如 PA-100K[14]，各个行人属性之间具有很强的依赖关系，例如，上装类型和下装类型的关系，性别与发型的关系，着装颜色之间的关系等。具体来说，比如，对于上装类型为西装的行人，他的下装极有可能是西裤，而不太可能是短裤等下装类型，那么即使图片中存在与行人下装类型相关的像素被其他障碍物部分遮挡的情况，也可以根据少部分像素信息与西装和西裤在语义上的依赖（搭配）关系，推测出行人的下装类型究竟是什么。同理，如果模型已经可以准确的判断出一个行人具有“长发”这个行人属性，那么该行人是女性的可能性一般要大于该行人是男性的可能性。那么，即使行人是背对于图像的观察点，仍然可以依据“长发”这个行人属性，以及其他的诸如着装等行人属性信息，来推测出该行人的具体性别。可见，无论是传统的多标签图像分类数据集，还是具有更强的属性间依赖关系的行人属性数据集，标签或者行人属性之间都存在着复杂的依赖关系。

在基于深度卷积神经网络的深度学习技术框架下，应该如何学习这种依赖关系呢？深度卷积神经网络作为一种监督学习技术，需要人工标注的监督信息作为学习目标，然而，研究者们所希望学习到的这种依赖关系是非常复杂的，它取决于标签或行人属性之间的语义角度上的联系，是一种脱离了数据集的统计信息以及图像像素信息的外部知识，这种依赖关系是无法像标签或行人属性一样使用由 0 和 1 组成的二值向量来表示的，甚至是无法使用数值向量来表示，也自然无法作为监督信息引导模型学习。其实，这种标签或行人属性之间的依赖关系甚至没有一个统一的定义标准，它取决于观察者的主观意识，不同的观察者极有可能给出不同的关系描述，这也使得这种依赖关系的学习更加的抽象和困难。

由于标签或行人属性之间的复杂的关系无法使用监督学习的方式进行建模，研究人员们另辟蹊径，想到了从数据集的统计信息的角度入手。标签或行人属性之间的复杂关系虽然多种多样，但是，无论什么标签或行人属性之间具有什么样的关系，数据集中都存在能够间接表达出这种关系的统计信息——具有依赖关系的标签或行人属性通常极有可能在一张图片中共同出现，研究者们将这种统计信息简称为“共现”关系。共现关系是一种统计信息，也就是说，可以从数理统计的角度去判断标签或行人属性之间是否存在某种依赖关系。最简单，直观的方式，就是观察条件概率。如果想分析标签或行人属性 X 与标签或行人属性 Y 之间是否存在某种依赖关系，可以观察条件概率 $P(X|Y)$以及 $P(Y|X)$，假设标签或行人属性 X 与标签或行人属性 Y 之间没有任何依赖关系，从概率角度来说，又可以称之为独立，那么，一般标签或行人属性 Y 对标签或行人属性 X 的出现概率不会有影响，即 $P(X|Y)$约等于 $P(X)$，同理，标签或行人属性 X 对标签或行人属性 Y 的出现概率不会有影响，即 $P(Y|X)$约等于 $P(Y)$。如果标签或行人属性的出现概率和条件概率不满足以上关系，甚至差别较大，那么标签或行人属性之间极有可能存在某种依赖关系。这种共现关系有一个特性，那就是“具有依赖关系”相对于“共同出现在同一张图片内”是即不充分也不必要的。不具有依赖关系的标签或者行人属性也有可能出现在同一张图片内，具有依赖关系的标签或者行人属性也未必会共同出现在同一张图片内。因此，仅凭单张图片并不能够反映绝对的共现关系，但是基于数理统计信息的条件概率可以度量这种共现关系。

至此，研究人员们将复杂多样的标签或行人属性依赖关系统一的抽象化为共现关系，并尝试对这种共现关系建模。随着深度学习技术的逐步发展，越来越来多的技术用于建模这种共现关系，从早期的循环神经网络，注意力机制，到近期的图神经网络，都在共现关系的建模中取得了实质性的突破。下一节将会简单介绍这些针对于标签或行人属性之间的共现关系的建模方式。

1.2.3 基于共现依赖关系的多标签图像识别与行人属性识别方法

近几年来，在图像的多标签分类问题以及行人属性识别研究领域中，研究者们已经将研究重点放在建模标签或行人属性的共现关系上。与此同时，最早诞生于自然语言处理领域中的注意力机制也逐渐在计算机视觉领域中获得了广泛的应用，并且取得了非常不错的效果。在 2019 年之前，循环神经网络是建模标签或行人属性之间的共现关系的主要方法[15,16]，注意力机制[17]也是提升多标签图像分类以及行人属性识别性能的方法之一。近两年来，以图卷积网络为代表的图神经网络也在逐步兴起，它突破了传统的神经网络对输入数据形式的限制，它能够建模非欧式的具有图结构的数据。由于标签与行人属性之间的共现依赖关系可以由图的形式表达，在 2019 年之后，有研究者尝试使用以图卷积网络为代表的图神经网络建模标签或行人属性之间的共现依赖关系[18]。然而，现有的这些模型处理行人属性识别任务时仍然存在两个问题：

(1) 无论是基于循环神经网络的模型还是基于图神经网络的模型，现有的模型结构设计仍存在局限性，导致它们无法准确的建模行人属性之间的共现关系。

(2) 行人属性的空间位置分布以及行人属性之间的相对位置关系存在一定的规律性，现有的基于注意力机制的模型并不具备有关于这种规律性的先验知识，因此这些模型仍然无法精确提取行人属性空间位置信息。

在第三章中将详细描述这些建模标签或行人属性之间的共现关系的方法以及优缺点，并详细阐述这些研究工作对本课题带来的启发。

1.3 研究内容及贡献

本课题的研究目标是设计一个行人属性识别模型，这个行人属性识别模型是在多标签图像分类模型的基础上，针对其建模行人属性共现依赖关系的能力不足和无法精确提取行人属性空间位置信息这两个缺陷进行进研究与改进，提升行人属性识别模型的性能。

1.3.1 共现依赖关系建模研究

目前看来，图卷积神经网络是建模标签或行人属性共现依赖关系的最有效的方法，其性能远超基于循环神经网络的模型。本课题将论文[18]中提出的基于图卷积神经网络的模型作为主要 Baseline 进行研究。对 Baseline 模型进行分析发现，现

有的图卷积神经网络在建模标签或行人属性关系时，存在以下两点问题：

(1) 在构建的标签或行人属性共现依赖关系图中，存在大量的有向边，其方向是从出现频率较低的标签或行人属性指向出现频率较高的标签或行人属性。这种模式不利于低频率的标签或行人属性从高频率的标签或行人属性中迁移特征信息，导致低频率标签或行人属性的识别性能不佳。

(2) 无论是在行人属性识别数据集上还是在多标签图像分类数据集上构造的图中，大量的节点都不具有从其它节点指向它自己的边，这会导致大量的节点不参与图卷积网络的消息传递过程，进而使图卷积神经网络退化成一般的前向传播神经网络，限制性能的提升。

针对以上两个问题，本课题提出了一种新的构造共现依赖关系图的方法。该方法依据反应共现依赖关系的条件概率，构造一对“非对称”的图，在这一对图上，使用图卷积神经网络建模标签或行人属性的共现依赖关系，实验结果证明，本课题中提出的方法可以有效的提升行人属性识别任务的性能。

1.3.2 空间位置信息提取研究

现如今，注意力机制仍然广泛的应用于计算机视觉以及自然语言处理领域。在行人属性识别任务中，行人属性在行人图片上的空间位置分布具有一定的规律性，具体来说，识别特定的行人属性时应关注于图像中特定的空间位置。所以，本课题进一步的探索了搜索与特定行人属性相关的图片空间位置的方法。

本课题提出了一种基于词向量语义指导的空间位置注意力机制的行人属性识别方法。这种注意力机制不同于时下最流行的自注意力机制，在这种注意力机制中，与行人属性相关的词向量将会作为注意力机制中的 Query，同时，图像的特征图将会作为注意力机制中的 Key 与 Value，这种方式相当于在图像的特征图中搜索与特定行人属性的词向量信息高度相关的空间位置，识别特定的行人属性时，可以更加有针对性的关注图像特征图中的部分区域。在本课题中，通过这种方式可以引入与行人属性的空间位置分布以及行人属性之间的相对位置关系相关的先验知识，进一步的提升识别性能。

1.4 论文组织结构

本文的组织结构如下：

第二章介绍本论文的相关技术背景，其中包含卷积神经网络，注意力机制，图卷积神经网络等技术的相关知识，以及性能评估指标与研发平台等。

第三章介绍本论文提出的基于非对称共现依赖关系图的共现关系建模方法以及它在多标签图像分类任务以及行人属性识别任务中的应用效果。

第四章介绍本论文提出的基于词向量语义指导的空间注意力机制模型的行人属性识别方法以及它在行人属性识别任务中的应用效果。

第五章对本论文的内容进行总结，阐述了本论文中的两方面工作的主要内容和贡献，并对未来可以进一步研究的内容做了介绍。

2　技术背景

本章将介绍本课题中使用到的深度学习技术的背景。本章首先介绍卷积神经网络，卷积神经网络几乎是所有计算机视觉任务中必不可少的用于提取特征的主干网络，本章将会介绍它的概念以及一个实例——残差网络。接下来，本章将会介绍注意力机制的基本思想，并介绍应用于计算机视觉领域的注意力机制模型。紧随其后，本章将介绍近几年来迅速发展的图卷积神经网络，图卷积网络在建模非欧式的数据时表现出了极为强大的能力，同时，它也是我们建模标签或行人属性之间的共现依赖关系的核心技术。最后，本章将介绍本课题中用于评估模型性能的指标，以及研发模型的平台。

2.1　卷积神经网络

本节首先简述卷积神经网络的发展历史与种类，并简述其原理。接下来，本节将会介绍一个具体的卷积神经网络实例——深度残差网络，深度残差网络是目前使用最为广泛的卷积神经网络之一，也是本课题中主干网络所采用的模型。

2.1.1　卷积神经网络基础

卷积神经网络(CNN)是一种广泛应用于计算机视觉领域的人工神经网络。二十世纪五十年代末，神经科学研究领域的研究者们发现，人类大脑的视觉系统在识别物体时类似于一个信息分级处理架构，较低的层级用于识别一些低级的特征，比如纹理，颜色，轮廓等，随着层级的逐渐升高，较高的层级将低级的特征进一步的组合成较高的语义特征并进行识别。受到人类视觉系统的启发，Yann Lecun 等人[19]在二十世纪九十年代首次提出了卷积神经网络，并成功的应用于手写体数字识别任务。但是，由于当时的数据量以及硬件计算能力的限制，卷积神经网络并没有迅速的发展起来，直到 2012 年，Hiton 等人[5]提出了 AlexNet，卷积神经网络才逐渐的成为学术界的研究热点。

卷积神经网络的核心是卷积层，除了一般的卷积层之外，还有很多中其他类型的卷积层，比如在 Xception 网络[20]中提出的深度可分离卷积，它广泛的应用于 MobileNet 系列的网络中[21-23]，再比如 AlexNet[5]中使用的分组卷积，它广泛的应用于 ShuffleNet 系列的网络中[24, 25]，此外，还有空洞卷积[26]，转置卷积[27]等。基本的卷积神经网络都是由大量的卷积层组成的，除此之外，卷积神经网络的构成

还包括激活层，池化层，批归一化层[28]，残差链接[6]等。现代的卷积神经网络已经不再需要研究人员们进行手工设计，通常使用神经架构搜索技术[29-32]可以在特定的架构空间内搜索符合特定条件的卷积神经网络架构，不过，这些搜索到的卷积神经网络架构仍然为上文中提到的基本组成元素的组合。

卷积神经网络之所以适合于处理计算机视觉领域中的图像数据，是因为它们能使用局部操作对表征进行分层抽象。这种方式之所以对图像数据有效是源于以下两点原因。首先，对于图像数据来说，局部操作至关重要，因为在图像中，目标对象通常具有局部性，一个目标对象通常只关联于一小部分紧密连接的图像像素区域，所以使用局部操作可以很好的提取目标对象的特征，相比于一般的提取全局特征的前向神经网络，局部操作不仅符合目标对象的局部性，避免全局特征引入不必要的噪声，还能够节约参数。另外，高维语义特征通常是多种低维特征的组合变换，使用多层卷积运算，可以逐步的将低维特征进行组合变换，进而渐渐抽象出高维语义特征，这种特征提取模式被称为分层抽象。

2.1.2 深度残差网络

深度残差网络(ResNet)[6]是计算机视觉领域中使用的最为广泛的图像特征提取网络。在深度残差网络诞生之前，学术界已经达成了一个普遍的共识，即增加卷积神经网络的深度能够使其提取更加复杂的特征模式，从而使性能获得提升。VGG网络[8]就是依靠增加网络深度获得了大幅度的性能提升。然而，深度残差网络的研究者们却发现当网络的深度增加到一定程度时，网络的性能就会发生饱和，如果此时继续增加网络的深度，还有可能会发生退化，导致性能降低，网络的性能可能会退化到还不如一个浅层卷积神经网络，并且，这种退化现象并不是由于过拟合而引起的，因为性能的下降不仅仅表现在测试集上，也表现在训练集上。图 2-1 展示了一个 20 层的卷积神经网络和一个 56 层的卷积神经网络训练期间在训练集上与测试集上的误差变化。

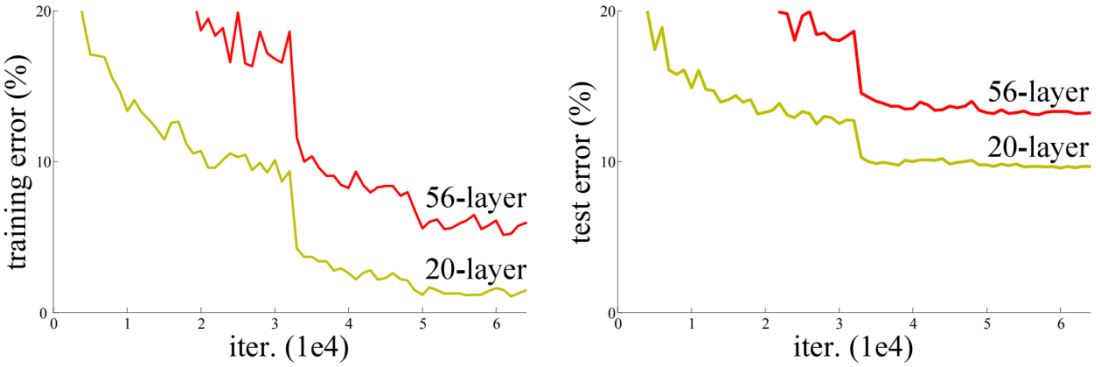

图 2-1 不同深度的卷积神经网络在训练期间的误差变化[6]

Figure 2-1 Error variation of convolutional neural networks with different depths during training[6]

深度残差网络的研究者们将网络退化问题归咎于深层卷积神经网络难以训练，当网络深度增加时，参数空间增大，优化的难度会显著增加，同时，梯度消失，梯度爆炸等问题可能也会伴随着出现，进一步的增加优化的难度。

深度残差网络的研究者们为了解决网络退化的问题，提出了一种方法，这种方法的核心在于改变网络的学习目标，传统的深度卷积神经网络中每一层的目标都是学习到一个映射关系，这个映射关系能够对输入的特征图进行特征提取并重新组合抽象成更复杂的高维语义特征，而在深度残差网络中，每一层不再学习这种映射关系，而是学习残差映射关系，这一层的输出会与输入通过一个“短路连接(shortcut)”组合在一起，形成一个整体的输出，以模拟原始卷积层的学习目标。现在用数学语言来描述这种方法的核心思想：假定某一卷积层在传统的卷积神经网络中想要学习一个关于输入 x 的映射 $H(x)$，那么在深度残差网络中，它的学习目标 $F(x)$就是 $H(x) - x$，该层的输出 $F(x)$将会通过一个短路连接与输入 x 相加，得到 $F(x) + x$，即 $H(x) - x + x = H(x)$。网络结构如图 2-2 所示。

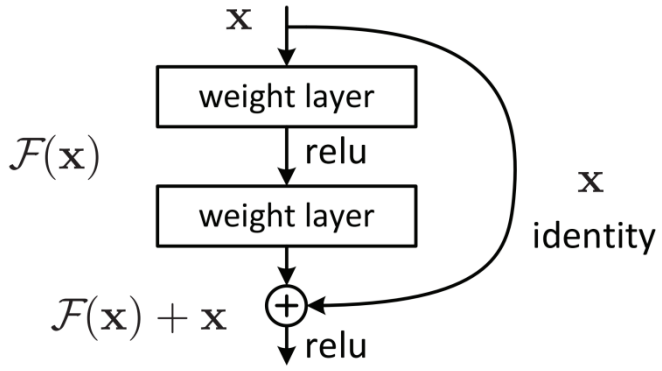

图 2-2 深度残差网络核心结构[6]

Figure 2-2 Core structure of deep residual network[6]

这种方法基于一个事实：假定需要在一个浅层卷积神经网络的基础上，通过堆叠新的卷积层构建一个深层卷积神经网络，一个最极端的情况是，堆叠的新的

卷积层没有学习到任何新的高维抽象的特征模式，而仅仅是将输入复制到输出，相当于学习了一个恒等映射，那么这个深层卷积神经网络的性能至少不会比浅层神经网络的性能要差。然而，即使是简单的恒等映射，对于卷积层来说也是很难通过梯度下降等优化算法精准的学习到的，所以研究者们通过短路连接这种方式显示的构造了恒等映射，再令卷积神经网络去学习残差映射，相当于降低了优化的难度。同时，短路连接还能够直接将深层的梯度回传给浅层，一定程度上缓解了梯度消失的问题。

深度残差网络使构建极深的卷积神经网络成为了可能。在本课题中分别选取了具有 18 层，50 层和 101 层的深度残差网络作为主干网络用来提取图像特征。

2.2 注意力机制

注意力机制的思想最早起源于自然语言处理领域中机器翻译任务[33, 34]。早期的机器翻译任务通常采用编码器-解码器(Encoder-Decoder)架构，编码器和解码器通常由循环神经网络以及它的更复杂的变体[35, 36]所构成。编码器-解码器架构完成机器翻译任务的一般思路是：首先通过编码器，将待翻译的句子编码成一个定长向量，这个定长向量将作为解码器的输入，由解码器解码，输出翻译后的句子。这个框架简单直观，易于理解，但是也存在着很大的局限性。最大的局限性源自于定长向量，定长向量主要存在以下两点问题。首先，无论多长的句子都会被编码为一个定长向量，这本身就有失偏颇，对于长句子来说，如果定长向量所能够容纳的信息不足以概括整个长句子的内容，翻译效果就会大打折扣。另外，许多研究表明，如果采用循环神经网络以及它的变体作为编码器，编码器输出的定长向量所包含的信息中，处于待翻译句子尾部的词汇在信息中所占的比重较大，换句话说，编码器输出的定长向量中更倾向于“记住”句子末尾的词汇，或者说更容易“忘记”句子开头的词汇。

为了解决上述的定长向量引发的两个问题，研究人员们提出了注意力机制。注意力机制的核心思想源自于一个客观事实，就是在翻译句子时，目标句子中不同的词汇与待翻译句子中不同的词汇存在着相互对应的关系，所以在翻译的过程中，每当需要翻译一个词汇时，都应该关注于待翻译句子中的不同词汇。举例说明上述客观事实，假定我们希望将英文句子 ”I love China.” 翻译为中文“我爱中国。”，在翻译第一个中文词汇“我”的时候，显然应该关注于英文句子中的 ”I”，同理，翻译“爱”时，应该关注于英文单词 ”love”，最后，翻译“中国”时，应该关注于英文单词 ”China”。在早期的机器翻译任务中，编码器-解码器架构中的编码器与解码器都是由循环神经网络或其变体构成的，在编码器中，循环神经网

络在每一个时间步输入一个词汇，并输出一个表示当前循环神经网络状态的隐向量，这个隐向量中包含了大量的关于当前时间步输入的词汇的信息，因此隐向量可以作为在解码器的翻译阶段所需要关注的局部信息。在解码器进行翻译的阶段，解码器中的循环神经网络或其变体在每一个时间步都应该关注于编码器输出的隐向量中与当前目标词汇高度相关的隐向量，为了体现“高度相关”，同时也为了让整个模型可导，能够以梯度下降的方式进行训练，解码器在每一个时间步，都会为所有的编码器隐向量动态的分配一组归一化的权重，并对所有编码器隐向量进行加权求和，作为最终要关注的隐向量，权重较大的隐向量即为当前时间步应该重点关注的隐向量，权重采用可导的方式进行计算，整个模型也能够以梯度下降优化算法进行训练。

注意力机制在机器翻译任务中大幅度的提高了性能，因此，它也被扩展到自然语言处理领域的许多其它的模型中。目前在自然语言处理领域中大放异彩的BERT 语言预训练模型[37]是基于 Transformer 架构[38]的，而 Transformer 架构的核心就是注意力机制。与此同时，在计算机视觉领域中，图像数据本身就存在很强的局部相关性，一些对象实体在图像中通常只和一部分紧密连接的像素点相关，所以图像数据本身也极其适合使用注意力机制建模。因此，部分研究人员将注意力机制引入了计算机视觉领域，目前，针对于对于图像数据的注意力机制大致可分为三类，一类是空间注意力机制，一类是通道注意力机制，最后一类是前两类注意力机制的结合。

空间注意力机制指的是在图像或特征图中有选择性的关注某些特定的空间位置。空间注意力机制的提出源自于一个客观事实——在图像中，特定的对象实体通常只和一部分紧密连接的局部像素点相关联，即使是在经过卷积层的运算得到的特征图上，这种对象实体与紧密连接的局部空间位置之间的关联也仍然存在，因此，如果对特定的对象实体感兴趣，就应该关注于特定的局部空间位置，空间注意力机制的典型代表就是 No-Local 网络[39]。通道注意力机制指的是在特征图中有选择性的关注某些特定的通道，忽略某些不重要的信息所对应的通道。通道注意力机制的提出源自于另外一个客观事实——在卷积神经网络中，对图像或特征图进行卷积运算可以看作是对特征的提取与重组。输出特征图的每一个通道，都是由一个特定的卷积核对原始输入进行卷积运算而得到的，这一步可以视作特征的提取，特征的信息包含于特定的卷积核中。将所有输出的单通道特征图沿着通道这个维度重新拼接在一起，组合成新的特征图，这个新的特征图包含了新的更加抽象复杂的特征模式，这一步可以视作特征重组。在不同的任务中，需要关注的特征可能是不同的，而通道注意力机制正是基于这个事实，不同的通道对应于不同的特征的提取结果，由于只关注部分特征，那么也可以只重点关注于这部分

特征所对应的通道，忽略其它的不重要的特征所对应的通道。通道注意力机制的典型代表就是 SENet[40]。许多研究表明，空间注意力机制与通道注意力机制都能够一定程度的提升模型性能，那么很自然的就可以想到将二者结合起来进一步的提升性能，这就是最后一类注意力机制，其典型代表就是 CBAM[41]。

在行人属性识别任务中，不同行人属性通常与行人图片上不同的空间位置相关联，而且，这些行人属性的空间位置之间通常存在着一定的规律性，比如，下装的空间位置几乎全部都位于上装的空间位置的下方，所以，空间注意力机制非常适合行人属性识别任务，本课题提出了一种基于词向量语义指导的空间注意力机制模块的行人属性识别方法，用于提升行人属性识别任务的性能。

2.3 图卷积神经网络

现代常用的图卷积神经网络[42]是标准的谱图卷积网络的一阶近似形式[43]。它最早被提出用来完成图结构的数据上的半监督分类任务。图卷积神经网络的前向传播过程可以总结为节点特征的聚合与变换，节点特征聚合可以使得一个节点的特征在图上沿着边传播到其他的节点中去，每一个节点都可以通过节点特征聚合获取图的局部或全局的拓扑结构信息。这种在图上传播信息的能力是图卷积神经网络能够建模图结构数据的最重要的因素。

图卷积神经网络将图结构的数据作为输入，图结构数据实际上分为两部分，一部分是图上节点的特征，另一部分是图的拓扑结构。图上节点的特征通常是由能够表征图节点所对应的实体的标量或者向量构成的。图的拓扑结构则是由一个被称为“邻接矩阵”的图元素来表达。假设图中有 N 个节点，那么邻接矩阵就是一个具有 N 行 N 列的二值矩阵，每一行或每一列都对应着一个节点。若图是一个无向图（边没有方向），并且节点 i 和节点 j 之间存在一条边，那么矩阵的第 i 行第 j 列与第 j 行第 i 列的值均为 1，若节点 i 和节点 j 之间不存在边，那么上述的两个值则均为 0。从上述定义中不难发现，无向图的邻接矩阵是一个对称矩阵。若图是一个有向图，邻接矩阵的定义会稍有不同，当节点 i 和节点 j 之间存在一条从节点 i 指向节点 j 的边时，邻接矩阵的第 j 行第 i 列的值为 1，否则为 0。首先，与无向图的邻接矩阵不同的是，有向图的邻接矩阵不一定是一个对称矩阵，只有在一种极其特殊的情况下，即如果两个节点之间存在一条从一个节点指向另一个节点的边，也存在一条从另一个节点指向原节点的边时，有向图的邻接矩阵才是对称矩阵，然而，实际数据中，几乎无法寻找到满足这种条件的有向图。另外，可以从整行以及整列的角度进一步的分析有向图的邻接矩阵。首先，观察一个特定的行，若第 i 行中有几个特定的元素为 1，比如第 i 行的第 j 列，第 k 列的元素为 1，这代

表节点 j 和节点 k 存在指向节点 i 的有向边，换句话说，节点 i 所对应的行中，指明了其他节点是否存在指向节点 i 的有向边。进一步，观察一个特定的列，若第 i 列中，有几个特定的元素为 1，比如第 i 列的第 m 行和第 n 行的元素为 1，代表了节点 i 存在指向节点 m 和节点 n 的有向边，换句话说，节点 i 所对应的列中，指明了节点 i 是否存在指向其他节点的有向边。

第三章将详细的探讨有关于有向图的图卷积运算的特性，为了便于在第三章中展开讨论，本章将提前介绍一些相关概念。对于有向图，有两个概念，分别称之为“入度”和“出度”。入度和出度这两个概念是针对于特定节点而言的，入度表示指向该节点的边的数量，而出度表示从该节点指出的边的数量。从有向图的邻接矩阵中可以很容易的获知入度与出度的信息，节点 i 的入度即为邻接矩阵第 i 行中元素 1 的个数，或者说是第 i 行所有元素的和，而节点 i 的出度即为邻接矩阵第 i 列中元素 1 的个数，或者说是第 i 列所有元素的和。对邻接矩阵按行或按列求和就可以得到所有节点的入度与出度信息。

图卷积神经网络的前向传播可以由以下公式（1）概述：

$$Z = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} X \theta \tag{1}$$

在公式（1）中，$A = A + I_N$，其中 I_N 是单位矩阵，$D_{ii} = \sum_j A_{ij}$ 是度矩阵。X 是节点特征矩阵。$\theta \in R^{d_{input} \times d_{output}}$ 是节点特征变换矩阵。d_{input} 和 d_{output} 分别是节点特征的输入维度和输出维度。公式（1）可以简写为以下公式（2）：

$$Z = AX\theta \tag{2}$$

其中，$\hat{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$ 被称之为标准化邻接矩阵。在上文中曾提到过，图卷积神经网络的前向传播过程可以概括为节点特征的聚合和变换。节点特征聚合隐含在矩阵乘法 $\hat{A}X$ 中。令 $M = \hat{A}X$，矩阵 M 的第 i 行就是第 i 个节点进行特征聚合后的结果。这个特征聚合的结果可以看成是第 i 个节点的特征与其邻居节点的特征的加权平均，权重由标准化邻接矩阵的第 i 行给出。

从公式（2）中可以了解到，如果将多个图卷积层堆叠在一起组成一个图卷积神经网络，图中的节点就可以聚合来自更远的邻居节点的特征。假设图卷积神经网络由 k 个图卷积层构成，那么图中的一个节点最多可以聚合它的 k 阶邻居节点的特征。通常，图卷积神经网络会控制图卷积层的层数，以控制图中的节点所能够感知到的到图拓扑结构的范围。有许多研究表明，图卷积神经网络的层数不宜过多，这主要源自于两点原因。首先，通常节点只聚合其二阶或三阶邻居节点的特征就已经可以有效的大幅度提升性能，进一步增加层数，性能的提升效果会变的极其微弱，但进行训练的成本却会更高，训练的难度也会增大，得不偿失。另外，当图卷积层数过多时，图卷积神经网络会产生严重的过度平滑现象[44]，过度平滑现象会导致图上所有的节点的特征趋于一致甚至变得几乎无法区分，最终有可能

导致性能的下降。

2.4 性能评估指标

在第一章介绍的背景中，已经说明了行人属性识别任务与多标签图像分类任务的关系，简而言之，行人属性识别任务就是多标签图像分类任务的一个具体应用，所以，用于评估多标签图像分类任务性能的指标同样也适用于行人属性识别任务。在本节中，将会介绍用于评估多标签图像分类以及行人属性识别任务的性能指标。在第三章以及第四章中，这些性能评估指标将会被用来评估本课题中提出的模型在多标签图像分类基准数据集以及行人属性识别基准数据集上的性能。

多标签图像分类任务或行人属性识别任务本质上可以看成是多个二分类或多分类任务的组合，二分类任务与多分类任务都有统一的性能评估指标，而多标签图像分类任务或行人属性识别任务的性能评估指标与这些基本的二分类任务与多分类任务的性能评估指标类似，相当于这些基本的性能评估指标在多标签上的延伸[45]。

在二分类任务或多分类任务中，Average Precision 是对 Precision-Recall 曲线上的 Precision 值求均值。对于 Precision-Recall 曲线来说，理论上标准的方法是使用积分来进行均值计算。然而在实际应用中，通常无法获取精确的 Precision-Recall 曲线，所以一般并不直接对该 Precision-Recall 曲线进行积分计算，而是简单的对在不同的 recall 值下测量得到的 Precision 值直接求平均。由于在多标签图像分类任务或行人属性识别任务中存在多个标签或行人属性，每一个标签或行人属性的分类或识别都可以独立的作为一个二分类或多分类任务，那么就可以为每一个标签或行人属性独立的计算出一个 Average Precision。在多标签图像分类任务或行人属性识别中，mean Average Precision (mAP)是所有的标签或行人属性的平均 Average Precision，它的定义如下公式（3）所示：

$$mAP = \frac{1}{C}\sum_{i}^{C} AP_i \tag{3}$$

其中，C为标签或行人属性的个数，AP_i是第i个类别或行人属性的Average Precision值。

除了 mAP 性能评估指标之外，还有两组性能评估指标也非常重要，这两组性能指标分别用于评估全局性能和每一类的平均性能。度量全局性能的指标包含 Overall Precision(OP)，Overall Recall(OR)与 Overall F1(OF1)，它们的定义如以下公式（4），（5），（6）所示：

$$OP = \frac{\sum_i N_i^C}{\sum_i N_i^P} \tag{4}$$

$$OR = \frac{\sum_i N_i^C}{\sum_i N_i^g} \tag{5}$$

$$OF1 = \frac{2 \times OP \times OR}{OP + OR} \tag{6}$$

其中，C 与上文 mAP 的定义中的 C 相同，N_i^C 表示真实的具有标签或行人属性 i 并且标签或行人属性 i 被模型正确的分类或识别的图片的数量，N_i^p 表示被模型预测具有标签或行人属性 i 的图片的数量，N_i^g 表示真实的具有标签或行人属性 i 的图片的数量。

评估每一类平均性能的评估指标包含 Per-class Precision (CP)，Per-class Recall (CR)与 Per-class F1 (CF1)，它们的定义如下公式（7），（8），（9）所示：

$$CP = \frac{1}{C}\sum_i \frac{N_i^C}{N_i^P} \tag{7}$$

$$CR = \frac{1}{C}\sum_i \frac{N_i^C}{N_i^g} \tag{8}$$

$$CF1 = \frac{2 \times CP \times CR}{CP + CR} \tag{9}$$

公式（7）、（8）和（9）中的符号与全局性能评估指标中出现的符号具有完全相同的物理意义。

从上述公式中可以发现，用于评估全局性能的评估指标就是将传统的二分类任务或多分类任务中的 Precision，Recall 以及 F1 score 扩展到了多个类别中，相当于把所有的类别当作一个整体的“正类”来看待。而用于评估每一类平均性能的评估指标则是将各个类别独立出来，分别计算每一个类别内部的 Precision，Recall 以及 F1 score，再计算所有类别的平均性能。如果数据集中存在类别分布不平衡的现象，导致部分类别出现频率极高，而其他类别的出现频率较低，此时，用于评估全局性能的评估指标 OP，OR 和 OF1 相比于评估每一类平均性能的评估指标 CP，CR 与 CF1 更容易受到出现频率较高的类别的影响。

2.5 研发平台

本小节将会简略介绍本课题中使用的研发平台，研发平台包括三个基本组成部分，分别是 Python 编程语言，Anaconda，以及 Pytorch。

Python 编程语言是一种解释型编程语言，它是一种支持面向对象，以及动态数据类型的高级程序设计语言。Python 语法结构简单，易于学习，代码也易于阅

读与维护，同时还能支持多种高级特性。Python 优势在于具有丰富功能的库，因此其在多个领域都得到了广泛应用，并且 Python 支持交互式编程，这使代码调试更加方便，对于非计算机专业的使用者来说更加友好。最后，Python 编程语言具有良好的可移植性，其代码可以不经过任何修改的情况下在大多数软硬件平台上运行。基于以上优点，Python 编程语言在数据科学以及人工智能领域几乎成为了最主流的编程语言工具。

Anaconda 是一个开源的 Python 发行版，其主要用于科学计算。Anaconda 中包含了多个功能丰富的包，例如 NumPy，pandas，以及 matplotlib 等。NumPy 主要用于进行多维数组与矩阵的运算，以及一些数学函数的运算。Pandas 是一个数据分析库，它通常以 DataFrame 格式来管理数据，并提供了许多高效的用于操作 DataFrame 接口。Matplotlib 是一个强大的数据可视化工具，它为数据可视化提供了多种不同的形式，是目前使用最为广泛的数据可视化工具之一。除了功能丰富的包之外，Anaconda 还提供了一个强大的包管理工具 conda，conda 支持创建虚拟环境，为高效开发奠定了基础。

Pytorch 是一个开源的基于 Python 的深度学习库，无论是在工业界还是在学术界，都有非常广泛的应用。Pytorch 本身其实是 torch 的 Python 版本，而 torch 是一个对多维数组数据进行计算的库，在深度学习领域有这广泛的应用，基于 Python 编程语言的 Pytorch 重写了 torch 库的许多方面，在追求简洁易用的基础上又能够同时保证效率，进一步提高了 torch 的灵活性和效率。Pytorch 的另外一个重要的特性的是支持动态计算图，相比于另一款主流深度学习框架 Tensorflow 的静态计算图来说，动态计算图易于调试，并且能够为其他的一些需要动态数据进行的分析提供更加直接的获取动态数据的方式，例如梯度分析等。得益于这种特性，Pytorch 目前在工业界与学术界都有着非常广泛的应用，相比于 Tensorflow 来说，Pytorch 的市场地位仍然在不断扩大。

2.6 本章小节

本章介绍了本课题中所使用的深度学习技术的背景。本章首先介绍了三种深度学习技术，分别为卷积神经网络，注意力机制，与图卷积神经网络。这三种深度学习技术是本课题中所提出的模型的基本组成部分，对于后面的章节进行展开讨论至关重要。最后，本章介绍了性能评估指标与研发平台，性能评估指标对于理解模型的性能提升也至关重要。

3 基于非对称共现依赖关系图的行人属性识别方法

在本章中将会介绍本课题中提出的基于非对称共现依赖关系图的行人属性识别方法。首先介绍现有模型以及这些模型存在的问题，并重点分析目前最先进的基于图卷积神经网络的模型中存在的问题。接下来本章将会介绍解决这些问题的基本思想并给出具体模型。最后，本章将会介绍给出的最终模型在多标签图像分类任务以及行人属性识别任务中的性能，并通过介绍消融实验充分的证明本课题所提出的方法的有效性。

3.1 现有模型及其问题

在第一章中介绍了多标签图像分类任务以及行人属性识别任务从传统方法过渡到深度学习方法的发展历史，并表明无论是在多标签图像分类中，还是在其具体的应用——行人属性识别任务中，建模标签或行人属性之间的共现依赖关系对于提升性能都是至关重要的。在第一章中已经说明了，现有的建模标签或行人属性之间的共现依赖关系的方法主要包括循环神经网络，图卷积神经网络等，理解这些方法的原理及局限性对于理解在本章中提出的基于非对称共现依赖关系图的共现关系建模方法至关重要，在本节中，将会详细介绍这些共现关系建模方法及其优缺点。

循环神经网络作为一种建模时空序列数据的方法，在早期的工作中被引入用来建模标签或行人属性之间的共现依赖关系。Wang 等人[15]首次提出了使用循环神经网络序列化的预测标签或行人属性的思想。但是，其缺陷也较为明显，原因是序列化的预测所有标签或行人属性具有一定的局限性，标签或行人属性之间的共现依赖关系并非如此简单，它并不是一种线性链式的依赖关系，而是一种成对的，非对称的，复杂的共现依赖关系，这种共现依赖关系是从整个数据集的统计信息中反映出来的，而并不是只局限在一张图片之内。

Wang 等人[16]则提出将循环神经网络与注意力机制结合起来，建模标签或行人属性之间的共现关系。这个模型的创新点在于提取每一个标签或行人属性在特征图上的空间位置信息并用于特征提取，为后续的研究进一步的拓宽了思路。但是，使用 LSTM 建模标签或行人属性之间的共现依赖关系，并进行序列化预测的方法与 Wang 等人[15]提出方法具有相同的局限性。

Zhu 等人[17]提出了空间正则化网络，他们的研究与 Wang 等人的研究[16]有一定的相似之处，因为他们都提出了一个观点，就是对一个标签或行人属性进行预测

时，需要考虑这个标签或行人属性在图像特征图上的空间位置信息，在特征图上从特定的空间位置提取和这个标签或行人属性强相关的图像特征，来进行预测。不过这篇论文与 Wang 等人的研究[16]相比，又提出了另一个观点，即不同的标签或行人属性所对应的不同的空间位置之间也存在着相对关系，这一点在行人属性识别中尤为突出，例如，上装和下装的相对位置关系，发型和眼镜的相对位置关系等。为了学习不同标签或行人属性的相对空间位置关系，该论文的作者们提出了空间正则化网络。这篇论文的最大创新点在于提出了不同标签或行人属性所处的不同的空间位置之间存在一定的关联，这个观点为行人属性识别工作提供了新的思路。这种空间位置相关性在行人属性识别数据集上尤为明显。不过，这篇论文所提出的模型并没有考虑从宏观角度上利用整个数据集中反应出来的统计信息建模标签或行人属性之间的共现依赖关系。所以这个模型仍有一定的局限性。

Chen 等人[18]提出了一个全新的建模标签或行人属性间共现依赖关系的方法。他们首先提出，可以通过类别或行人属性出现的条件概率来描述类别或行人属性之间的共现依赖关系，进而可以把复杂的共现依赖关系用图(Graph)的形式表示，在此基础上，使用图卷积网络学习标签或行人属性之间的共现依赖关系。具体的方法如下，首先对整个数据集上的标签或行人属性的出现频率进行统计，并利用这些频率估计标签或行人属性出现的条件概率，设定一个条件概率阈值用来去除统计噪声，接下来，依据上一步得到的条件概率构建能反应标签或行人属性共现依赖关系的图。用图卷积网络将图上每一个对应于特定的标签或行人属性的节点映射到一个特征向量中，每一个特征向量就是这个标签或行人属性所对应的分类器向量，它们将会用于最终的图像分类。分类器向量是通过图卷积网络运算得到的，使用图卷积网络能够将图拓扑结构中反映出来共现依赖关系映射到向量的数值中。这篇论文是一篇极具开创性的论文，在近几年图神经网络迅速发展的背景下，它首次将图卷积网络引入多标签分类以及行人属性识别任务中，并取得了较大的性能提升，这篇论文所提出的模型也是本课题中行人属性识别研究工作的基础。在下一节中，将会详细分析这个机遇图卷积神经网络的模型的局限性。

3.2 基于图卷积神经网络的模型的局限性

本课题深入的研究了 Chen 等人所提出的模型 ML-GCN[18]，并发现了其中的一些缺陷，在本节当中，将会详细描述 Chen 等人提出 ML-GCN 模型存在的缺陷。

在 ML-GCN 模型中，研究者们使用条件概率来度量标签或行人属性之间的共现依赖关系，并使用有向图的形式来表达所有的标签或行人属性之间的共现依赖关系。在有向图中，节点代表某一个特定的标签或行人属性，有向边则表达了共

现依赖关系，举例说明，若节点 *A* 与节点 *B* 之间存在一条由 *A* 指向 *B* 的有向边，那么就代表在标签或行人属性 *A* 出现的条件下，标签或行人属性 *B* 出现的可能性很大，换句话说，就是在标签或行人属性 *A* 出现的条件下，标签或行人属性 *B* 出现的概率大于一个预先设定的阈值。图卷积神经网络将会对这个表征共现依赖关系的有向图进行建模，并且为每一个节点所对应的标签或行人属性输出其分类器向量。在第二章第三节中曾经提到过，图卷积神经网络的前向传播过程可以概括为节点特征的聚合与变换，在有向图的节点特征聚合过程中，边的方向代表了节点特征的聚合方向，根据上面的例子可以推断出，若节点 *A* 与节点 *B* 之间存在一条由 *A* 指向 *B* 的有向边，那么节点 *B* 将会聚合节点 *A* 的特征到它自己的特征中去，但是，反过来却相反，节点 *A* 不会聚合节点 *B* 的特征，图 3-1，3-2 和 3-3 展示了有向图的节点特征聚合过程。这种在图上通过节点特征聚合传播信息的能力使图卷积神经网络能够有效的建模图结构的数据，所以边的方向——节点特征聚合的方向，对于建模共现依赖关系图是至关重要的。

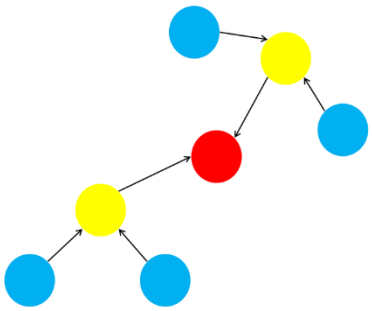

图 3-1　节点特征初始状态

Figure 3-1 Initial state of node features

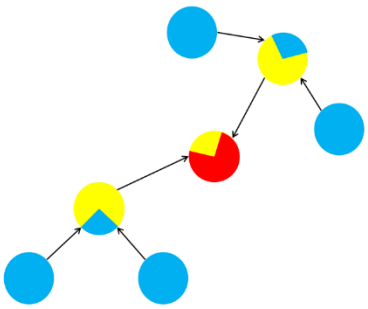

图 3-2　经过一次节点特征聚合的节点特征

Figure 3-2 Node features after a node feature aggregation

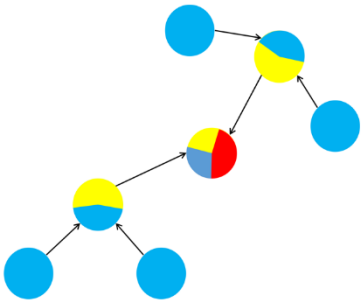

图 3-3 经过两次节点特征聚合的节点特征

Figure 3-3 Node features after twice node feature aggregation

由图 3-1，3-2，3-3 可以观察到，在有向图中，节点特征会沿着边的方向进行聚合。由于蓝色的节点存在指向黄色节点的边，因此黄色节点会聚合蓝色节点的特征信息，但是反过来蓝色节点却不会聚合黄色节点的信息。另外，经过多次节点特征聚合，节点可以聚合到它更远的邻居节点的特征信息，例如在图 3-3 中，红色节点经过两次节点特征聚合可以聚合到它的二阶邻居节点（蓝色节点）的特征信息。

在本课题的前期的调研过程中，通过仔细的观察与分析 ML-GCN 模型在 MS-COCO 数据集以及 PA-100k 数据集上构造的共现依赖关系图，发现 ML-GCN 模型仍然存在两个问题，这两个问题的具体描述如下：

(1) ML-GCN 模型在许多多标签图像分类基准数据集上或行人属性识别基准数据集上所构造的共现依赖关系有向图中，大量的有向边是从低频标签或行人属性对应的节点指向高频标签或行人属性对应的节点。比如，在多标签图像分类基准数据集 MS-COCO 数据集上，共有 80 种不同的标签，其中，” person ”是出现频率最高的标签，” person ”标签对应的节点有 55 条入边，这意味着 ” person ”标签的特征将会由其他 55 个低频标签的特征进行增强，然而，这 55 个低频标签却并没有从高频标签的特征中迁移特征信息，这将会导致模型所能够学习到的有关于低频标签的信息不足，进而影响低频标签的识别性能。

(2) 第二个问题在于，ML-GCN 模型在许多多标签图像分类基准数据集上或行人属性识别基准数据集上所构造的共现依赖关系有向图中，大量的节点并不存在从其它的节点指向自己边。换句话说，这些节点在图卷积神经网络的前向传播过程中，仅仅实施节点特征变换，并不实施节点特征聚合，图 3-1，3-2 和 3-3 中的蓝色节点就是一个很好的示例。对于节点特征来说，

如果仅仅实施节点特征变换而不实施节点特征聚合，相当于图卷积神经网络退化为了一个一般的前向传播神经网络，显然，这种现象的出现会导致大量的标签或行人属性无法利用图卷积神经网络的关系建模能力。在多标签图像分类基准数据集 MS-COCO 上构造的共现依赖关系图上共有 80 个节点，其中高达 57 个节点没有从其他节点指向自己的边，在行人属性识别基准数据集 PA-100k 上构造的共现依赖关系图上共有 26 个节点，其中高达 14 个节点没有从其他节点指向自己的边。图 3-4 展示了在 PA-100k 数据集上构造的有向图。这种现象非常普遍，通常在一个数据集中，只有少数标签或行人属性能够利用到图卷积神经网络的关系建模能力。

以上提到的两个问题之间存在一定联系，在对数据集的统计结果以及构造的共现依赖关系图中发现，不具有从其他节点指向自己的边的节点所对应的标签或行人属性大多数为低频标签或行人属性，因此，寻求一个能够同时解决这两个问题的方法是有可能的。

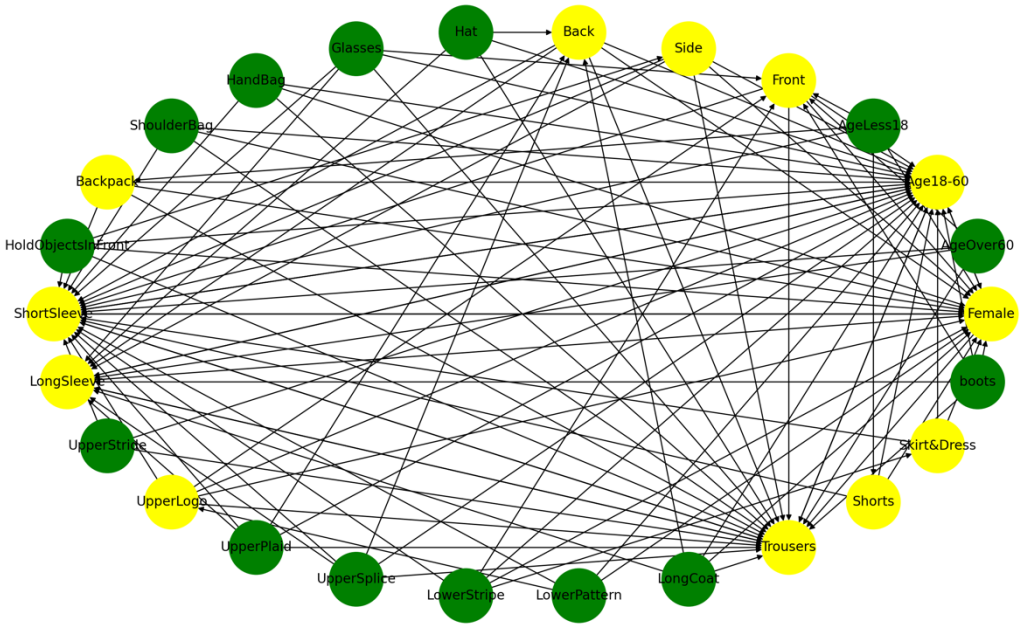

图 3-4 在 PA-100k 数据集上构造的共现依赖关系图，超过半数的节点（绿色节点）不存在从其他节点指向自己的边。

Figure 3-4 In the co-occurrence dependency graph constructed on PA-100k dataset, more than half of nodes (green nodes) do not have edges pointing to themselves from other nodes.

3.3 基本思想

上一节中介绍了现有的基于图卷积神经网络的共现依赖关系建模方法的局限

性，在本节中，将会介绍本课题中解决上述问题的基本思路。

对于上一节中提到的第一个问题，即多数有向边是从低频标签或行人属性指向高频标签或行人属性，从直觉上来说，一个最直接的方法就是反转边的方向，这样许多聚合方向就会变为从高频标签或行人属性指向低频标签或行人属性。如果从共现依赖关系图的构造过程这个角度来说明这种最直接的方法，就是当条件概率 $P(B \mid A)$大于一个特定的阈值时，将会在共现依赖关系图中定义一个从 B 指向 A 的有向边，而不是像 ML-GCN 模型中那样定义一个从 A 指向 B 的有向边，A 和 B 即为标签或行人属性。为了探究这种边的定义方式对于性能的影响，本课题进行了一系列消融实验，在多个基准数据集上的实验结果充分的说明了通过反转边的方向可以提升性能。通过对比高频标签或行人属性与低频标签或行人属性的性能提升可以发现低频标签或行人属性的性能提升更大。因此，可以推测高频标签或行人属性对低频标签或行人属性做了特征信息的补充，致使低频标签或行人属性的识别性能提升，从而提升了总体性能。上一节中已经说明，第一个问题与第二个问题存在一定的关联性，所以第二个问题也可以通过这种方式得到一定程度的缓解。当有向边的方向被转置了之后，大多数节点都存在从其它节点指向自己的有向边，或者说是入度大于 0。比如，在多标签图像分类基准数据集 MS-COCO 上，反转边的方向后，仅仅有 15 个节点不存在由其它的节点指向自己的有向边，而在原始的 ML-GCN 模型中，这样的节点的数量高达 57 个，超过了节点总数 80 的一半，而在行人属性识别基准数据集 PA-100k 上，转置有向边的方向后，所有的节点都存在从其它节点指向自己的有向边。

虽然转置边的方向可以使整体性能获得一定的提升，但是通过对各个标签或行人属性的分类或识别性能的提升程度进行观察可以发现，虽然多数的标签或行人属性的分类或识别性能得到了提升，但是有个别的标签或行人属性的分类或识别性能却出现了下降，这个现象说明，并不是所有的标签或行人属性都适合这种边的方向的定义方式。因此，本课题提出，在这种边的方向的定义方式的基础之上，与原有的 ML-GCN 模型中边的方向的定义方式相结合，以满足个别标签或行人属性对于边的方向的定义方式的需求，相结合的方式还可以进一步缓解上一节中提到的第二个问题，从而进一步提升性能。为了增加模型的灵活性，本课题中提出对于两种边的方向的定义方式使用两个不同的条件概率阈值，以构造一对非对称的共现依赖关系图，并分别使用两个不同的图卷积神经网络建模标签或行人属性的共现依赖关系，最后将两个图卷积神经网络计算得到的节点特征向量进行融合，作为最终的标签或行人属性的分类器向量。在多标签图像分类基准数据集以及行人属性识别基准数据集上的实验结果证明，这种组合的方式可以获得更高的性能。

3.4　模型结构

本节介绍基于上文所描述的思想所构造的模型。本节首先从整体角度概述提出的模型，然后重点描述非对称共现依赖关系图的构造方法，最后介绍模型的各部分细节。

3.4.1　模型结构概述

上一节中介绍了解决现有方法中存在的问题的基本思想，本节将会介绍基于这种思想的具体化模型。

模型结构图如图 3-5 所示。从整体上来说，模型划分为两个部分，它们分别是全卷积神经网络部分和非对称共现依赖关系图建模部分。全卷积神经网络用于提取图像特征图 $F \in R^{H \times W \times C}$，接下来图像特征图通过全局最大池化(GMP)或者全局平均池化(GAP)压缩为一个特征向量 $f \in R^{C}$。H，W 和 C 分别代表图像特征图的高，宽与通道数。在非对称共现依赖关系图建模部分，首先通过上文所述的两种边的方向的定义方式构造的一对非对称的共现依赖关系图，接下来使用两个不同的图卷积神经网络分别对这一对非对称共现依赖关系图中的两个图进行节点特征提取，并使用线性变换将两个图卷积神经网络输出的节点特征进行融合，形成最终的分类器 $W = \{w_i\}_{i=1}^{N}$，N 为标签或行人属性的个数。最终，图像特征向量将会与每一个标签或行人属性的分类器向量 w_i 做内积得到一个标量，并通过 Sigmoid 激活函数归一化，这个归一化的值即为标签或行人属性 i 出现的概率。

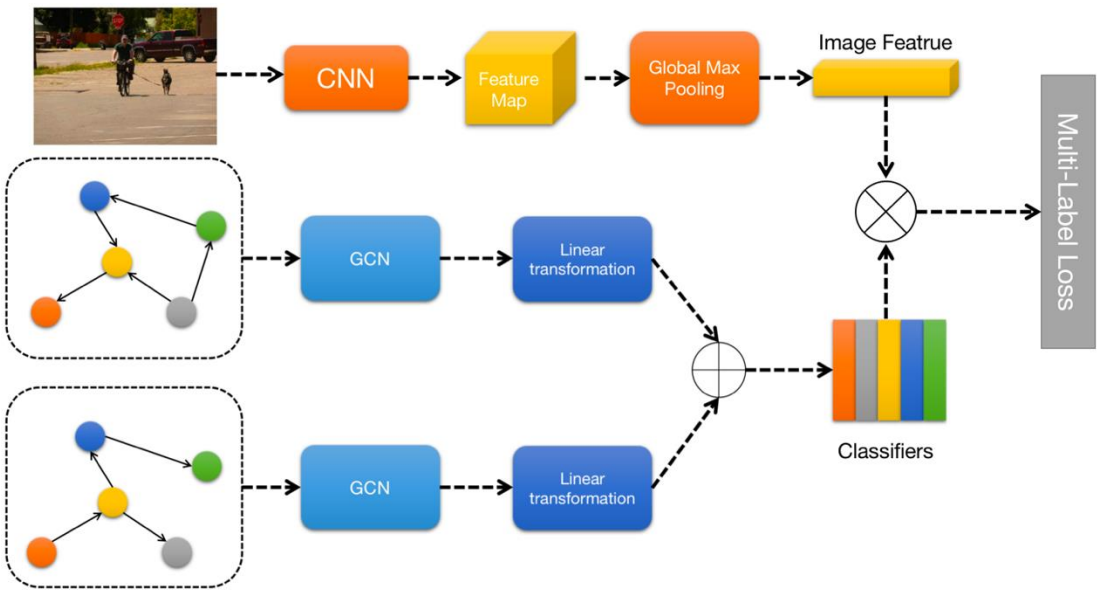

图 3-5 模型结构图

Figure 3-5 Overview of model

3.4.2　非对称共现依赖关系图构造方法

在多标签图像分类任务以及行人属性识别任务中，标签或行人属性并不是独立出现的，它们之间具有一定的依赖关系，某些标签或行人属性的出现可能会依赖于其它的标签或行人属性。度量这种共现依赖关系的方式之一就是使用条件概率。比如，当数据集中存在标签或行人属性 L_i 与 L_j，条件概率 $P(L_i \mid L_j)$代表在标签或行人属性 L_j 出现的条件下，L_i 出现的可能性。由于在数据集当中，训练集的标签是已知的，所以可以使用训练集来估计数据集上所有标签或行人属性出现的条件概率。具体的估计方法如下：首先统计训练集上标签或行人属性 L_i 与 L_j 的出现次数 M_i 与 M_j，与此同时，L_i 与 L_j 出现在同一张图片内的次数 M_{ij} 也可以被统计出来。接下来，可以使用如下（10）和（11）计算公式估计标签或行人属性 L_i 与 L_j 之间条件概率：

$$P(L_i|L_j) = M_{ij}/M_j \quad (10)$$

$$P(L_j|L_i) = M_{ij}/M_i \quad (11)$$

两个非对称的共现依赖关系图可以由统计得到的条件概率来构造。首先，这两个非对称的共现依赖关系图都是有向图，并具有完全相同的节点，每一个节点都对应于一个特定的标签或行人属性。这两个有向图在边的方向的定义上恰好完全相反。从直觉上来说，一个最直接的方法就是当条件概率不为 0，即 $P(L_i \mid L_j) > 0$ 时，在第一个有向图上定义一条从 L_j 指向 L_i 的有向边，而在第二个有向图上定义一条从 L_i 指向 L_j 的有向边。然而，这种直观的方法却存在一些非常明显的问题。首先，如果 $P(L_i \mid L_j) > 0$，可以推测 $M_{ij} > 0$，那么进一步可以推出 $\mathrm{P}(L_j \mid L_i) > 0$。换句话说，对于任意一个图，如果两个节点之间存在一条有向边，无论它的方向如何，一定都存在另一条有向边，其方向与原来的有向边的方向相反。这将会消除有向图的方向性，并导致构造出来的两个图完全相同，这就失去了使用两种不同的边的方向的定义方式的意义。除此之外，由于条件概率是由统计方法估计出来的，其结果当中必然存在噪声，一些极小但非零的条件概率有极有可能是统计噪声。处于不同区间的条件概率数量的统计结果在图 3-6，3-7 中。由图 3-6，3-7 可以观察到，绝大多数的概率值都非常小，这些概率是由噪声引起，并不能够反应标签或行人属性之间的共现依赖关系。因此，为了解决这个问题，本课题中采用了类似 ML-GCN 模型中的方法，一个条件概率阈值 τ 被引入用于对条件概率进行过滤。引入条件概率阈值之后的方法是当条件概率 $P(L_i \mid L_j) \geq \tau$ 时，在第一个有向图上定义一条从 L_j 指向 L_i 的有向边，而在第二个有向图上定义一条从 L_i 指向 L_j 的有向边。条件概率具有非对称性，通过设定合理的阈值，在一对标签上统计得到的两个条件概率中通常只有一个满足条件或者都不满足条件，这有效的维持了

有向图的方向性，同时还可以过滤掉极小但不为零的噪声概率值。通过这种方式所构造出的两个共现依赖关系图中，所有边的方向恰好相反，换句话说，它们的邻接矩阵恰好关于对角线对称，也就是说，节点特征聚合的方向在这两个共现依赖关系图中是完全对称的。

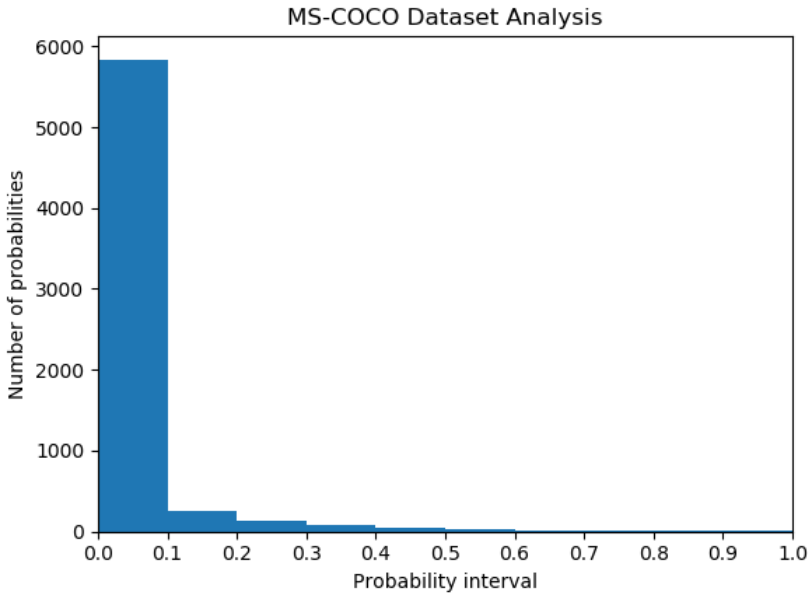

图 3-6 MS-COCO 数据集条件概率分布情况

Figure 3-6 Conditional probability distribution of MS-COCO dataset

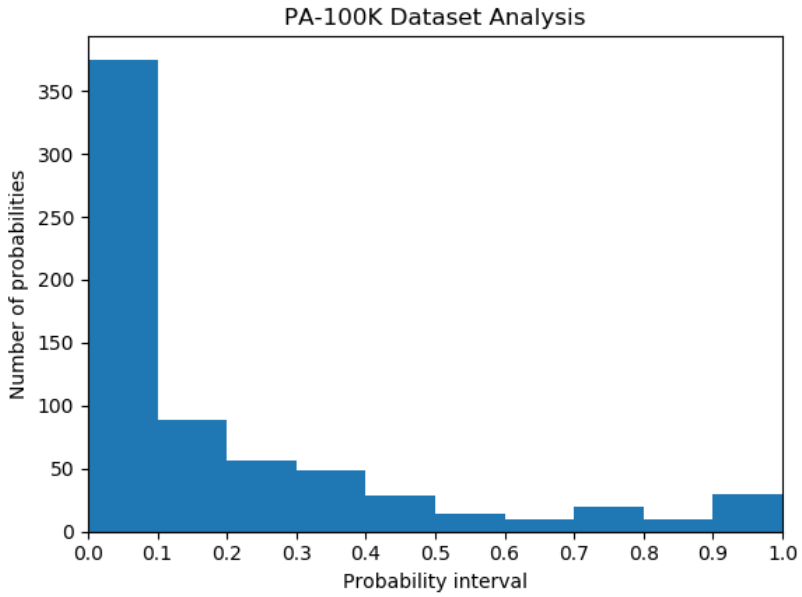

图 3-7 PA-100k 数据集条件概率分布情况

Figure 3-7 Conditional probability distribution of PA-100k dataset

为了增加模型的灵活性，本课题进一步扩展了上述的方案，引入了两个阈值参数 τ_1 和 τ_2 分别用于构造两个共现依赖关系图。具体方案如下：当 $P(L_i \mid L_j) \geq \tau_1$ 时，在第一个共现依赖关系图上定义一条从 L_j 指向 L_i 的有向边，当 $P(L_i \mid L_j) \geq \tau_2$ 时，在第二个共现依赖关系图上定义一条从 L_i 指向 L_j 的有向边。通过引入不同的阈值，可以使构造的两个共现依赖关系图的形式更加多样化。当两个阈值的值不相等时，构造的两个共现依赖关系图的边的方向不再完全对称，比如，当 $\tau_1<\tau_2$ 时，在第一个共现依赖关系上如果存在一条从 L_j 指向 L_i 的有向边，并不代表在第二个共现依赖关系图上一定会存在一条从 L_i 指向 L_j 的有向边，因为条件概率 $P(L_i \mid L_j)$ 与两个阈值参数 τ_1 和 τ_2 的关系有可能满足 $\tau_1< P(L_i \mid L_j) <\tau_2$，此时第二个共现依赖关系图上不存在从 L_i 指向 L_j 的有向边。通过这种方式所构造出的两个共现依赖关系图中，所有边的方向不再完全对称，换句话说，它们的邻接矩阵不再关于对角线对称，也就是说，节点特征聚合的方向在这两个共现依赖关系图中是非对称的，因此，本课题将构造的这一对有向图称为非对称共现依赖关系图。

3.4.3 其他模型细节

上一小节介绍了基于非对称共现依赖关系图的行人属性识别模型的核心部分，即非对称共现依赖关系图的构建。在本节中，将会介绍除了核心部分之外其它的模型细节，其它模型细节包括两个方面，一个方面是图卷积神经网络与特征融合，另一方面是深度卷积神经网络。

图卷积神经网络用于提取上一小节中提出的非对称共现依赖关系图中的节点特征，对于每一个共现依赖关系图，一个两层的图卷积神经网络被用来提取它的节点特征，每一个图卷积神经网络的输入节点特征均为它所对应的标签或行人属性的 300 维的 GloVe 词向量[46]，若标签或行人属性包含多个词，则输入的节点特征为标签或行人属性中所有词的词向量的平均。词向量中包含了标签或行人属性的部分特征信息，并且能够一定程度上反应标签或行人属性之间的语义关系。在两层图卷积神经网络之间，采用负斜率为 0.2 的 LeakyReLU 函数[47]作为非线性激活函数。仅仅使用两层图卷积神经网络的原因是为了避免过深的图卷积神经网络导致的过度平滑问题。最后，两个图卷积神经网络输出的节点特征将会被融合在一起并作为最终的分类器向量。模型中分别使用两个线性变换将两个图卷积神经网络输出的节点特征映射到分类器参数空间：

$$W = Z_1 W_1 + Z_2 W_2 \tag{12}$$

Z_1 和 Z_2 分别是两个图卷积网神经网路输出的节点特征，W_1 和 W_2 是两个线性变换矩阵。对于每一个标签或行人属性，其分类器向量是一个 C 维向量，C 也是

图像特征向量的维数。

得益于强大的局部空间特征提取能力以及对特征的组合与抽象能力，深度卷积神经网络在计算机视觉领域的许多任务中都发挥了重要作用。在基于非对称共现依赖关系图的模型中，深度卷积神经网络作为主干网络用于提取图像的特征。在以往的多标签图像分类以及行人属性识别模型中，ResNet-101 是使用的最为频繁的主干网络。为了进行公平的对比，在多标签图像分类基准数据集 MS-COCO 上，本课题在实验中也采取 ResNet-101 作为主干网络。在实验中发现，ResNet-101 对 MS-COCO 数据集表现出了一定的过拟合现象，因此，为了尽可能的减少主干网络过拟合对性能的影响，并证明基于共现依赖关系图的模型在不同主干网络下的效用，在 MS-COCO 的实验中不仅采用 ResNet-101 作为主干网络，也采用了 ResNet-50 作为 Baseline 模型与基于非对称共现依赖关系图的模型的主干网络。在行人属性识别基准数据集 PA-100k 上，ResNet-50 将会作为主干网络用于进行特征提取。无论是 ResNet-50 还是 ResNet-101，它们都在 ImageNet 数据集上进行了预训练。

3.5 实验验证

本节将会介绍基于非对称共现依赖关系图的模型在多标签图像分类任务以及行人属性识别任务中的性能结果。

3.5.1 实验设置

ML-GCN 模型是本实验中最主要的 Baseline 模型，为了与 ML-GCN 模型进行公平的对比，本实验中使用与 ML-GCN 相同的图卷积神经网络超参数设置，两层图卷积神经网络输出的节点特征维度分别为 1024 与 2048。由于作为主干网络的 ResNet-50 与 ResNet-101 输出的图像特征均为 2048 维，所以节点特征融合过程中的变换矩阵的输出维度也设置为 2048。在非对称共现依赖关系图的构造过程中，由于其中第一个共现依赖关系图的构造方法与 ML-GCN 模型中构造方法一致，所以对于这个共现依赖关系图，在实验中使用了与 ML-GCN 中相同的条件概率阈值，这个阈值 τ_1 设置为 0.4。在实验中发现，若将 τ_2 设置为略大于或等于 τ_1 的值均可以获得较好性能，所以在下文中，将报告 τ_2 设置为 0.6 时所获得的实验性能结果，这是因为当 τ_2 设置为 0.6 时所获得的性能更好。在训练阶段，使用与 ML-GCN 模型完全相同的数据预处理方法，图像将会被随机剪裁一次，并重置分辨率大小，重置后的分辨率为 448×448，最后再经过一次随机水平翻转，作为主干卷积神经网

络的输入。在训练阶段使用的神经网络优化算法为小批量随机梯度下降算法(Mini Batch SGD)，其动量参数设置为 0.9，权重衰减参数设置为 10^{-4}。在训练过程中使用余弦退火学习率衰减策略[48]作为学习率衰减策略，初始学习率设置为 0.01，最小学习率设置为 10^{-6}。模型将会训练 100 个 epoch。在没有其它任何额外说明的时候，基于非对称共现依赖关系图建模的模型与 Baseline 模型均使用上述实验设置进行实验。

3.5.1 实验结果

多标签图像分类基准数据集 MS-COCO 最早被提出用来测试目标检测模型的性能，在近些年来，它也被用于评估多标签图像分类模型的性能。MS-COCO 数据集中，训练集包含 82081 张图片，验证集包含 40504 张图片，在数据集中共有 80 种标签。由于测试集的标签数据并未被公开，实验中将验证集作为测试集用于评估模型性能。表 3-1 中展示了多种不同的多标签图像分类模型在 MS-COCO 数据集上的性能结果。除了 ML-GCN 模型外，表格中还报告了其它一些模型的性能结果，这些模型包括 CNN-RNN[15]，RNN-Attention[16]，Order-Free RNN[49]，ML-ZSL[50]，SRN[17]，Multi-Evidence[51]等。由于一些早期的多标签图像分类模型本身具有一定的限制，它们不可以为每一张图片动态的调整其预测出的标签的数量，所以一般这一类模型均使用 Top-3 性能指标，Top-3 性能指标指的是对于任意一张图片，将出现概率最高的三个标签作为模型为当前图片预测的标签，而其它的标签，即使概率再高，也不会作为模型为当前图片预测的标签。表 3-1 右半部分展示了所有模型的 Top-3 性能结果。在表 3-1 中，“-”代表该模型并未报告这个性能指标。从表 3-1 中可以观察到，无论主干网络是 ResNet-50 还是 ResNet-101，基于非对称共现依赖关系图的模型在所有的重要性能指标上(mAP，CF1，OF1)均优于其它的 Baseline 模型。

行人属性识别基准数据集 PA-100k 是由 598 个室外监控摄像头拍摄的照片所构成的。这个数据集包含了 100000 张行人图片，整个数据集被随机的划分为训练集，验证集与测试集，它们的数量之间的比例为 8:1:1，每一张行人图片都被标柱了 26 个行人属性。对于 PA-100k 数据集，实验中采用 ResNet-50 作为提取图片特征向量的主干网络。表 3-2 展示了基于非对称共现依赖关系图的模型与 ML-GCN 模型的性能结果。从表 3-2 中可以观察到，基于共现依赖关系图的模型在各个性能指标上都显著的优于 ML-GCN 模型。性能的大幅度提升符合预期，原因是 PA-100k 数据集上行人属性之间的共现依赖关系要远强于 MS-COCO 数据集上标签之间的共现依赖关系。多标签图像分类基准数据集 MS-COCO 中的图片来自于范围很大

的自然场景或人造场景，标签的种类繁多，横跨很多不同的领域，标签之间的相互联系本身就不是特别紧密。而行人属性识别数据集 PA-100k 中的图片均为单个行人的图片，行人属性之间的联系非常紧密，共现依赖关系更强，而基于非对称共现依赖关系图的模型能够非常全面而又准确的建模这种共现依赖关系并利用这种共现依赖关系辅助行人属性的识别。从标签或行人属性出现的条件概率中也可以说明 PA-100k 数据集的行人属性之间的共现依赖关系要强于 MS-COCO 数据集的标签之间的共现依赖关系，在 MS-COCO 数据集上，所有值大于等于 0.4 的条件概率的平均值仅为 0.58，而在 PA-100k 数据集上，所有值大于等于 0.4 的条件概率的平均值高达 0.70。表 3-2 中的结果充分的证明了基于非对称共现依赖关系图的模型能够充分利用行人属性之间的共现依赖关系。

表 3-1 MS-COCO 数据集性能评估结果

Table 3-1 Performance evaluation results of MS-COCO dataset

Methods	**All**							**Top-3**					
	mAP	***CP***	***CR***	***CF1***	***OP***	***OR***	***OF1***	***CP***	***CR***	***CF1***	***OP***	***OR***	***OF1***
CNN-RNN	61.2	-	-	-	-	-	-	66.0	55.6	60.4	69.2	66.4	67.8
RNN-Attention	-	-	-	-	-	-	-	79.1	58.7	67.4	80.4	63.0	72.0
Order-Free RNN	-	-	-	-	-	-	-	71.6	54.8	62.1	74.2	62.2	67.7
ML-ZSL	-	-	-	-	-	-	-	74.1	64.5	69.0	-	-	-
SRN	77.1	81.6	65.4	71.2	82.7	68.9	75.8	85.2	58.8	67.4	87.4	62.5	72.9
ResNet-101	77.3	80.2	66.7	72.8	83.9	70.8	76.8	84.1	59.4	69.7	89.1	62.8	73.6
Multi-Evidence	-	80.4	70.2	74.9	85.2	72.5	78.4	84.5	62.2	70.6	89.1	64.3	74.7
ML-GCN (ResNet50)	*81.0*	***84.0***	*69.5*	*76.0*	***85.8***	*72.9*	*78.8*	*87.1*	*62.2*	*72.6*	***90.3***	*64.7*	*75.4*
Ours (ResNet50)	***81.4***	*83.6*	***70.5***	***76.5***	*84.0*	***74.5***	***79.0***	***88.4***	***64.3***	***73.2***	*89.5*	***65.6***	***75.7***
ML-GCN(ResNet101)	82.3	83.3	**71.9**	77.2	84.7	**75.2**	79.7	87.3	**63.8**	73.7	90.1	**65.9**	76.1
Ours (ResNet101)	**82.9**	**86.8**	70.4	**77.7**	**87.6**	74.0	**80.2**	**90.1**	63.2	**74.3**	**91.3**	65.8	**76.5**

表 3-2 PA-100k 数据集性能评估结果

Table 3-2 Performance evaluation results of MS-COCO dataset

Methods	**All**						
	mAP	***CP***	***CR***	***CF1***	***OP***	***OR***	***OF1***
ML-GCN (ResNet50)	63.7	69.7	52.6	60.0	85.1	79.8	82.3
Ours (ResNet50)	**67.0**	**71.5**	**55.2**	**62.3**	**87.9**	**82.3**	**84.4**

3.5.2 消融实验

本小节介绍消融实验的结果与分析。设置消融实验的目的是为了探索两种不同的边的方向的定义方式对性能的影响。本课题基于 Baseline 模型 ML-GCN 设计

了一组对照实验，Baseline 模型 ML-GCN 使用了 3.4.2 小节中介绍的第一种边的方向的定义方式构造一个共现依赖关系图，为了排除阈值等其它因素的影响，本课题在 ML-GCN 模型的基础之上，仅仅将其构造的共现依赖关系图中所有的边的方向进行反转，得到了一个用于对比的模型，显然这个将边的方向反转之后的模型使用的是第二种边的方向的定义方式。这两个模型在 MS-COCO 数据集以及 PA-100k 数据集上的性能在表 3-3，表 3-4 和表 3-5 中。在 MS-COCO 数据集上，使用第二种边的方向的定义方式的性能要略微好于使用第一种边的方向的定义方式。在 PA-100k 数据集上，使用第二种边的方向的定义方式的性能显著的优于使用第一种边的方向的定义方式。然而，无论使用哪一种边的定义方式，模型的性能均无法达到同时使用两种方式的基于非对称共现依赖关系图的模型的性能。

表 3-3 MS-COCO 数据集消融实验结果(ResNet-50)

Table 3-3 Ablation results of MS-COCO dataset (ResNet-50)

Definition	**All**						
	mAP	***CP***	***CR***	***CF1***	***OP***	***OR***	***OF1***
First Definition	81.0	**84.0**	69.5	76.0	**85.8**	72.9	78.78
Second Definition	**81.3**	81.4	**72.0**	**76.4**	83.2	**75.0**	**78.81**

表 3-4 MS-COCO 数据集消融实验结果(ResNet-101)

Table 3-4 Ablation results of MS-COCO dataset (ResNet-101)

Definition	**All**						
	mAP	***CP***	***CR***	***CF1***	***OP***	***OR***	***OF1***
First Definition	82.3	**83.3**	71.9	77.2	**84.7**	75.2	79.7
Second Definition	**82.6**	80.7	**74.2**	**77.3**	82.8	**76.0**	**79.3**

表 3-5 PA-100k 数据集消融实验结果

Table 3-5 Ablation results of PA-100k dataset

Definition	**All**						
	mAP	***CP***	***CR***	***CF1***	***OP***	***OR***	***OF1***
First Definition	63.7	69.7	52.6	60.0	85.1	79.8	82.3
Second Definition	**66.3**	**70.4**	**56.0**	**62.4**	**86.4**	**80.3**	**83.2**

对于行人属性识别基准数据集 PA-100k，本课题进一步的分析了使用第二种边的方向的定义方式对高频行人属性与低频行人属性的识别性能的提升。各个行人属性的出现频率以及识别性能（使用 Average Precision 度量）的提升幅度在图 3-8，3-9 中。从图 3-8，3-9 中可以观察到，使用第二种边的方向的定义方式可以有效的提升大多数低频行人属性的识别性能。然而，在图 3-9 中也可以观察到，存在个别

的几个行人属性，其识别性能不仅没有提升，反而出现了下降。这个现象说明了使用第二种边的方向的定义方式对于某些行人属性来说是不合理的，换句话说，每一种边的方向的定义方式，都有其适合的标签或行人属性，但是从整体性能的角度来说，使用第二种边的方向的定义方式会更合适一些，所以很自然的就可以想到将两种方式结合起来进一步提升性能。

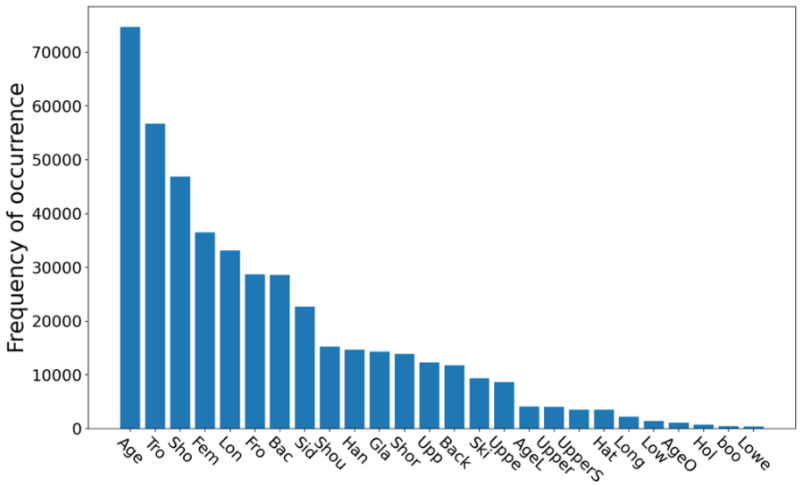

图 3-8 PA-100k 中各个行人属性的出现频率（行人属性名称采用简写）

Figure 3-8 The frequencies of pedestrian attributes in PA-100k dataset (pedestrian attribute names are abbreviated)

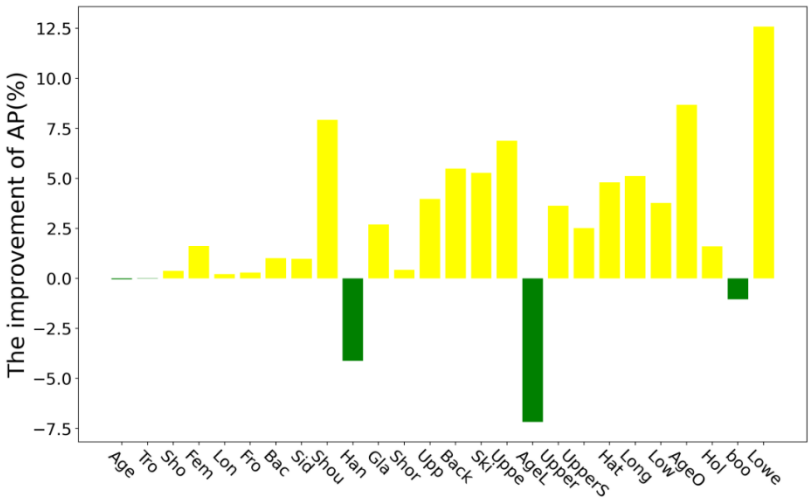

图 3-9 各个行人属性的识别性能提升（行人属性名采用简写）

Figure 3-9 Performance improvement of every pedestrian attribute (pedestrian attribute names are abbreviated)

3.6 本章小结

无论是在多标签图像分类任务中，还是行人属性识别任务中，建模标签或行人属性之间的共现依赖关系都至关重要。本章中提出，使用图卷积神经网络建模共现依赖关系时，使用高频标签或行人属性的特征信息对低频标签或行人属性的特征信息进行补充可以提升低频标签或行人属性的分类或识别性能。基于上述理论，本课题提出使用两种不同的边的方向的定义方式构造一对非对称的共现依赖关系图，接下来使用一个统一的基于图卷积神经网络模型建模共现依赖关系。由于图卷积神经网络具有强大的关系建模能力，模型可以学习到隐含了共现依赖关系信息的分类器向量。这种分类器向量致使深度卷积神经网络提取图像特征时拉近具有共现依赖关系的标签或行人属性在特征空间中的距离，进一步提升了模型的性能。使用两种不同的边的方向的定义方式定义一对非对称共现依赖关系图时，引入了两个不同的阈值用于过滤条件概率，可以增加模型的灵活性，并进一步的提升性能。实验结果表明，通过这种方式可以有效提升多标签图像分类任务以及行人属性识别任务的性能。

4　基于词向量语义指导的空间注意力机制的行人属性识别方法

本章将会介绍本课题中提出的第二个提升行人属性识别任务性能的方法——注意力机制。本章首先详细的描述对行人属性识别数据集的观察，并对一些注意力机制可以发挥作用的场景进行总结。接下来，本章将会具体的描述如何利用基于词向量语义指导的空间注意力机制模块改进常规图像分类模型，并详细描述了基于词向量语义指导的空间注意力机制模块的细节。最后，本章将会介绍实验结果并进行总结。

4.1　基本思想

在本章中，将会介绍提升行人属性识别任务的性能的另一个方法——注意力机制。为什么使用注意力机制可以提升行人属性识别的性能？这源自于对行人属性识别数据集的观察。通常，行人属性识别数据集中的行人图片源自于室内或室外监控摄像头拍摄的照片，这些照片经过人工筛选与剪裁，最终只保留下来与单一行人相关的像素部分，几乎去除了所有的背景。因此，不同行人图片中的行人属性具有类似的分布规律模式，比如，行人上装属性的信息通常来自于图像的中上部分，而行人下装属性的信息通常来自远图像的中下部等。行人图片中这些相似的模式可以简单的概括为：不同的行人属性通常与图像中特定的区域相互关联。基于以上的观察，还可以进一步推断出另外一个比较重要结论，即不同的行人属性的空间位置之间存在相对关系，比如，行人上装属性所对应的空间位置一般在行人下装属性所对应的空间位置的上方。

通过对行人属性识别数据的观察，发现不同的行人属性通常与不同的图像空间位置相关联，并且不同行人属性所对应的不同的空间位置之间还存在相对位置关系。那么应该如何利用这些空间位置信息提升行人属性识别任务的性能呢？本节概括总结了一些特定场景，在这些场景下，空间位置信息将发挥重要作用，这些特定的场景如下：

(1) 多视角场景：拍摄行人图片的摄像头可能以多种不同视角拍摄行人图片，比如，行人有可能是正对摄像头，有可能是背对摄像头，这种视角上的差异可能会导致某些行人属性像素信息的减少或缺失，比如，行人正对摄像头时，摄像头可以全面的拍摄到包括行人面部特征以及上下装信息的图片，但是当行人侧对摄像头时，图片中有关于行人面部以及上装的像素信息会

有所减少，这种像素信息的减少或缺失将导致部分行人属性的识别性能下降。

(2) 遮挡：部分行人图片中存在遮挡现象，遮挡可能是由于其它的行人或物体所导致的。遮挡现象同样会导致与部分行人属性相关的像素信息的减少或缺失，进而影响这部分行人属性的识别性能。

(3) 光照条件：拍摄行人图片时，环境中的光照条件有可能不同，这可能会影响行人属性识别任务的性能。当光照强度极强，比如行人附近有光源的时候，强烈的光线可能会掩盖与行人属性相关的部分像素，影响识别性能。当光照强度较弱，比如在夜晚进行拍摄时，行人属性相关的像素也极有可能被黑暗所掩盖，造成同样的问题。

(4) 低分辨率：由于行人属性识别数据集中的行人图片一般是从视频监控图像中剪裁得到的，所以行人图片的分辨率一般较低，某些在物理空间上占比较小的行人属性，比如发型，眼镜，面部特征等行人属性只能关联极少的像素，导致这些行人属性的特征难以被卷积神经网络提取。

(5) 模糊：模糊是行人图片中一种很常见的现象，由于摄像头在拍摄行人图片时，行人很有可能处于运动状态，所以拍摄的行人图片很可能会出现模糊的现象，部分行人属性所对应的像素发生模糊现象会干扰该行人属性特征的提取。

对于上述的几个场景来说，行人属性的空间位置信息至关重要，当行人图片中因视角，遮挡，光照条件，分辨率，模糊等因素导致部分行人属性所对应的像素信息减少时，利用行人属性的空间位置分布规律以及相对位置关系可以辅助模型准确的定位像素信息减少的行人属性的所关联的空间位置，从而使模型仍然能够准确的提取像素信息减少的行人属性的特征，保证这些行人属性仍然能够被正常的识别。即使未出现部分行人属性所对应的像素信息减少的情况，行人属性的空间位置分布规律以及相对位置关系也能够发挥作用，它可以辅助模型精准的定位特定的行人属性，并提取与该行人属性强相关的特征，从而进一步提升行人属性识别任务的性能。

由上文所述可知，行人属性的空间位置分布规律以及相对位置关系对于行人属性识别任务至关重要，那么应该如何提取行人属性的空间位置信息？在第二章第二节中已经介绍了关于注意力机制的相关知识，其中，在计算机视觉领域中广泛应用的空间注意力机制尤其适合于提取行人属性的空间位置信息，因为空间注意力机制的核心思想是对于不同的目标对象关注图像或特征图上不同的空间位置，而行人属性识别任务中的目标对象——行人属性恰好与不同的空间位置相关联，因此，可以利用注意力机制可以有效的提取行人属性的空间位置信息。

现有的多标签图像分类模型或行人属性识别模型从空间位置信息提取的角度可以分为两类。第一类模型[6, 15, 18]几乎完全忽视了空间位置信息的重要性，这一类模型通常对于图像特征图不做任何额外的处理，而是直接使用全局最大池化(Global Max Pooling, GMP)或全局平均池化(Global Average Pooling, GAP)操作将图像特征图压缩为一个图像特征向量，识别任何行人属性时都使用同样的全局特征向量，而没有针对特定的行人属性从特定空间位置中提取与其高度相关的图像特征。第二类模型[16, 17, 39, 52, 53]通常使用注意力机制提取行人属性的空间位置信息，但是这些模型普遍存在一个问题：行人图片本身的监督信号只包含行人属性，并不包含行人属性所对应的图像像素区域，因此，对于注意力机制模型来说，提取特定的行人属性的特征时，它并不了解哪些空间位置需要关注，只能通过弱监督信号——行人属性本身自行发掘行人属性的空间位置规律。这会导致注意力机制模块难以训练，无法精准的定位行人属性的空间位置。若行人图片受到多视角，遮挡，光照等因素的影响时，性能就会受到更严重的影响。

为了解决上述模型中存在的问题，本课题中提出一种基于词向量语义指导的空间注意力机制模块的行人属性识别方法。对于缺乏监督信号的问题，本课题提出使用词向量作为弥补，指导模型在特征图中搜寻与行人属性相关的空间位置。行人属性所对应的词向量包含了与行人属性相关的信息，引入词向量相当于增加了模型的先验知识，在先验知识的指导下模型能够更加准确的提取与行人属性相关的空间位置信息，缓解了由于缺乏监督信号导致的定位不准确的问题。

4.2　模型总览

本节将会介绍常规图像分类模型以及如何利用注意力机制模块改进常规图像分类模型。理解常规图像分类模型至关重要，因为理解了常规图像分类模型后才能够理解注意力机制的优势，同时，常规图像分类模型也是展开对比实验的基础模型，因此，本节首先介绍常规分类模型及其优缺点，再介绍如何利用注意力机制模块对常规图像分类模型进行改进。

常规图像分类模型如图 4-1 所示。常规图像分类模型一般通过深度卷积神经网络主干对图片进行特征提取，获取一张图片的特征图，并使用全局最大池化或全局平均池化将特征图压缩为一个特征向量，最后使用这个特征向量与每一个分类器向量计算内积并归一化，获取各个类别的概率分布。

在图 4-1 中，H，W 和 C 分别为深度卷积神经网络输出的特征图的高，宽与通道数。全局最大池化或全局平均池化是在每一个通道上实施的，对于全局最大池化来说，输出的特征向量的第 c 维的值为特征图第 c 个通道上 $H\times W$ 个值的最大值，

而对于全局平均池化来说，输出的特征向量的第 c 维的值为特征图第 c 个通道上 $H\times W$ 个值的平均值。

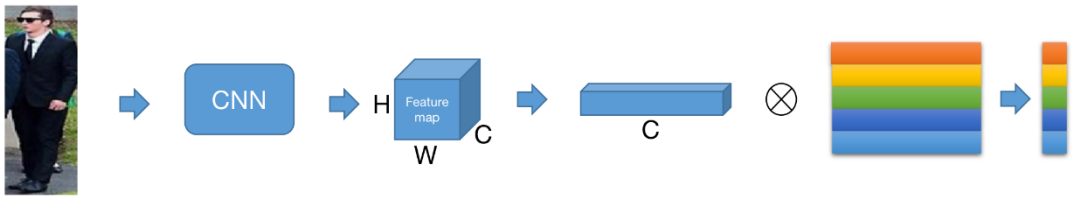

图 4-1 常规图像分类别模型

Figure 4-1 Conventional image classification model

使用常规模型进行行人属性识别时，无论识别哪一个行人属性，都是用完全相同的图像特征向量进行识别，图像的特征向量是对图像特征图进行全局最大池化或全局平均池化得到的，它包含了整个行人图片中所有行人属性的特征信息。然而，在识别某一个特定的行人属性时，图像特征向量中包含的其它行人属性的信息就成为了冗余，甚至还有可能导致当前要识别的行人属性的信息被其它的行人属性的信息所掩盖。因此，对于每一个行人属性，应该从图像的特征图中提取与这个行人属性强相关的那一部分信息，汇总为一个专门用来表征这个行人属性的特征向量，并用这个特征向量进行该行人属性的识别。在本课题中，使用注意力机制模块来完成这个任务。

本课题对图 4-1 所示的一般的模型进行了改进，用一个注意力机制模块替换原有模型中的全局最大池化/全局平均池化操作，注意力机制模块不再输出单个图像的特征向量，而是为每一个行人属性都输出一个专门用于表征这个行人属性的特征向量，改进的模型结构如图 4-2 所示。图 4-2 中的虚线框所包含的部分即为注意力机制模块。

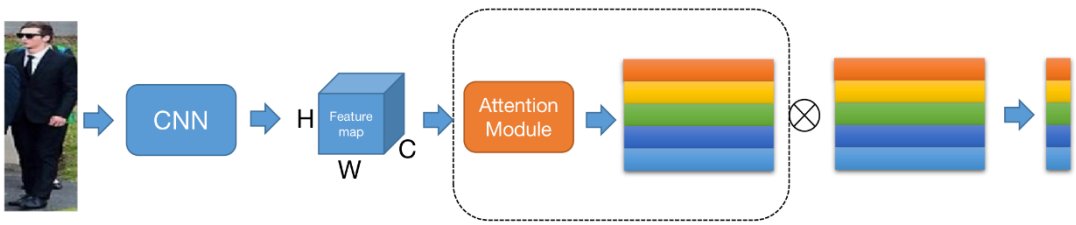

图 4-2 使用注意力机制模块改进的常规图像分类模型

Figure 4-2 Improved conventional image classification model with attention module

4.3 基于词向量语义指导的注意力机制模块

本小节将会详细的描述基于词向量语义指导的注意力机制模块。注意力机制本身并没有一个固定不变的形式，它是一种思想，在不同的领域和任务当中有不一样的具体的表现形式，现如今，研究者们在许多注意力机制研究的基础上，对注意力机制的一般性形式做了概括，这个概括一般被称之为 QKV 模型。

QKV 模型中具有三个要素——查询(Query)，键(Key)，值(Value)。通常来说，Key 与 Value 的数量是相等的，他们具有一一对应的关系，Query 的数量可以与 Key 和 Value 的数量不相等，在某些特殊情况中[37, 38]，Query，Key 与 Value 的数量可以相等。在 QKV 模型中，Query 通常是注意力机制思想中所需要关注的局部信息的线索，注意力机制思想希望根据 Query 这个线索，从 Value 中寻找与这个线索高度相关的 Value 作为注意力机制的输出。那么如何判定一个 Query 是否与某一个 Value 高度相关呢？这就需要 Key，由于 Key 与 Value 是一一对应的，可以将 Key 视为 Value 的概括总结，或者说 Key 代表了 Value。对于每一个 Query，都通过一个函数计算它与所有的 Key 的相关程度，这个函数对于每一个 Key 都输出一个注意力系数，一般情况下会对所有的 Key 所对应的注意力系数进行归一化，归一化系数的大小反映了 Query 与 Key 的相关程度，其实也就反映了 Query 与 Value 的相关程度，最后使用这一组归一化系数对所有的 Value 进行加权平均，作为 QKV 模型对当前 Query 的输出结果，其中与 Query 高度相关的 Value 具有更大的注意力系数，所以在输出结果中的占比也就更大。

假定 Query 的数量为 M 个，即 $Q=\{q_i\}_{i=1}^{M}$，Key 与 Value 的数量为 N 个，即 $K=\{k_j\}_{j=1}^{N}$，$V=\{v_j\}_{j=1}^{N}$，其中 k_j 与 v_j 具有一一对应的关系。那么注意力机制 QKV 模型的一般计算过程如下，首先，对于每一个 q_i，通过一个函数 f 与每一个 k_i 进行运算得到 N 个标量，$Si=\{s_{i1}, s_{i2}, \ldots, s_{ij}, \ldots, s_{iN}\}$，其中 $s_{ij}=f(q_i, k_j)$，接下来，对 N 个标量使用 Softmax 函数进行归一化，得到 $A_i=\{a_{i1}, a_{i2}, \ldots, a_{ij}, \ldots, a_{iN}\}$，满足 $\sum_j a_{ij}=1$。最后，计算 Value 的加权平均值作为注意力机制 QKV 模型的对 q_i 的输出，即 $\sum_j a_{ij}v_j$。

基于词向量语义指导的注意力机制模块的定义也符合上述的注意力机制 QKV 模型。基于词向量语义指导的注意力机制模块的根本目的在于从图像特征图上的局部空间位置提取与特定行人属性强相关的特征，也就是说，注意力模块所关注的局部特征信息是有关于特定行人属性的，那么行人属性即为局部特征信息的线索，因此，应该将行人属性作为 QKV 模型中 Query，由于每一个行人属性通常与不同的局部空间位置相关联，那么每一个行人属性应该对应于一个 Query。Query 作为线索，其中应当包含有关于它所对应的行人属性的信息，因此，在本课题中，采用行人属性的 300 维的 GloVe 词向量作为 Query，词向量给出了行人属性的语义

信息，这种语义信息告知注意力机制模块要去特征图中寻找一个什么样的空间区域。

行人属性的特征信息通常隐含于图像特征图上的局部空间当中，因此，Value 应该定义为图像特征图上局部空间的特征，具体来说，若图像的特征图是一个 $H\times W\times C$ 维的张量，应该以每一个空间位置上的特征作为最基本的局部空间的特征，那么可以将这个 $H\times W\times C$ 维的张量分解为 HW 个 C 维的向量，每一个 C 维向量都对应于一个 Value，这种方式简单直观，可扩展性很极强。接下来阐述 Key 的定义，Key 与 Value 具有一一对应的关系，Key 作为 Value 的代表，应当包含 Value 中对于 Query 来说最感兴趣的特征信息，为了简单起见，在基于词向量语义指导的注意力机制模块中，定义 Key=Value，也就是说 $H\times W\times C$ 维的张量分解成的每一个 C 维的向量即是 Key 又是 Value。最后，需要定义计算非归一化的注意力系数的函数，这个函数的定义方式多种多样，在不同的领域和任务当中可以定义为多种不同的形式，在基于词向量语义指导的注意力机制模块中使用简单的内积运算作为这个函数，即 $f(q_i, k_j)=q_i\cdot k_j$，然而，使用内积的形式存在一个问题，两个向量可以进行内积运算的前提是两个向量的维度需要相同，然而在前文中提到 Query 是 300 维的 GloVe 词向量，Key 是图像特征图上某个空间位置的特征向量，这个向量的维度为 C，C 的值通常为 1024 或 2048 等，因此，在进行内积运算之前，基于词向量语义指导的注意力机制模块会使用一个线性变换矩阵 W 将 300 维的 Query 词向量变换为 C 维向量，由于这个线性变换过程是可导的，所以线性变换矩阵 W 中的参数可以通过梯度下降优化算法在训练过程中得到。图 4-3 为基于词向量语义指导的空间注意力机制模块的工作流程图。

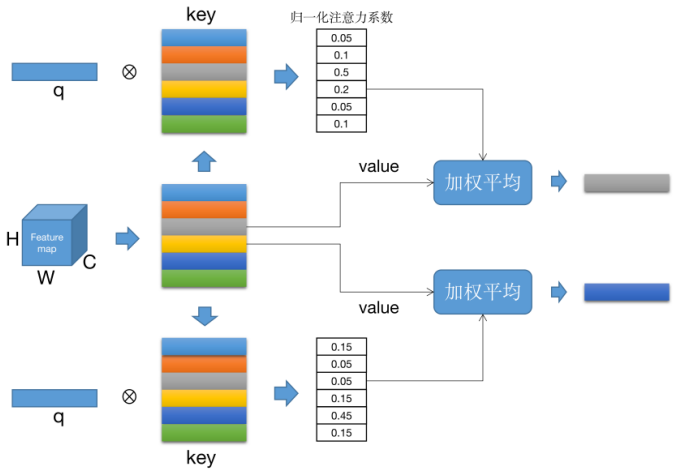

图 4-3 注意力机制模块工作流程

Figure 4-3 Workflow of attention module

4.4 实验验证

本节将会介绍使用基于词向量语义指导的空间注意力机制模块改进的模型在行人属性识别基准数据集 PA-100k 上的性能效果。另外，本节当中还介绍了同时使用基于词向量语义指导的空间注意力机制模块与第三章中提出的基于非对共现依赖关系图的方法改进的模型在行人属性识别基准数据集 PA-100k 上的性能效果。

4.4.1 基准模型设置

在本节介绍的实验中，所有的模型均选取 ResNet-18 作为深度卷积神经网络主干，这是因为在实验中发现对于 PA-100k 这种规模较小的数据集来说，使用过深的卷积神经网络容易引起过拟合现象，这会影响性能结果并干扰对性能结果数据的分析，因此，在本章中使用 ResNet-18 作为深度卷积神经网络主干。

为了验证基于词向量语义指导的空间注意力机制模块的有效性，实验中一共设置了三个模型：

(1) Baseline 模型：这个模型就是图 4-1 中的常规图像分类模型，该模型没有添加任何能够带来性能提升的神经网络结构，因此使用这个模型作为 Baseline 进行对比可以很容易的观察到基于词向量语义指导的空间注意力机制模块所带来的性能提升。

(2) 基于空间注意力机制模块改建的模型：这个模型与图 4-2 中的模型结构基本相同。唯一不同的就是该模型并未使用词向量作为注意力机制模块的 Query，而是直接随机初始化所有的 Query 向量并通过神经网络反向传播学习所有 Query 向量的参数。这个模型并未引入任何有关于行人属性的先验知识，设置这个模型的目的是为了验证使用词向量作为行人属性先验知识的有效性。

(3) 基于词向量语义指导的空间注意力机制模块改进的模型：该模型为图 4-2 中的模型，该模型引入了词向量作为有关于行人属性特征的先验知识，并依据这种先验知识在图像的特征图上搜寻与特定行人属性高度相关的图像特征。

除了以上三个模型之外，基于词向量语义指导的空间注意力机制模块还可以与第三章中提出基于非对称共现依赖关系图的行人属性识别方法进行结合，前者是对图像特征图进行处理，后者则是对分类器模块进行改进，所以这两个改进方法可以不加冲突的结合在一起，进一步提升行人属性识别任务的性能。因此，在实验部分，另外设置了两个模型用于评估这两种改进方法带来的性能提升。第一

个模型是对常规图像分类模型中的分类器模块使用基于非对称共现依赖关系图的方法进行改进的模型。第二个模型则是同时使用基于非对称共现依赖关系图的方法与基于词向量语义指导的空间注意力机制模块进行改进的模型。

4.4.2 实验结果

本节将会介绍上一小节中提出的五个模型在行人属性识别基准数据集 PA-100k 上的性能评估结果。实验中所有的超参数设置以及数据预处理方式与第三章中的实验完全相同。

表 4-1 展示了常规图像分类模型与使用和不使用词向量语义指导的空间注意力机制模型的的性能评估结果。由表 4-1 可以观察到，相比于不使用注意力机制的常规图像分类模型来说，使用注意力机制可以提升行人属性识别任务的性能。另外，引入了词向量语义指导的空间注意力机制模型的性能要优于不使用词向量语义指导的空间注意力机制模型的性能，这充分的说明了引入先验知识的有效性。词向量中所包含的有关于行人属性的先验知识可以指导注意力机制模块准确的提取与特定行人属性高度相关的特征，进而提升行人属性识别任务的性能。

表 4-1 注意力机制模块性能评估结果

Table 4-1 Performance evaluation results of attention module

Methods	**All**						
	mAP	***CP***	***CR***	***CF1***	***OP***	***OR***	***OF1***
Baseline	62.4	68.8	49.3	57.4	86.1	76.6	81.1
Attention without GloVe	63.5	68.1	**51.2**	58.45	85.3	**78.1**	81.6
Attention with GloVe	**64.3**	**75.1**	47.9	**58.46**	**86.3**	77.8	**81.9**

表 4-2 非对称共现依赖关系图+注意力机制模块性能评估结果

Table 4-2 Performance evaluation results of graph &attention module

Methods	**All**						
	mAP	***CP***	***CR***	***CF1***	***OP***	***OR***	***OF1***
Graph	64.6	69.0	52.6	59.7	84.7	79.0	81.8
Graph & Attention	**65.0**	**72.0**	**53.2**	**61.2**	**87.1**	**80.1**	**83.4**

表 4-2 展示了上一小节中提出的最后两个模型的性能评估结果。通过对比表 4-1 和 4-2 可以发现，常规图像分类模型的性能最差，单独使用基于词向量语义指

导的空间注意力机制模块或单独使用基于非对称共现依赖关系图的方法对常规图像分类模型进行改进均可以一定程度的提升性能，但是同时使用这两种改进方法所得到的性能提升更大。这个实验结果说明了行人属性识别任务的性能可以通过同时使用本课题中提出的两种改进方法进行提升，单独使用其中任意的一个方法也是有效的，但是同时使用两种方法通常能获得更优的性能。

4.5　本章小结

本章首先通过对行人属性识别数据集中的行人图片进行观察，分析了行人图片的一般图像布局模式以及行人图片中存在的问题，并阐述了使用注意力机制解决这些行人图片中存在的问题的基本思想。接下来介绍了如何通过注意力机制改进传统的图像分类模型，并详细的描述了基于词向量语义指导的空间注意力机制模块的设计思想与具体结构。最后，本章通过在行人属性识别基准数据集上进行实验，充分有效的证明了基于词向量语义指导的空间注意力机制模块可以提升行人属性识别任务的性能。本章中提出的基于词向量语义指导的空间注意力模块仅仅引入了少量的参数，并未大幅度增加模型的复杂度，但能够有效的提升识别性能。基于词向量语义指导的空间注意力模块不仅可以应用于行人属性识别任务，还可以应用于其他的隶属于图像分类的多种应用当中，具有很强的可扩展性，另外，基于词向量语义指导的空间注意力模块可以与第三章中提出的基于非对称共现依赖关系图的改进方法进行结合，进一步分提高行人属性识别任务的性能。

5 结论

为了进一步提升现代行人属性识别模型的性能以达到投入应用的标准，本文从两个方面探索了提升现代行人属性识别模型的性能的方法。首先通过对现有的基于图卷积神经网络的行人属性识别方法进行分析，发现了其中存在的两个问题，即低频行人属性特征信息不足与多数节点特征不发生聚合的问题，进而提出了一种新的定义共现依赖关系图中边的方式，并与原有模型进行结合，以解决原有模型中存在的问题。另外，通过对行人属性识别数据集的观察，发现了行人图片中存在的空间位置分布规律，这种空间位置信息可以在行人图片受到干扰时准确的提取行人属性特征。本文进一步分析了现有模型在提取空间位置信息中存在的问题，并提出使用词向量语义指导空间注意力机制模块提取与行人属性高度相关的图像特征的思想。

本文的主要工作及贡献点如下：

(1) 提出了一种基于非对称共现依赖关系图的行人属性识别模型，这个模型不仅能够在行人属性识别任务中应用，也可以应用于许多不同的基于多标签图像分类的应用当中。使用 ResNet-50 作为主干网络的情况下，在行人属性识别基准数据集上，这种方法能够将 mAP 提升约 3.3%。使用 ResNet-101 作为主干网络的情况下，在多标签图像分类基准数据集上，这种方法也能够将 mAP 提升约 0.6%以上。

(2) 提出了一种基于词向量语义指导的空间注意力机制模块用于改进现有的行人属性识别模型。增加了该模块之后仅仅引入了极少量的参数，在使用 ResNet-18 作为主干网络的情况下，能够将 mAP 提升约 1.9%。

(3) 将基于非对称共现依赖关系图的模型与基于词向量语义指导的空间注意力机制模块进行结合，进一步提升行人属性识别模型的性能。使用 ResNet-18 作为主干网络的情况下，相比于一般的图像分类模型能够将 mAP 提升约 2.6%，相比于只使用基于非对称共现依赖关系图的方法进行改进的模型来说，mAP 能够提升约 0.4%，相比于只使用基于词向量语义指导的空间注意力机制模块改进的模型来说，mAP 能够提升约 0.6%。

在本文所描述的实验中发现，如果数据集中出现类别分布不均衡的情况，使用图卷积神经网络从高频类别的特征中向低频类别的特征中迁移特征信息可以提升低频类别的识别性能。这个发现为解决类别不平衡问题提出了一种新的思路，未来的工作可以着眼于探究如何利用图卷积神经网络的特征传播能力为低频类别的特征进行信息补充。另外一方面，无论是图卷积神经网络还是注意力机制模块，

都可以选择很多种不同的形式，不同的形式有不同的优点，例如有的形式具有更高的准确度而有的形式复杂度更低，适合部署在线上进行推理，未来的工作仍可以继续探究何种形式更适合于行人属性识别任务，以进一步提升模型的准确度与推理速度，达到上线应用的标准。

参考文献

[1] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]. //IEEE Computer Society Conference on Computer Vision & Pattern Recognition. 2005: 886-893.

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

[3] Chang C C, Cj L. LIBSVM: a Library for Support Vector Machines[J]. ACM transactions on intelligent systems and technology (TIST), 2011, 2(3):27.

[4] Lafferty J, Mccallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. //Proc. 18th International Conf. on Machine Learning. 2001: 282-289.

[5] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[C]. //NIPS. Curran Associates Inc. 2012:1097-1105.

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

[7] Huang G, Liu Z, Laurens V, et al. Densely Connected Convolutional Networks[C]. //IEEE Conference on Computer Vision & Pattern Recognition. 2017: 4700-4780.

[8] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]. // International Conference on Learning Representations (ICLR). 2014: 1-8.

[9] Szegedy C, Wei L, Jia Y, et al. Going Deeper with Convolutions[C]. //IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[10] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[C]. //IEEE Conference on Computer Vision and Pattern Recognition. 2016:2818-2826.

[11] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[C]. //AAAI Conference on Artificial Intelligence. 2017: 4278-4284.

[12] Jia D, Wei D, Socher R, et al. ImageNet: A Large-Scale Hierarchical Image Database[C]. //Proc of IEEE Computer Vision & Pattern Recognition, 2009:248-255.

[13] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context[C]. // European Conference on Computer Vision. 2014:740-755.

[14] Liu X, Zhao H, Tian M, et al. HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis[C]. // IEEE International Conference on Computer Vision (ICCV). 2017:350-359.

[15] Jiang W, Yi Y, Mao J, et al. CNN-RNN: A Unified Framework for Multi-label Image Classification[C]. //IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:2285-2294.

[16] Wang Z, Chen T, Li G, et al. Multi-label Image Recognition by Recurrently Discovering

Attentional Regions[C]. //IEEE International Conference on Computer Vision. 2017:464-472.

[17] Feng Z, Li H, Ouyang W, et al. Learning Spatial Regularization with Image-Level Supervisions for Multi-label Image Classification[C]. //IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:5513-5522.

[18] Chen Z M, Wei X S, Wang P, et al. Multi-Label Image Recognition with Graph Convolutional Networks[C]. // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 5172-5181.

[19] Lecun Y, Boser B, Denker J, et al. Backpropagation Applied to Handwritten Zip Code Recognition[J]. Neural Computation, 1989, 1(4):541-551.

[20] Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions[C]. //IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:1800-1807.

[21] Howard A, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[22] Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]. //IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 4510-4520.

[23] Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3[C]. //IEEE/CVF International Conference on Computer Vision (ICCV). 2020: 1313-1324.

[24] Zhang X, Zhou X, Lin M, et al. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices[C]. //IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 6648-6856.

[25] Ma N, Zhang X, Zheng H T, et al. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design[C]. //European Conference on Computer Vision. 2018: 122-138.

[26] F Yu, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.

[27] Zeiler M D, Fergus R. Visualizing and Understanding Convolutional Neural Networks[C]. //European Conference on Computer Vision. 2014: 818-833.

[28] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[C]. //International Conference on International Conference on Machine Learning. 2015: 448-456.

[29] Zoph B, Le Q V. Neural Architecture Search with Reinforcement learning[C]. //International Conference on Learning Representations. 2017.

[30] Zoph B, Vasudevan V, Shlens J, et al. Learning Transferable Architectures for Scalable Image Recognition[J]. //IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 8697-8710.

[31] Tan M, Le Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks[C]. //International Conference on Machine Learning. 2019:6105-6114.

[32] Tan M, Chen B, Pang R, et al. MnasNet: Platform-Aware Neural Architecture Search for Mobile[C]. //IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 2815-2823.

[33] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align

and Translate[C]. //International Conference on Learning Representations. 2015.

[34] Luong M T, Pham H, Manning C D. Effective Approaches to Attention-based Neural Machine Translation[C]. //Conference on Empirical Methods in Natural Language Processing (EMNLP). 2015: 1412-1421.

[35] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[36] Cho K, Merrienboer B V, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]. //Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1724-1734.

[37] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. //Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). 2019: 4171-4186.

[38] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[C]. //Conference on Neural Information Processing Systems (NIPS). 2017: 6000–6010.

[39] Wang X, Girshick R, Gupta A, et al. Non-local Neural Networks[C]. //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7794-7803.

[40] Jie H, Li S, Gang S, et al. Squeeze-and-Excitation Networks[C]. //IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.

[41] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[C]. // European Conference on Computer Vision. 2018: 3-19.

[42] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks[C]. // International Conference on Learning Representations. 2017: 1-10.

[43] Remi G C. Wavelets on Graphs via Spectral Graph Theory[J]. Applied and Computational Harmonic Analysis, 2011, 30(2):129-150.

[44] Li Q, Han Z, Wu X M. Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning[C]. // AAAI Conference on Artificial Intelligence. 2018: 3538-2545.

[45] Wu X Z, Zhou Z H. A Unified View of Multi-Label Performance Measures[J]. arXiv preprint arXiv:1609.00288, 2016.

[46] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation[C]. //Conference on Empirical Methods in Natural Language Processing. 2014: 1532-1543.

[47] Al M, Ay H, Ay N. Rectifier Nonlinearities Improve Neural Network Acoustic Models[C]. // International Conference on Machine Learning. 2013: 1-6.

[48] Loshchilov I, Hutter F. SGDR: Stochastic Gradient Descent with Warm Restarts[C]. // International Conference on Learning Representations. 2017.

[49] Chen S F, Chen Y C, Yeh C K, et al. Order-Free RNN with Visual Attention for Multi-Label Classification[C]. //AAAI Conference on Artificial Intelligence. 2018: 6714-6721.

[50] Lee C W, Fang W, Yeh C K, et al. Multi-Label Zero-Shot Learning with Structured Knowledge Graphs[C]. // IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 1576-1585.

[51] Ge W, Yang S, Yu Y. Multi-Evidence Filtering and Fusion for Multi-Label Classification,

Object Detection and Semantic Segmentation Based on Weakly Supervised Learning[C]. //IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 1277-1286.

[52] 李洋, 许华虎, 卞敏捷.基于 CNN-ATT-ConvLSTM 的行人属性识别[J].计算机应用与软件, 2021, 38(04):152-158.

[53] 胡剑波, 任劼, 郑江滨. 基于注意力模型的行人属性识别方法[J].科学技术创新,2021, (05):63-65.

作者简历及攻读硕士学位期间取得的研究成果

一、作者简历

戚余航，男，1996 年 8 月生。2014 年 9 月至 2018 年 6 月就读于大连海事大学信息科学与技术学院电子信息工程专业，取得工学学士学位。2018 年 9 月至 2021 年 6 月就读于北京交通大学电子与信息工程学院通信与信息系统专业，研究方向是信息网络，取得工学硕士学位。攻读硕士学位期间，主要从事行人属性识别领域的工作。

二、发表论文

［1］**Qi Y**, Guo Y and Chen Y. Multi-label Image Recognition with Asymmetric Co-occurrence Dependency Graphs. 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA), 2021, pp. 287-294, doi: 10.1109/ICBDA51983.2021.9403091.

三、参与科研项目

[1]基于图表示与注意力机制的行人属性识别

[2]基于异质信息网络的跨领域推荐系统

[3]基于深度学习的习题理解和应用

.

.

.

.

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和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通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了谢意。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文数据集

表 1.1： 数据集页

关键词*	密级*	中图分类号	UDC	论文资助
行人属性识别；图神经网络；共现关系；词向量；空间注意力	公开			
学位授予单位名称*		学位授予单位代码*	学位类别*	学位级别*
北京交通大学		10004	工学	硕士
论文题名*		并列题名		论文语种*
基于图表示与注意力机制的行人属性识别算法研究				中文
作者姓名*	戚余航		学号*	18120118
培养单位名称*		培养单位代码*	培养单位地址	邮编
北京交通大学		10004	北京市海淀区西直门外上园村 3 号	100044
学科专业*		研究方向*	学制*	学位授予年*
通信与信息系统		信息网络	3	2021
论文提交日期*	2021.5.7			
导师姓名*	郭宇春		职称*	教授
评阅人	答辩委员会主席*		答辩委员会成员	
电子版论文提交格式 文本（ ） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ） 推荐格式：application/msword；application/pdf				
电子版论文出版（发布）者		电子版论文出版（发布）地		权限声明
论文总页数*	49			
共 33 项，其中带*为必填数据，为 21 项。				