摘　要

多标签图像分类的研究目标是准确预测出给定样本中存在的多个目标对象，广泛应用于图像检索、自动注释和智能监控等实际场景。由于多标签图像的复杂性，目前多标签图像分类的研究存在两个关键的问题需要解决：一、如何挖掘类别之间的依赖关系。二、多标签图像中的对象大小尺度不同，特征提取过程中小目标的特征信息易丢失，导致小目标分类准确率低的问题。于是，针对上述两个问题，本文提出了一种新的多标签图像分类模型。

同时，由于复合表情识别本质上也属于多标签图像分类问题，本文对提出的多标签图像分类模型进行一定的修改，将其应用于复合表情识别中。接下来将对本文的主要工作内容进行概括：

（1）为了更好地挖掘类别之间的依赖关系，本文提出了一个多头图注意力模块。具体而言，首先构建一个初始标签图，图的节点表示各个类别标签的特征向量，图节点之间的边表示类别标签之间的相关性，然后通过图注意力网络的自注意力机制学习图节点之间的相关性，并进行特征交互，最终得到包含标签相关性的判别特征用于最终的分类，从而提升模型的分类准确率。

（2）为了提高多标签图像中小目标的分类准确率，本文提出了一个多尺度语义注意力模块。该模块通过使用多个不同卷积层输出的特征图进行特征融合，用于增强小目标的特征信息，从而提升小目标的分类准确率。同时使用标签词嵌入向量指导学习在特征图中与各个类别相关的特征信息，得到特定于类别的特征向量，并将其作为多头图注意力模块的输入。

（3）在复合表情识别任务中，一个复合表情标签通常是由多个基本表情标签组成，如“惊恐”是由“惊讶”和“恐惧”这两个基本表情标签组成的，这正符合多标签图像的定义。基于以上分析，本文将复合表情数据集转换成多标签数据集，然后利用所提出的多标签图像分类模型来解决复合表情识别问题。

（4）本文在 MS COCO 和 VOC 2007 这两个多标签图像数据集和 JAFFE 表情数据集上进行了实验，验证了模型的有效性。在 MS COCO 和 VOC 2007 数据集上的实验结果表明，本文提出的模型相比相同环境下的 ML-GCN 模型在 mAP 指标上分别取得了 2.7%和 1.6%的提升，并且在 JAFFE 表数据集上的 mAP 指标相比 DBM-DACNN 模型也有 5.1%的提升。

关键词：多标签图像分类；标签相关性；图注意力网络；复合表情识别

Abstract

The research goal of multi-label image classification is to accurately predict multiple target objects in a given sample, which is widely used in practical scenarios such as image retrieval, automatic annotation, and intelligent monitoring. Due to the complexity of multi-label images, there are two key problems to be solved in the current research on multi-label image classification: First, how to mine the dependencies between categories. Second, the size and scale of objects in multi-label images are different, and the feature information of small objects is easy to lose during the feature extraction process, resulting in the problem of low classification accuracy of small objects. Therefore, in response to the above two problems, this paper proposes a new multi-label image classification model.

At the same time, since the composite expression recognition is essentially a multi-label image classification problem, this paper makes some modifications to the proposed multi-label image classification model and applies it to the composite expression recognition. The main content of this paper will be summarized in the following:

(1) To better mine the dependencies between categories, this paper proposes a multi-head graph attention module. Specifically, first construct an initial label graph, the nodes of the graph represent the feature vectors of each category label, and the edges between graph nodes represent the correlation between category labels, and then learn the graph through the self-attention mechanism of the graph attention network The correlation between nodes, and feature interaction, and finally obtain the discriminant features including label correlation for the final classification, thereby improving the classification accuracy of the model.

(2) To improve the classification accuracy of small objects in multi-label images, this paper proposes a multi-scale semantic attention module. This module uses the feature maps output by multiple different convolutional layers for feature fusion to enhance the feature information of small targets, thereby improving the classification accuracy of small targets. At the same time, the label word embedding vector is used to guide the learning of feature information related to each category in the feature map,

and the category-specific feature vector is obtained, which is used as the input of the multi-head map attention module.

(3) In the compound expression recognition task, a compound expression label is usually composed of multiple basic expression labels, such as "panic" is composed of two basic expression labels "surprise" and "fear", which is in line with the multi-label image. definition. Based on the above analysis, this paper converts the composite expression dataset into a multi-label dataset, and then uses the proposed multi-label image classification model to solve the composite expression recognition problem.

(4) In this paper, experiments are conducted on two multi-label image datasets, MS COCO and VOC 2007, and the JAFFE expression dataset, to verify the effectiveness of the model. The experimental results on the MS COCO and VOC 2007 datasets show that the proposed model has achieved 2.7% and 1.6% improvement in mAP indicators compared with the ML-GCN model in the same environment, and the JAFFE table dataset Compared with the DBM-DACNN model, the mAP index is also improved by 5.1%.

Key words: Multi-label image classification; Label relevance; Graph Attention network; Compound expression recognition

目　录

图目录

表目录

1 绪论

1.1 研究背景与意义

随着大数据时代的到来和深度学习的快速发展，人工智能技术已成为了大众关注的焦点，并将新的科技革命寄希望于人工智能中。计算机视觉是人工智能领域中一个重要的分支，其研究的主要目的是让计算机能够像人眼一样识别现实世界中存在的事物。而图像分类是计算机视觉领域中的一个基础方向，也是很多应用领域的基础，其目标是准确识别给定图像中物体的具体类别。图像分类还可进一步细分为单标签图像分类和多标签图像分类。其中单标签图像分类[1]指的是给定样本只有一个类别标签，需要计算机预测出该样本的类别标签，如图 1−1 所示。多标签图像分类[2]则指的是给定样本包含多个类别标签，需要计算机同时预测出给定样本中包含的所有类别标签，如图 1-2 所示。相比于单标签分类，多标签图像分类更具有实际意义，因为在现实世界中的自然场景中，通常都是多个对象同时存在的。

随着许多强大的深度学习模型[3-6]的出现，单标签图像分类的准确率得到了显著的提升。于是，许多研究者将目光转向了更具挑战性的多标签图像分类。但是相比于单标签图像来说，多标签图像分类包含的对象更多且更复杂，也就意味着传统的单标签图像分类模型无法很好地应用到多标签图像分类中。所以，研究者们开始研究新的方法来应对多标签图像分类中的挑战，并且取得了很大的进展。最初研究人员提出一种简单直接的解决办法就是将多标签图像分类问题转换成一组二元分类问题[7-9]。然而，这种孤立地处理每个对象忽略了对象之间的共现关系，因为在现实世界中共同出现的对象之间可能存在依赖关系，即标签相关性[10-12]。以运动场景为例，当一张图片中同时出现人、羽毛球拍等类别时，可以推测该图片中存在羽毛球这个类别的可能性较大。这是因为在同一个场景中，人、羽毛球拍和羽毛球这三个类别通常会同时出现，这表明了它们之间存在相关性。因此，在多标签分类任务中建模标签之间的相关性对于提升分类性能至关重要。通过对标签之间相关性的建模，我们可以更准确地预测图片中可能存在的标签，从而提高分类准确性。

因为多标签图像包含多个不同尺度大小的目标对象，由此引发的一个问题：在多标签图像中小目标对象的分类准确率低。小目标指的是：在图像中所占像素点很少的类别。而这类小目标对象在图像特征提取过程中，该类别的信息很

容易在低分辨率特征图中丢失，从而导致模型的分类准确率低。由此可见，小目标对象的特征丢失也是影响多标签图像分类性能的一个重要因素。

因此，如何更好地挖掘标签之间的依赖关系以及增强小目标的特征信息，是多标签图像分类研究过程中需要解决的关键问题。并且在进行多标签图像研究的过程中，发现复合表情本质上也属于多标签图像。于是，为了进一步提升复合表情的识别准确率，我们将所提出的多标签图像分类模型进行一定的修改并将其应用于复合表情识别任务中，这对复合表情识别研究具有重要的意义。

面部表情是表达人类情感最具表现力的方式之一[13]。准确地识别人脸表情具有极大的现实意义。过去的几十年间，人脸表情识别一直是热门的研究方向。早期的人脸表情识别研究主要还是关注于开心、惊讶、愤怒、厌恶、悲伤和恐惧这六类基本表情[14]。但是，在现实生活中人们的情绪表达有时非常复杂，单纯地使用一种基本表情标签来定义并不准确，因此就需要依靠多种基本表情组成的复合标签来诠释。例如在突然见到好久不见的朋友时流露出来的惊喜，则由惊讶和开心组成；惊恐则是由惊讶和恐惧两种基本表情组成，这类表情被定义为复合表情。尽管“复合表情”的概念早就已经提出，但大多数研究人员还没有将其作为一个多输出问题来处理。本文深入分析了复合表情的定义，发现复合表情标签可以拆分为多个基本表情标签，即复合表情图像也可以看作多标签图像。于是为了更好地提升复合表情识别的准确率，本文将复合表情识别问题转换成多标签图像分类问题，并运用本文提出的多标签图像分类模型来解决复合表情识别问题。

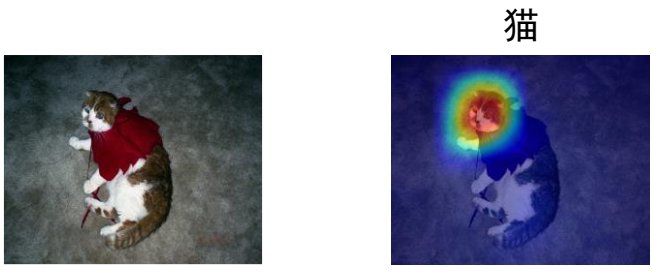


图 1-1 单标签图像分类

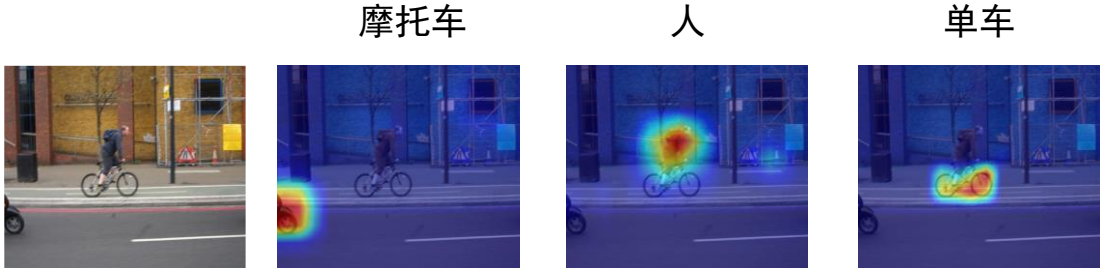


图 1-2 多标签图像分类

1.2 国内外研究现状

1.2.1 多标签图像分类研究现状

由于深度学习的不断发展，越来越多的研究人员将深度学习模型应用于单标签图像分类，并且极大地提升了单标签图像分类的准确率。然而现实世界的自然场景下获取的图像通常都包含不止一个目标对象，因此多标签图像分类比单标签图像更具挑战性和实际应用价值。于是，许多的研究者开始关注多标签图像分类，并探索各种方法来提高其准确性。因为深度学习模型展现出卓越的图像特征提取能力，所以现在的研究者大多都倾向于使用深度学习模型作为基础网络。而在基于深度学习模型的方法中，本文按照解决问题的思路和方法，将其分为以下三类：

（1）基于候选框的多标签图像分类方法：其主要思想是从图像中提取所有可能的候选区域，然后将每个候选区域作为目标图像进行单标签图像分类。2014 年，Girshick 等[15]和 Oquab 等[16]分别提出了两种基于候选框的多标签图像分类和检测方法。虽然在分类准确率上，上述两种方法都有很大的提升，但是它们极其依赖数据集中标注的真实边界框，当换成一个没有进行标注边界框信息的心多标签图像数据集时，上述两种方法的性能可能会下降。为了使模型不再依赖于数据集，Wei 等人[17]在 2015 年提出了 HCP 模型。该模型不再需要数据集标注真实边界框，而是借助于 Bing[18]和 EdgeBoxes[19]等方法在图像中自动生产边界框或分段的假设池，然后将所有候选区域输入到一个共享的 CNN 网络中提取各个候选区域图像特征，最后通过在 CNN 中添加一个新的池化层来预测不同标签的概率分布。

虽然 HCP 不再依赖于数据集的标注边界框信息，但是该模型需要生成大量的候选区域，十分浪费计算资源。而且多标签图像分类的目标是预测一组对象类别而不是生成所有可能对象的准确空间位置。于是，2020 年，Gao 等人[20]提出 MCAR 模型。该模型首先将整张图片作为输入一个 CNN 网络中提出全局的图像特征，然后通过一个 MCAR 模块利用注意力集中自动学习输入图像中存在目标对象的候选区域，然后将得到的候选区域再次输入同一个 CNN 网络中提取图像特征，得到候选区域的局部特征。然后结合全局图像特征和候选区域的局部特征，进行多标签分类。该模型可以通过注意力机制自适应地生成目标区域，不需要对数据集进行标注边界框也不需要采用边界框生成技术，但是该模型需要在训练之前预先定义每张图像生成目标区域的数量。

（2）基于视觉注意力机制的多标签图像分类方法：注意力机制作为一种简洁有效的处理方法，被广泛应用于分类、分割等各类视觉任务中，因此一些研究者也将注意力机制引入多标签分类领域，用于隐式地建模不同标签之间的空间关系。其主要思想是感兴趣的类别可能只位于图像的某些空间区域，使用注意力机制关注标签在图像中特定区域的位置，更好地学习每个标签最具有分辨的图像特征，从而提高分类性能。2017 年，Zhu 等人[21]提出了一种名为 SRN 的多标签图像分类模型，该模型基于注意力机制，利用标签之间的语义和空间关系进行监督。对于多标签图像，SRN 生成所有标签的注意力图，并利用可学习的卷积网络捕获它们之间的依赖关系，从而提供空间正则化，提高多标签分类准确率。类似的，为了解决基于候选框的多标签图像分类方法导致冗余计算的问题， Wang 等人[22]在 2017 年提出了一种循环记忆注意力机制模块，通过引入一个空间转换层，以达到自适应在卷积特征映射中搜索到语义感知区域的目的，同时使用一个 LSTM 网络[23]对这些捕捉到的语义感知区域通过序列预测的方式以进行多标签预测。但是由于注意力区域仅通过图像级监督学习，该模型只能粗略地定位目标对象存在的区域。人类观察图像的注意力通常是从图像中央位置的对象开始然后转移到下一个显著的对象目标上。受这一方式的启发，并通过引入循环注意力机制和强化学习，在 2018 年 Chen 等人[24]提出了一种端到端的循环注意力强化学习框架，用于多标签图像分类。该框架由全卷积网络和 LSTM 循环注意力感知模块组成。全卷积网络用于提取图像深度特征表示，LSTM 循环注意力模块用于迭代搜索与标签相关的区域并预测这些区域的标签得分。LSTM 网络可以"记住"之前迭代的信息，从而自然地捕获注意力区域之间的上下文依赖关系，从而提高分类准确率。该方法同样不需要数据集标注真实边界框，仅图像级别的监督即可进行端到端的训练。

人类眼睛的视觉感知对图像的变换具有很强的不变性。如一张图像进行旋转、缩放、反转等操作以后，人类照样仍然可以准确识别图像。受这一特性的启发，Guo 等人[25]提出了 VAC 模型。该模型采用双分支网络，分别以原始图像和其转换图像作为输入。为了衡量两个分支之间的注意力热图的一致性，该模型引入了一种新的注意力一致性损失。然后通过两个损失函数来优化网络模型，提升模型对图像变换的感知不变性。上述采用注意力机制关注感兴趣区域的方法，智能通过图像级的监督，缺乏明确的语义指导。针对上述问题，在 2020 年，You 等人[26]提出一种新的基于图嵌入的交叉跨模态注意力网络。该模型利用一种新的图嵌入方法(ASGE)来捕获标签之间的语义关系，然后提出了跨模态注意力机制(CMA)方法，该方法使用学习到的标签语义图来指导学习与标签相关的

注意力图，然后将图像特征分解为与标签相关的向量，最后将这些向量用于最终的分类。

（3）基于标签相关性的多标签图像分类方法：由于对象通常在现实世界中共同出现，共同出现的对象之间可能存在联系，因此，在多标签图像分类任务中，建模标签之间的依赖关系是提高分类性能的关键[27]。于是，越来越多的研究者专注于如何更好地探索和挖掘标签之间的依赖关系。2016 年，Wang 等人[28]引入递归神经网络来显式地建模标签之间的依赖关系，结合 CNN，提出一个 CNN-RNN 模型。该模型首先通过 CNN 主干网络提取图像特征，但是由于 CNN 提取的图像特征并不包含标签之间的依赖关系，于是该模型使用 LSTM 来对标签之间的依赖关系进行建模，挖掘标签之间的关系，从而提升模型的多标签分类准确率。LSTM 模块被用来捕获标签依赖关系，但由于它是一种序列模型，需要将标签排序并按照特定顺序输入，这会限制 CNN-RNN 模型的灵活性。于是，Chen 等人[29]提出了 Order-Free RNN 模型,该模型不需要预先确定用于预测的标签顺序，而是使用引入的 LSTM 模型依次学习标签依赖性，从而减少标签顺序的约束。并且该模型中引入的注意力模块可以使模型专注于与每个标签关联的图像区域，从而即使对象的尺寸较小，也会获得更好的预测。

随着图神经网络的发展以及图神经网络在计算机视觉领域逐渐兴起后，一些研究者开始将 GCN[30]引入到多标签分类任务中。2019 年，Chen 等人[31]首次使用 GCN 建模标签依赖性，提出了 ML-GCN 模型来捕获多标签图像分类的标签相关性。该模型使用标签词向量和统计得到的标签共现概率构建一个标签图，并使用 GCN 学习一个包含标签依赖关系的分类器。为了显式建模标签依赖关系，该模型设计了标签相关矩阵，用于指导 GCN 中节点之间的信息传播。此外，还提出了一种重新加权方法，重新分配中心节点和邻居节点的权重系数，以缓解过度平滑的问题。重新分配中心节点和邻居节点的权重系数。然后将卷积网络提取的图像特征进行全局池化，然后再与 GCN 训练得到的分类器结合，得到最后预测标签的概率分布。但是 ML-GCN 的标签图需要人工统计训练集的标签共现信息设计得到的，对于不同的数据集进行测试时，需要重新制定标签关系图，使得该方法的灵活性较差。于是，Li 等人[32]提出了 A-GCN 来解决 ML-GCN 中不灵活的问题。该模型通过使用一个自适应标签图模块学习标签相关性，然后生产标签相关矩阵，得到标签图。然后利用学习得到的标签图输入 GCN 中，学习得到具有标签依赖关系的分类器用于最终的分类，以此来进行多标签图像分类。并且由于标签共现在当前流行的标签数据集中是稀疏的，还引入了一个稀疏相关性约束。

2020 年，Zhang 等人[33]提出了一种 KSSNet 模型。该模型提出了一种新的标签相关性建模方法，由于标签共现概率得到的统计图会受到噪声的干扰，通过在统计图的基础上叠加知识先验图，然后在最终的叠加图上应用多层 GCN。并且通过在网络中添加多个卷积层与 GCN 之间的横向连接，将标签相关性的信息注入主干 CNN，用于特征学习过程中的标签感知。

同年，Chen[34]等人提出了一种特定于语义的图表示学习模型 SSGRL。虽然该模型也是采用标签共现概率构建标签图，但是 ML-GCN 的思路并不相同。SSGRL 侧重于图像特征的学习。通过将图像特征解耦为特定于类别的特征向量，然后通过图传播网络来学习标签之间的相关性。该模型的具体实现方式是首先利用低秩双线性池化方式，用标签词嵌入向量指导学习特定于类别的特征向量，然后将得到的特定于类别的特征向量输入到门控图神经网络中进行特征交互，学习到包含标签相关信息的特征向量，提升模型的分类准确率。

然而像 ML-GCN 和 SSGRL 这两种方法为整个数据集构建一个全局标签图可能会导致大多数常见数据集中的概率偏差问题，于是 Ye 等人[35]提出了一种注意力驱动动态图卷积网络(ADD-GCN)。该模型利用内容感知的类别表示来构造动态图表示。首先，ADD-GCN 通过 SAM 模块将卷积网络输出的特征图分解成多个内容感知类别表示。然后，将其输入到 D-GCN 模块中，生成判别向量用于多标签分类。Chen 等人在 ML-GCN 的基础上进一步提出 P-GCN[36]，通过一个标签感知调制模块将整体图像表示分解为与不同类别相关联的一组特征，并采用 GCN 学习相互依赖的类别特征用于分类，提升多标签图像分类的准确率。Nguyen 等人[37]提出了一种 MGTN 模型，该模型将图分为多个子图，然后在子图上学习标签之间的相关性，从而更好地挖掘标签之间的相关性。

1.2.2 复合表情识别研究现状

虽然人脸表情识别已经有几十年的研究了，但是以往的人脸表情识别研究中，主要关注的还是基本表情识别。但是，现实生活中的人脸表情往往是由多种基本表情组合而成的，而不仅仅是单一表情类别的表达，这些复杂的表情被称为复合表情[38]。如图 1-3 中第二行所示。因为复合表情是现实生活中真实存在的情绪表达，为了使复合表情识别任务更接近于现实，也更有应用价值，复合表情的图像样本都来自于现实世界的自然场景，而这种采集数据的方法也带来了一些问题，即自然场景下获取的复合表情样本标注困难，导致复合表情数据集的缺乏。于是，复合表情识别研究仍然极具挑战性。

图 1-3 基本表情与复合表情样本

由于大数据时代，许多大规模人脸表情数据集的完善，以及计算机硬件的发展使得计算机处理数据的能力极大地提高，所以使用基于深度卷积网络提取人脸表情特征的方法的识别准确率大幅度的提升，并且无需繁琐的手工设计。于是，在表情识别领域中，手工提取表情特征的方法逐渐被深度学习模型的特征提取方法所替代。而随着基本表情的识别准确率不断提高，许多的研究者则开始将目光聚焦于更具挑战性的复合表情识别。Benitez-Quiroz 等人[39]在 2017 年提出了一种利用面部表情图像中检测到的 AU 单元进行复合表情人脸识别的方法。然而，为了达到有效识别人脸表情的目的，该方法首先需要提高人脸表情中 AU 单元的检测准确率。Deng 等人[40]通过改进深度卷积网络提出了 DLP-CNN 模型，用于复合表情识别，并取得了较好的效果。Deng 等人[41]在 2018 年提出了 Deep Bi-Manifold CNN，这是一种新型的深度流形学习网络模型，能够保留深度特征的局部相关性和表情标签的流形结构，从而更好地学习复合表情的特征信息。

1.3 本文的研究内容

本文通过同时考虑多标签图像分类目前存在的两个问题：如何挖掘标签之间的相关性以及如何提高小目标的分类准确率，提出了一种新的多标签图像分类模型 MSSGAT(Multi scale Semantic Attention Mechanism and Graph Attention Network for Multi label Image Classification)。并且为了提高复合表情识别的性能，将本文提出的多标签分类模型进行一定的修改，然后用于复合表情识别。接下来对本文的主要研究内容进行详细的介绍。

（1）因为在现实世界中，同时出现的多个目标对象之间往往存在一定的联系。通过学习标签之间的相关性，可以更好地捕捉目标对象之间的关联信息，从而提高分类准确率。为了更好地挖掘标签之间的相关性，本文提出使用图注意力网络（Graph Attention Networks, GAT）[42]来建模标签之间的相关性。具体而言，本文通过一个多尺度语义注意力模块将多标签图像特征图分解成特定于类别的特征向量，将其作为图节点，然后借助 GAT 中的自注意力机制对每张图像中分解得到的特定于类别的特征向量进行相关性计算，可以更好地捕获每张图像中类别之间的关系，从而增强特征的判别能力。最后本文将在两个多标签数据集中进行实验，通过与其他模型的实验结果对比证明本文使用图注意力网络自适应挖掘标签相关性的方法能够更好地提升模型的分类准确率。

（2）针对多标签图像中的小目标对象分类准确率低的问题，由于多标签图像在提取图像特征过程中，图像中小目标对象的特征信息很容易丢失，导致无法很好识别小目标。为了更好地学习小目标的特征表示，提升模型对小目标的分类准确率，本文提出了一个多尺度语义注意力模块。该模块首先将多个尺度的图像特征图进行融合，以增强小目标的特征信息。然后使用通过标签词嵌入向量将特征图分解为各个类别的特征向量，使模型在训练过程中更好获取每个标签的特征表示，并将其作为图的节点特征输入到多头图注意力模块中。

（3）以往的面部表情识别主要集中于传统的基本表情识别，而在现实世界中，人类的面部情绪表达往往并不仅是单一表情类别的表达，而是需要多种基本表情组合而成的复合表情标签来诠释。而对于这种表情图像来说，一个表情样本包含着多个基本表情标签，这刚好符合多标签图像的定义。本文将复合表情识别视为多标签图像分类问题，并对本文提出的多标签图像分类模型进行一定的修改，将其应用于复合表情识别中。在该模型中，我们利用语义注意力模块将表情特征分解成各个类别的特征向量，并利用图注意力网络中的自注意力机制自适应地挖掘表情标签之间的相关性，用于提升模型的识别准确率。最后本文在一个复合表情数据集上进行实验，并通过与现有的其他模型对比可知本文提出的模型的识别准确率有很大的提升。

1.4 本文的组织结构

本文分为五个章节，每个章节的内容安排如下：

第一章 绪论

本章介绍了多标签图像分类和复合表情识别的研究背景和意义，并概述了这两个研究方向的国内外研究现状，还简单叙述了相关方法。最后对本文的主要工作以及内容的结构安排进行了说明。

第二章 相关技术介绍

本章介绍了图卷积网络的相关概念和理论知识，还详细介绍了一些经典的图卷积网络模型，接着详细介绍了几种主流的多标签图像分类模型，最后还介绍了复合表情识别的一些内容，以及详细介绍了表情识别中三个关键步骤的相关知识。

第三章 基于 GAT 和多尺度语义注意力机制的多标签图像分类模型

本章将对基于 GAT 和多尺度语义注意力机制的多标签图像分类模型进行详细的介绍。首先对模型的整个模型图进行说明，接着详细介绍了该模型的三个关键模块，然后介绍了用于模型训练的损失函数、实验中用到的数据集和实验设置，最后对在两个数据集上的实验结果进行对比与分析。

第四章 多标签图像分类模型在复合表情识别上的应用

本章首先对复合表情进行分析，通过分析发现与多标签图像分类任务之间的关系，进而提出了本文基于 MSSGAT 模型进行修改的复合表情识别模型，并详细介绍了该模型的总体架构和各个模块的功能。随后，介绍了模型训练所使用的损失函数和实验设置，包括数据预处理和模型优化等内容。最后，详细介绍了本文在实验中所使用的数据集，并对实验结果进行了对比分析和讨论。

第五章 总结与展望

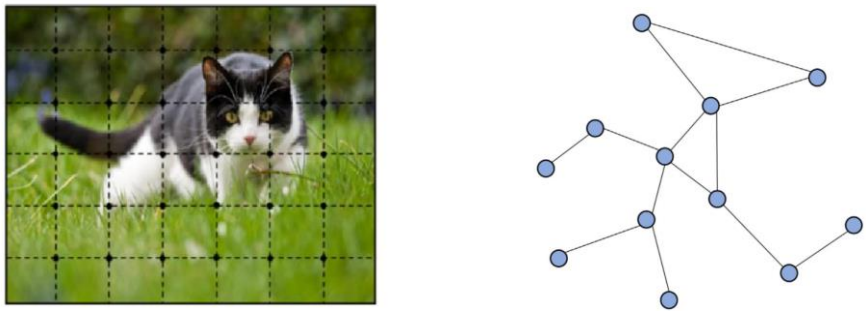
本章是本文的总结部分，将对本文的研究工作进行总结并提出未来的研究方向。首先对本文主要工作内容与创新进行总结，并在此基础上对本文工作的不足以及对未来如何针对这些不足加以改进进行了展望。

2 相关技术介绍

2.1 引言

本章将介绍一些本文工作涉及到的相关知识。首先，由于本文提出的多标签图像分类模型运用了图注意力网络，于是将对图卷积网络的一些相关知识进行详细的讲解。然后详细介绍两种基于图神经网络的多标签图像分类模型。本文的研究还涉及到了复合表情识别，在本章的结尾对复合表情的概念进行了说明，并详细说明了表情识别中三个关键步骤的相关知识。

2.2 图卷积网络



(a) 网格结构　　(b)图结构

图 2-1 网格结构和图结构数据

图卷积网络是一种最近发展十分迅速的图神经网络。图卷积网络的优势在于可以很好地处理非欧式空间的图结构数据，并且能够深度挖掘其数据的特征。众所周知，卷积网络在处理欧式空间的网格结构数据（如图像和视频数据）时，能够很好地提取数据的特征，这归功于卷积核拥有的特征机制：加权平均机制和参数共享机制。但是需要我们注意的是，这也导致了卷积神经网络的局限性，它只能处理欧式数据。如图 2-1(a)所示。然而，现实生活中存在着大量的非欧式空间数据，它们大多数以图的形式存在，如社交网络、交通运输网络，计算机网络以及分子结构等，如图 2-1(b)所示。因为传统的卷积网络无法应用于非欧空间的图结构数据中，所以许多研究者开始思考如何在图结构数据上进行卷积操作。于是，图卷积神经网络就此诞生，其本质就是扩展 CNN，使其可以应用于非欧空间数据。图卷积网络可以按照处理数据的方式分为两种不同的图卷积方法：基于频域的图卷积和基于空间域的图卷积。其中基于频域的方法源于

卷积定理的启示，它通过将空域信号转换为频域信号，然后进行图卷积操作；而基于空间域的方法则是通过定义聚合函数，在图数据上直接进行卷积操作。本节将分别对这两种方法进行介绍。

（1）基于频域的图卷积网络（Spectral-based Graph Convolutional Networks）：

基于频域的图卷积网络从卷积定理中得到启发，从而实现图卷积操作。即两个信号在空域做卷积运算后的傅里叶变化等于其经过傅里叶变换后的乘积。于是给定两个信号 f 和 g：

$$F(f*g)=F(f)\bullet F(g) \tag{2-1}$$

也可以对上式进行傅里叶逆变换得到下式：

$$f*g=F^{-1}(F(f)\bullet F(g)) \tag{2-2}$$

在上述公式中，F 和 F^{-1} 分别代表着傅里叶变换和逆变换操作。*和 $\bullet$ 则分别表示在空域的卷积操作和在频域的乘积操作。

于是，通过卷积定理的傅里叶变换将空域中的图数据和图卷积操作变换到频域中，然后在频域内进行乘积操作，接着通过傅里叶逆变换将在频域的乘积转换回原来的空域，这就完成了一次图卷积操作。其中，图结构数据进行傅里叶变换需要拉普拉斯矩阵，接下来将详细介绍图拉普拉斯矩阵的相关知识。

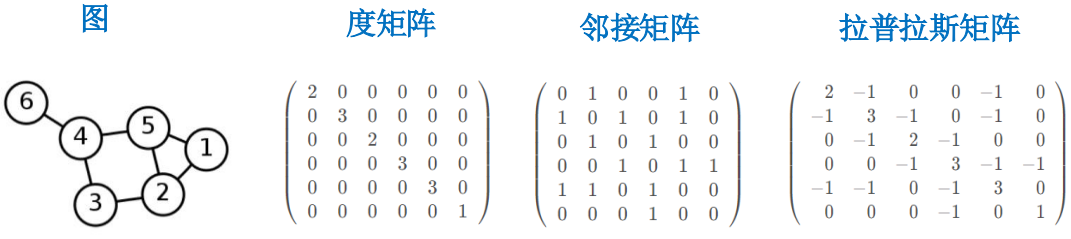


图 2-2 拉普拉斯矩阵

首先通过给定的图 G 构造出邻接矩阵 A 和度矩阵 D，然后通过计算 $L=D-A$ 得到图 G 的拉普拉斯矩阵 L，如图 2-2 所示。然后对其进行特征分解：

$$L=U\Lambda U^{T} \tag{2-3}$$

$$U=\{u_1,u_2,u_3,...,u_n\} \tag{2-4}$$

$$\Lambda=\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix} \tag{2-5}$$

其中，U 是由 L 的 n 个特征向量构成的矩阵。Λ 则是 n 个特征值组成的对角矩阵。然后使用分解得到的 U 矩阵对图信号 f 进行傅里叶变换：

$$\hat{f} = U^T f \tag{2-6}$$

其中，f 是图信号在空域下的输入。$\hat{f}$ 则是 f 经过傅里叶变换后在频域的表示。U^T 是特征矩阵 U 的转置。对 $\hat{f}$ 进行傅里叶逆变化：

$$f = U\hat{f} \tag{2-7}$$

最后，经过推导得到图卷积的计算公式为：

$$(f * g)_G = U((U^T f) \odot (U^T \mathrm{g})) \tag{2-8}$$

然后定义一个滤波器 $g_\theta = diag(U^T \mathrm{g})$，此时图卷积网络的公式可以简化为：

$$x * g_\theta = U g_\theta U^T x \tag{2-9}$$

基于频域的方法都是基于这个公式进行研究的，各种基于频域的方法之间的区别在于选择的滤波器g_θ不同。如 Bruna 等人[43]通过使用对角矩阵来代替谱域的卷积核，提出了第一代基于谱图的图卷积网络，然而这种方法的缺点就是计算复杂度大。2016 年，Defferrard 等人[44]提出了 ChebNet，用切比雪夫多项式来近似图信号的频谱，实现了图卷积操作。相比于谱卷积网络，切比雪夫网络具有更低的计算复杂度，并满足空间局部性。2017 年，Kipf 等人[30]进一步简化了 ChebNet，并提出了一个称为 GCN 的模型（Graph Convolutional Network）。GCN 采用 1 阶切比雪夫多项式进行图卷积，且每个卷积核只有一个参数，因此大大降低了模型的复杂度。其图卷积层公式如下所示：

$$H^{l+1} = f(H^l, A) = \sigma((\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^l W^l)) \tag{2-10}$$

其中，$\tilde{A} = A + I_N$，$\tilde{D} = \sum_j \tilde{A}_{ij}$。$H^l$ 表示第 l 个图卷积层的输入，W^l 表示模型需要学习的权重矩阵。$\sigma(\bullet)$ 代表激活函数。

于是，可以通过堆叠多个图卷积层构造出一个图卷积网络来建模和学习图节点之间的复杂关系，如图 2-3 所示。但是，这种方法也存在明显的局限性。基于频域的图卷积网络需要将图结构数据从空域转换到频域中进行卷积操作，所以该类方法都需要对拉普拉斯进行特征分解，得到特征矩阵 U，从而对图数

据进行傅里叶变换。因此基于谱域的方法的计算成本与图的大小成正比，于是它们很难应用于大型图。

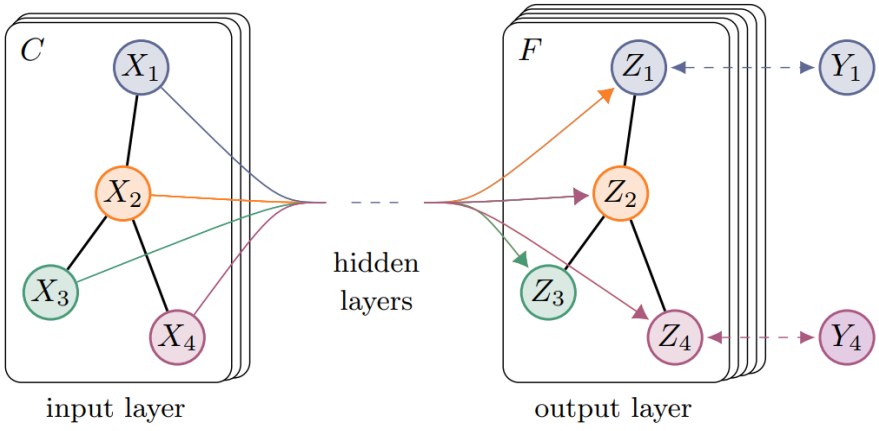
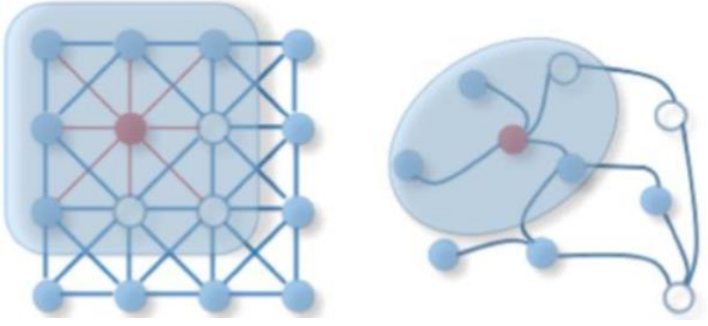


图 2-3 图卷积网络



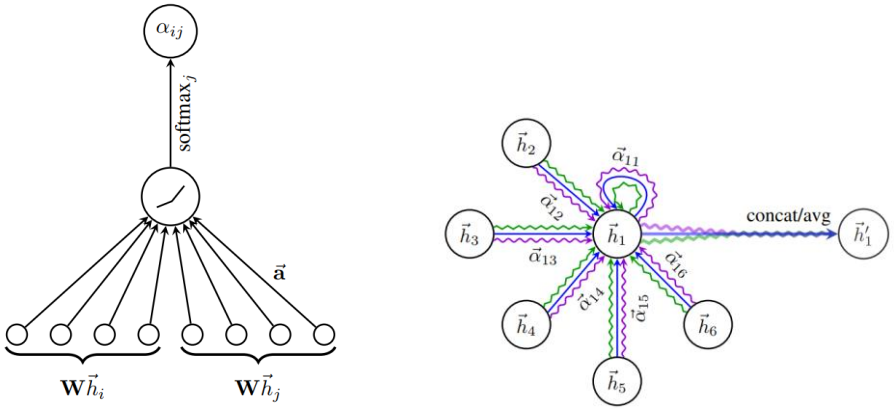
(a) 2D 卷积　　(b) 图卷积

图 2-4 卷积和图卷积

（2）基于空间域的图卷积网络(Spatial-based Graph Convolutional Networks):

基于空间域的图卷积网络的想法是定义一个函数直接在图结构数据上进行卷积操作。如图 2-4(a)所示，可以把 2D 图像中的每个像素点都看成是一个图节点，那么每个像素都与其附近的像素有边相连，图像就是一个规则的图结构，在该图上进行卷积操作，就是将红色节点和他它的 8 个邻居节点进行加权求和操作。于是，为了得到图 2-4(b)中红色中心节点的特征表示，其最简单一种解决方法就是对红色中心节点和邻居节点取平均值然后求和。但是，对于 2D 图像而言其中每个像素点的邻居节点是有序的，并且具有固定的邻居数量。而对于图来说，每个中心节点的邻居节点不仅无序而且数量不固定，所以基于空间域的图卷积操作就需要解决这些可变性的问题。在基于频域的图卷积网络中，可以将简化后的 GCN 中的滤波器看作图节点的聚合函数，在此基础上得到启发，从而引出了基于空域的图卷积网络[45-46]。

2017 年，Veličković 等人[42]提出了一种基于空间域的图卷积网络，即图注意力网络（GAT）。该网络引入了自注意力机制，通过在模型聚合节点特征的过程中自适应地关注与中心节点关联性更高的邻居节点，从而赋予其更高的权重。这种方法使得网络可以更加准确地聚合节点信息，提高了模型的性能。图注意力网络的模型结构如图 2-5 所示。



(a) 注意力机制　　(b) 多头注意力机制

图 2-5 图注意力机制

在图中 $\vec{h}_i \in R^F$ 代表着节点 i 的特征，图中所有节点的特征表示集合为 $h=\{\vec{h}_1,\vec{h}_2,...,\vec{h}_n\}$。然后，使用自注意力机制计算两个节点之间的注意力系数。计算公式如下所示：

$$e_{ij}=a(W\vec{h}_i,W\vec{h}_j) \tag{2-11}$$

其中，$W\in R^{F'\times F}$ 图注意力层中需要学习的权重矩阵，e_{ij} 为节点 j 与中心节点 i 之间的注意力系数。由于每个中心节点的邻居节点可能有多个，所有我们将节点的邻居节点表示为集合 N_i，依次计算 $j\in N_i$ 对于中心节点 i 的注意力系数 e_{ij}，然后对其进行归一化：

$$\alpha_{ij}=soft\max_j(e_{ij})=\frac{\exp(e_{ij})}{\sum_{k\in N_i}\exp(e_{ik})} \tag{2-12}$$

其中 α_{ij} 为归一化后的注意力系数，于是根据上述公式得到最终的图注意力机制为：

$$\alpha_{ij}=\frac{\exp(Leaky\,\mathrm{Re}\,LU(\vec{a}^T[W\vec{h}_i\,||\,W\vec{h}_j]))}{\sum_{k\in N_i}\exp(Leaky\,\mathrm{Re}\,LU(\vec{a}^T[W\vec{h}_i\,||\,W\vec{h}_k]))} \tag{2-13}$$

在上式中，$\bullet^T$ 代表转置操作，||则是表示连接操作。得到注意力权重后，可以通过加权求和的方式对中心节点的特征与邻居节点的特征进行聚合，得到新的中心节点特征表示：

$$\vec{h}_i' = \sigma(\sum_{j \in N_i} \alpha_{ij} W \vec{h}_j) \tag{2-14}$$

然后，可以使用多头注意力机制进一步提升模型的性能。具体而言，就是使用多个独立的注意力层来进行公式(2-14)的计算，最后将这多个单头注意力的输出特征连接起来，得到最终的输出：

$$\vec{h}_i' = \sigma(\frac{1}{K}\sum_{k=1}^{K}\sum_{j \in N_i} \alpha_{ij}^k W^k \vec{h}_j) \tag{2-15}$$

由于图注意力网络需要计算每个中心节点与邻居节点之间的相关性权重，于是在处理大规模图结构数据时，存在一定的不足。于是，为了能够适应大规模图数据，Hamilton 等人提出了图采样聚合网络（GraphSAGE）[47]。GraphSAGE 不同于 GAT，为了能够处理大规模图，它在进行特征聚合的过程中不会考虑所有的邻居节点，而是使用采样操作，对邻居节点进行采样。通过聚合固定数量的邻居节点的特征来更新中心节点的特征。采样的方式不仅可以降低处理的复杂度，也可以降低计算的难度，从而使模型可以很好地处理大规模的图结构数据。Atwood 等人[48]则利用随机游走得到的 K 跳转移概率来定义节点之间的权重，提出了扩散卷积神经网络（Diffusion Convolution Neural Networks，DCNN）。

综上所述，图卷积网络能够有效地处理图结构数据，并通过在图上进行卷积操作，将节点之间的信息进行交互和传递，从而学习每个节点的表示。这种表示不仅考虑了节点自身的特征，还包含了其邻居节点的信息。而多标签图像中各个标签之间的关系天然符合图结构，因此可以使用图这种数据结构来对标签之间相关性进行建模。并且通过使用 GCN 在图上进行卷积操作，将每个节点的信息与其邻居节点的信息相结合，实现标签之间的信息交互和传递，从而捕捉到标签之间的依赖关系，并将这些依赖关系纳入模型中进行预测，提升模型的分类准确率。

2.3 多标签图像分类模型介绍

多标签图像分类旨在使计算机可以准确预测出中给定图像中存在的多个对象标签。由于在现实世界中同时出现的目标对象之间通常存在依赖关系，因此可以通过挖掘目标标签之间的依赖关系来提升模型的分类准确率。随着图神经

网络的快速发展，图神经网络可以很好地建模图这种关系型数据结构的能力受到了许多研究者的关注，于是越来越多的人开始将图神经网络应用于多标签图像分类领域中，并取得了巨大的成功。于是，本节接下来将详细介绍两种使用图神经网络来挖掘标签相关性的多标签图像分类模型。

2.3.1 基于图神经网络的多标签图像分类模型

2019 年，Chen 等人[31]首次在多标签图像分类领域引入 GCN 并提出了 ML-GCN 模型，用于挖掘标签之间的依赖关系。该模型通过堆叠图卷积层学习标签之间的相关性，将标签的词向量通过学习映射成包含标签依赖关系的分类器，用于对图像特征进行多标签分类。ML-GCN 整个网络是端到端的。该模型的整体框架如图 2-6 所示。

整个 ML-GCN 模型由两个部分组成。其中上半部分为常规的图像特征提取模块，通常使用 CNN 网络来提取图像的高级特征表示，该模型使用的是 RseNet 作为该部分的特征提前网络。然后将得到的图像特征输入一个全局最大池化层，最终得到图像的特征表示。即给定图像 I，图像特征输出为 x，如下所示：

$$x = f_{GMP}(f_{cnn}(I;\theta_{cnn})) \in R^D \tag{2-16}$$

其中 θ_{cnn} 表示 CNN 的模型参数。

模型的下半部分为分类器学习模块。在输入该模块之前，需要构建出一个标签图。ML-GCN 使用标签词嵌入向量表示图节点，通过统计数据集标签共现概率构造图的二元邻接矩阵 A。首先通过统计训练集中不同标签共同出现的次数得到矩阵 $M \in R^{C\times C}$，C 表示类别的数量。然后，通过以下公式得到条件概率：

$$P_i = M_i \big| N_i \tag{2-17}$$

然后通过一个阈值 τ 来对相关性 P 进行二值化，如下所示：

$$A_{ij} = \begin{cases} 1 & \text{if } P_{ij} < \tau \\ 0 & \text{if } P_{ij} >= \tau \end{cases} \tag{2-18}$$

最终得到图的邻接矩阵。接着通过卷积层对标签图进行特征交互：

$$H^{l+1} = h(\hat{A}H^l W^l) \tag{2-19}$$

其中，H^l 表示第 l 层节点的特征表示，W^l 表示第 l 层的权重矩阵，H^{l+1} 表示第 l+1 层节点的特征表示，$h(\cdot)$为激活函数，$\hat{A}$ 表示标准化后的邻接矩阵，同时论文中为了防止过度平滑的问题，还提出了重加权的邻接矩阵构造方式。

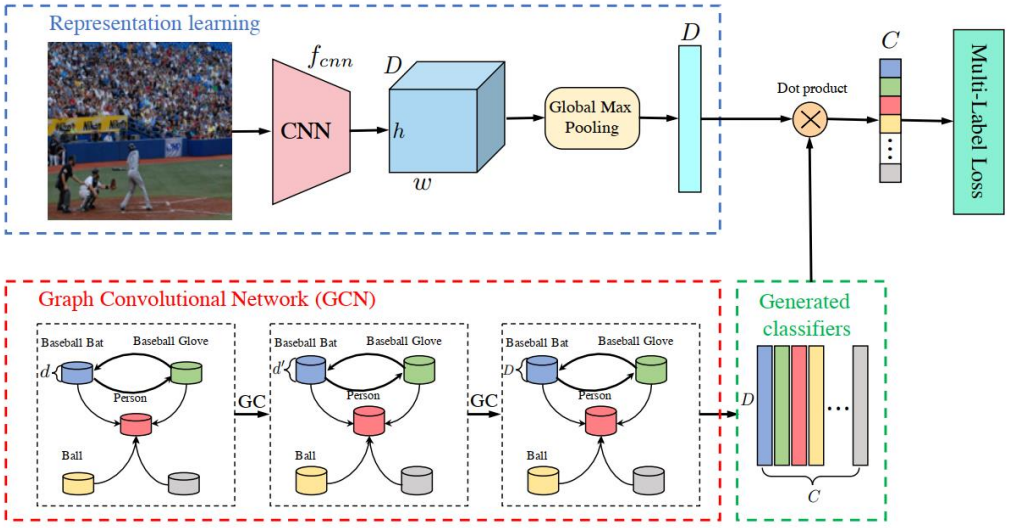


图 2-6 ML-GCN 模型

ML-GCN 的主要思想是利用图卷积网络学习一个包含标签相关信息的分类器，用于对经过全局池化后的图像特征进行分类。而 SSGRL[34]则与之不同。SSGRL 是运用图传播网络对图像进行解耦后的特征进行特征交互，挖掘标签相关性，使得最终输出的图节点特征包含标签相关性信息，然后进行分类。SSGRL 的模型如图 2-7 所示。

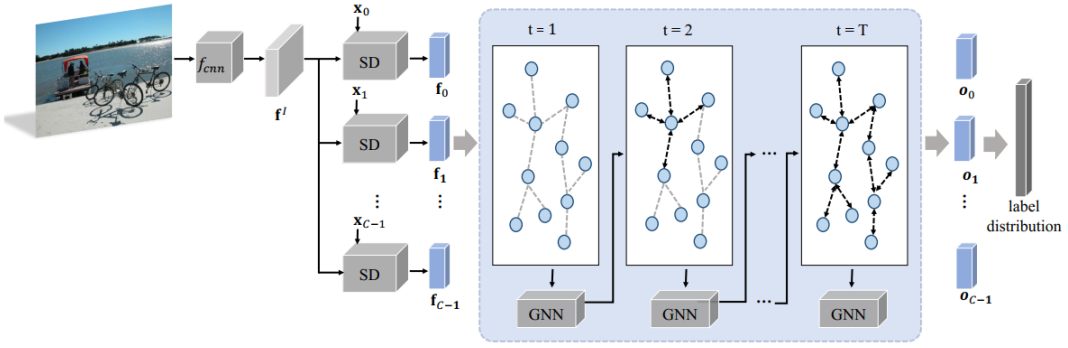


图 2-7 SSGRL 模型

该模型首先使用 CNN 基础网络提取图像特征，接着使用经过 Glove[49]预训练得到的标签词嵌入向量和低秩双线性池化[50]对图像特征进行特征解耦，将图像特征解耦为特定于类别的特征向量。给定输入图像 I，经过 CNN 网络得到图

像特征向量 f^I，然后输入每个标签的词嵌入向量 x_c，接着利用低秩双线性池化融合相应位置的图像特征 f_{wh}^I 和词向量 x_c：

$$\tilde{f}_{c,wh}^I = P^T(\tanh(U^T f_{wh}^I) \odot (V^T x_c)) + b \tag{2-20}$$

其中 P、U 和 V 均为模型需要学习的权重参数，b 表示偏置。然后获取每个像素点对应类别的注意力系数：

$$\tilde{a}_{c,wh} = \varphi_a(\tilde{f}_{c,wh}^I) \tag{2-21}$$

得到所有位置的注意力系数后，对其进行归一化：

$$a_{c,wh} = \frac{\exp(\tilde{a}_{c,wh})}{\sum_{w',h'} \tilde{a}_{c,w'h'}} \tag{2-22}$$

最后，对所有位置执行加权平均池化，以获得特征向量：

$$f_c = \sum a_{c,wh} f_{c,wh} \tag{2-23}$$

其中，f_c 表示特定于类别 C 的特征向量。然后将每个特定于类别的特征向量 f_c 作为图节点的特征向量，而图节点之间边的构造方式与 ML-GCN 相同。为了学习标签相关的特征表示，SSGRL 采用了门控图神经网络来进行特征交互，如下所示：

$$z_c^t = \sigma(W^z a_c^t + U^z h_c^{t-1}) \tag{2-24}$$

$$r_c^t = \sigma(W^r a_c^t + U^r h_c^{t-1}) \tag{2-25}$$

$$\tilde{h}_c^t = \tanh(W a_c^t + U(r_c^t \odot h_c^{t-1})) \tag{2-26}$$

$$h_c^t = (1 - z_c^t) \odot h_c^{t-1} + z_c^t \odot \tilde{h}_c^t \tag{2-27}$$

上述公式中的 W^z、U^z、W^r、U^r、W 和 U 均是可学习的权重参数。

2.4 复合表情识别介绍

经过深入的研究与分析，发现复合表情识别问题也是一种多标签图像分类问题，并且使用单标签标注复合表情图像获得的识别效果并不理想，于是本文在多标签图像分类的研究基础上，将复合表情图像转换成多标签图像，从而将

其转换成多标签图像分类任务，然后引入多标签图像分类模型进行复合表情识别。于是，接下来本节将对表情识别的相关知识进行详细介绍。

人脸表情识别任务是指给定一张人脸表情图像，让计算机能够准确地识别出该图像所表达的情感状态。复合人脸表情和基本人脸表情的表情识别步骤并没有太大的区别，主要的不同点在于它们所包含的表情类别数量不同。复合表情图像可能同时包含多个基本表情标签。人脸表情识别的步骤如下图 2-8 所示。

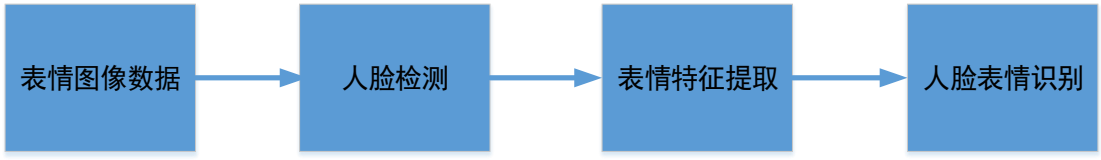


图 2-8 人脸表情识别流程

（1）人脸检测。由于在自然场景下收集到的表情样本通常包含许多噪声，这些噪声会对表情识别的准确率有一定的影响。为了降低这些噪声对识别性能的影响，通常会对输入的表情样本进行预处理操作。人脸检测这个步骤就是为了去除掉会对人脸表情识别造成干扰的背景区域。该步骤主要就是在给定的人脸表情样本中找到人脸所作的区域，然后将包含人脸的区域划分出来，剔除掉与人脸无关的背景区域。

（2）表情特征提取。在人脸表情识别任务中，特征提取是一个非常重要的步骤。不同的特征提取方法会提取到不同的人脸表情特征，这些特征会对最终的分类结果产生影响。因此，人脸表情特征提取在人脸表情识别中起着至关重要的作用，许多研究者都致力于对其进行优化和改进。

（3）表情分类。表情分类是人脸表情识别中的最后一步，它的目的是对从表情图像中提取出的表情特征进行分类，从而预测该表情图像所属的表情标签。

2.4.1 人脸检测

人脸检测是的主要目的就是找到表情图像中人脸所在的区域。因为人脸表情样本中包含许多干扰因素，只有通过人脸检测准确找到人脸所在的位置，才能够去除非面部区域和背景区域，从而保证提取到有效的人脸表情特征。

在人脸检测的早期，研究者通常使用模板匹配方法[51]来检测图像中人脸所在的区域。该方法使用标准人脸模板来描述面部特征，并计算输入图像子窗口与模板之间的相关值。然后，将计算得到的相关值与设定的阈值进行比较，以

检测输入图像中是否存在人脸。模板匹配技术优点是实现比较简单，但是缺点也很明显，即并不适用于所有场景。

经典的基于统计模型的方法有 Viola-Jones 算法[52]。该算法使用 Haar-like 特征对人脸进行描述，然后通过 AdaBoost 算法的思想，训练多个弱分类器组成一个强分类器，并将多个强分类器级联在一起，形成最终的用于人脸检测的分类器。该方法不仅取得了很好的检测效果，还提高人脸检测的速度。因此，Viola-Jones 算法在深度学习运用于人脸检测之前一直是主流的人脸检测方法。

虽然相比于基于模板的方法，Viola-Jones 算法在人脸检测任务中有了很大的效果提升，但其稳定性较差，并且该方法在一些复杂场景下进行人脸检测所取得的效果不是很理想。随着深度学习在人工智能这个领域中不断展现出它的能力，研究者们开始提出将深度学习引入人脸检测领域，于是人脸检测领域有了全新的发展方向。基于深度学习的方法不仅能够提升稳定性，还可以增强人脸检测的精度。如 2015 年，Li 等人[53]通过对 Viola-Jones 算法进行改进， 提出了一个新的模型 Cascade CNN。该模型通过引入深度神经网络来解决开放场景中进行人脸检测时会出现的一些问题。并且 Cascade CNN 模型同样采用了级联的思想，通过使用三个不同的网络模块构建了一个三阶级联的网络。通过这种方式，网络的最初阶段可以设置的较为简单，可以在保持较高召回率的同时过滤掉大量的非人脸窗口，从而降低网络的负担和提升整体检测速度，并获得高质量的人脸框定位。Zhang 等人[54]提出了一个深度级联多任务框架（MTCNN），旨在解决人脸检测和关键点检测问题。该模型包含三个不同的网络结构。MTCNN 模型采用了级联的思想，每个网络结构都有不同的目标和损失函数，并且通过多任务学习的方式共同优化。该模型能够快速、准确地检测图像中的人脸和关键点位置。

2.4.2 表情特征提取

表情特征提取是表情识别的关键步骤之一，因为提取到的表情特征的质量直接影响着表情识别的准确率。人脸面部表情的变化会导致面部肌肉也出现一定的变化，而人脸表情特征提取的目的就是通过相关的算法从给定的人脸面部表情样本中提出那些有效的表情特征，然后使用分类器对这些提取到的表情特征进行分类。可以按照表情特征提取方式的不同，分为传统的手工提取特征和基于深度学习的特征。

（1）手工提取的特征

传统的手工特征可以进一步细分为基于纹理的特征和基于几何的特征。其中，基于纹理的特征中 Gabor 小波法[54]和局部二值模式（LBP）[56]是两种最常用的方法。Gabor 小波法虽然能够有效地提取图像的纹理特征信息，但是该算法也存在一定的缺陷，该方法计算得到的特征维数过高以及数据存在冗余。LBP 算法可以对图像的灰度变化和角度随机性具有更好的鲁棒性，因此在提取人脸表情特征方面表现较好。然而，该方法的缺点在于容易受到噪声的干扰，并且其产生的高维直方图对算法的实时性有一定的影响。在基于几何特征的提取方法中，其中经典的方法有 SIFT（尺度不变特征变换）[57]。它通过在不同的尺度空间中对关键点位置进行局部描述，可以在人脸特征中获得较高的区分性。

（2）基于深度学习的特征

由于深度学习在图像分类研究领域取得了巨大的成就为表情识别研究提供了一种新的思路，许多研究人员开始将深度神经网络应用到表情识别领域中。不同于传统的手工提取特征，深度神经网络可以通过学习提取图像的特征，不仅避免了复杂的手工提取特征过程，还能够提取一些手工提取方法无法提取的关键特征。Veena Mayya 等人[58]提出使用深度卷积网络用于识别面部表情。由于 GPU 的使用，人脸图像特征提取时间显著减少，并且在当时实现了最高的识别率。Mollahosseini 等人[59]设计了一个更深层次的神经网络，通过使用 Inception 结构来识别人脸表情。除了单一模型，人们还采用多网络集成的方法。Liu 等人[60] 分别训练了三个不同卷积层的神经网络。然后对三个子网络的输出进行平均，得到最终的预测值。Connie 等人[61]通过将传统的 SIFT 特征与 CNN 从图像中提取的特征相结合，以进一步提高了表情识别的准确率。

2.4.3 表情分类

表情分类是表情识别任务过程中最后一个步骤。该步骤主要任务是使用分类器对通过模型学习到的人脸表情图像特征进行识别，将其准确划分到对应的人类表情类别。常用的机器学习分类方法有 SVM、KNN 和 HMM 等。SVM[62]利用核函数将低维的数据映射到高维空间，以获得最优的分类平面。KNN 算法[63]是通过寻找样本空间中 K 个最相似的样本来进行分类或回归，K 值的选择、距离度量和分类决策规则是非常重要的三个要素。而KNN算法的缺陷主要有两个，一是算法的时间、空间复杂度较高，特别是在处理大规模数据时，效率较低；二是 K 值的选取会影响算法的分类结果，不同的 K 值可能会得到不同的分类结果。

不同于传统的机器学习的分类方法，基于深度学习的分类方法是可以与提取特征同时进行的。基于深度学习的分类方法其实就是由多个全连接层组成的分类器。通过卷积网络提取得到图像特征后输入到全连接层中进行分类，然后使用激活函数计算该图像被认为归属于每个表情类别的概率，最后计算模型分类的损失并通过最小化损失函数优化分类器，从而得到一个分类效果很好的分类器来执行从而执行人脸表情识别任务。

对于 KNN 算法而言，给定的数据量会对其性能产生影响，因为 KNN 算法是一种基于距离度量的算法，需要通过计算样本之间的距离来确定最近邻的样本。因此，如果给定的数据量不足，可能会导致算法无法准确地找到最近邻，从而影响算法的性能。而对于基于深度学习的分类方法而言，训练的数据量同样会对分类器的性能造成影响。因此，在实际的运用中，我们需要针对所在的应用场景，来选择出适合该场景的最佳的分类器，以提高模型的识别性能。

2.5 本章小结

本章主要是对于本文工作涉及到的一些相关知识的介绍。由于本文引入图注意力网络自适应地挖掘标签之间的相关性，于是本章首先对图卷积网络的一些概念以及理论进行详细介绍，并且介绍了几种经典的图卷积网络模型。然后本章对当前主流的基于图神经网络的多标签图像分类算法中经典具有代表性的方法如 ML-GCN、SSGRL，详细介绍了它们的模型以及这些方法的核心思路。在本章的最后，我们还详细介绍了复合表情的相关知识，包括表情识别过程中的三个关键步骤。

3 基于 GAT 和多尺度语义注意力机制的多标签图像分类模型

3.1 引言

由于在现实世界中同时出现的目标对象之间通常存在依赖关系，因此可以通过挖掘目标标签之间的依赖关系来提升模型的分类准确率。并且由于多标签图像中包含各种尺度大小的目标对象，在提取特征的过程中小目标的特征信息可能会丢失，导致小目标在多标签识别中的性能往往较低，特别是对于具有挑战性的数据集。针对上述的两个问题，本章提出了一种新的多标签图像分类模型 MSSGAT。本章首先介绍 MSSGAT 的模型框架，并分别详细介绍组成模型的三个关键模块和损失函数。然后在两个多标签图像数据集上对本章提出的模型进行测试，并将实验结果与其他现有的模型进行对比分析。

3.2 MSSGAT 模型

3.2.1 模型总体框架

由于多标签图像通常包含多个目标对象，因此在进行多标签图像分类时，需要考虑对多个标签进行分类。这样导致的一个问题是，输出组合的数量呈指数形式增加，使得多标签图像分类相对于单标签图像分类更具有难度。因为同时出现的对象之间可能存在一定的联系，所以多标签图像分类可以通过挖掘多个标签之间的相关性来辅助模型进行预测，从而提高模型的分类准确率。于是，本章通过使用图注意力网络来自适应地挖掘标签之间的依赖关系，提出了一个新的多标签图像分类模型 MSSGAT。主要思想是借助图注意力网络中的自注意力机制自适应地学习每张图像中各个类别的相关性权重，更好地捕获和利用类别之间的相关性学习标签相关的特征，更好地进行结果预测。

本章提出的 MSSGAT 模型主要包含特征提取模块、多尺度语义注意力模块和多头图注意力模块这三个关键模块。整个网络模块的框架如图 3-1 所示。在特征提取模块中，本章选用 ResNet-101[64]来提取图像特征，并选取最后三个卷积层的输出作为该模块的输出。在多尺度语义注意力模块中，本文将三个不同尺度的图像特征图进行特征融合，然后利用词嵌入向量指导学习特定于各个类

别的特征向量。在多头图注意力模块中，利用图注意力层的自注意力机制根据各个特定于类别的特征向量自适应地挖掘标签之间的依赖关系，得到最终包含标签相关信息的判别向量用于分类。

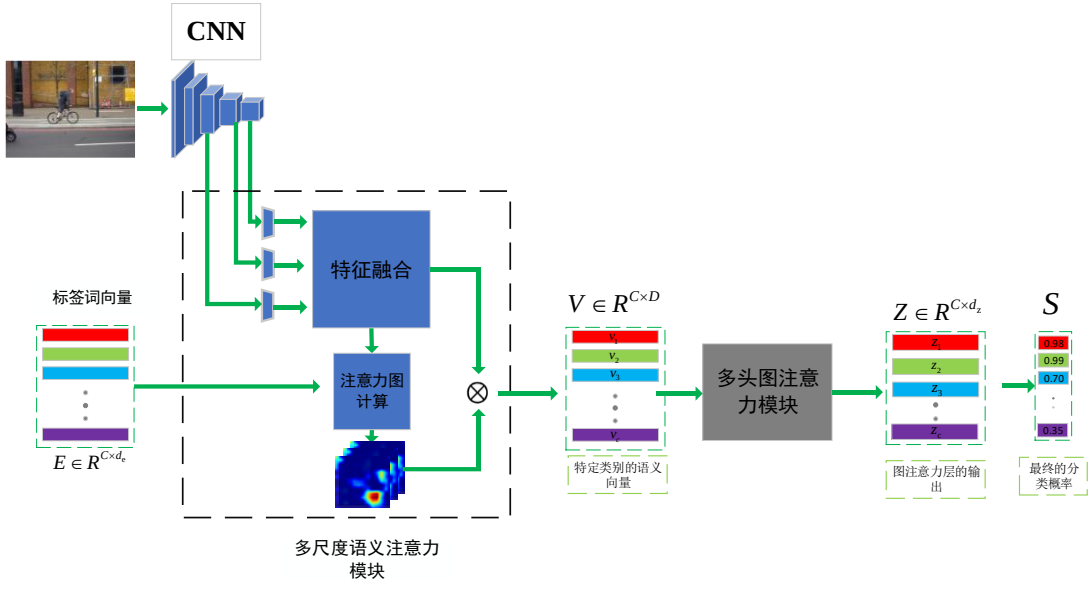


图 3-1 MSSGAT 模型图

3.2.2 特征提取模块

通过对现在主流的多标签分类模型的研究，本文使用 ResNet-101[64]作为 MSSGAT 模型的特征提取主干网络。因为 ResNet 兼具简单与性能优越两大优点，所以在现在的计算机视觉领域中很多任务都纷纷使用 ResNet 来提取图像特征。

自从深度卷积神经网络提出之后，很多人都认为，搭建的神经网络越深，模型获取到的有用信息就越多，提取到的图像特征就越丰富，也就能够提升模型的性能。可是根据实验表明，当卷积网络达到一定深度后，继续加深网络不仅不会使模型的性能得到提升，反而会使得模型再训练过程中难以收敛，导致模型准确率下降。因为不断加深网络会导致了梯度消失和梯度爆炸。针对这个问题，何凯明等人提出了一种深度残差网络，也就是本文所使用的 ResNet，它可以使模型在尽可能加深的同时不会降低模型的性能。

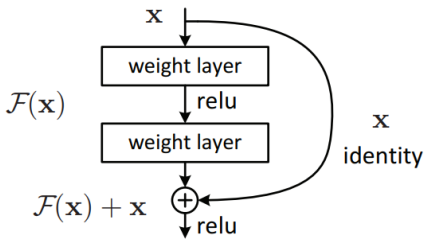


图 3-2 残差构建块

残差网络中提出了两种映射方式：恒等映射和残差映射。恒等映射是指输入和输出的特征图大小和通道数都相同，残差映射则是对输入进行一定的变换后再与输出相加，从而得到最终的特征图。如图 3-2 所示，右侧标有 *X* 的曲线代表着恒等映射，而残差指的则是 *F(X)*部分，最后的输出则是 *F(X)+X*。

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

图 3-3 ResNet 结构图

如图 3-3 所示，按照网络层数的不同可分五种 ResNet。并且观察各个卷积层中的残差块数量可以发现，ResNet-18/34 这两种网络的卷积块是由两层卷积组成的，而 ResNet-50/101/152 这三种网络使用的是三层卷积的卷积块，分别对应图 3-4 中两种不同设计方式的残差块。ResNet-18 和 ResNet-34 采用的是图 3-4 左侧的残差块，每个残差块由两个 3x3 的卷积层组成。而 ResNet-50、ResNet-101 和 ResNet-152 采用的是图 3-4 右侧的残差块，每个残差块由一个 1x1 的卷积层、一个 3x3 的卷积层和一个 1x1 的卷积层组成。之所以在 ResNet-50/101/152 网络中使用的残差块与 ResNet18/34 不同，是为了减少计算和参数量。

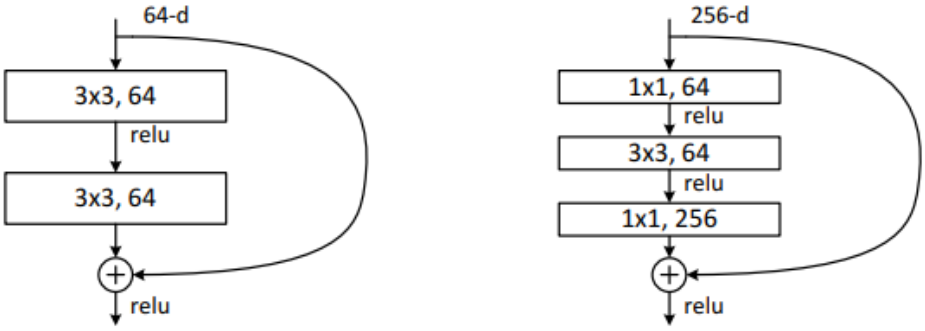


图 3-4 深度残差函数

在本章提出的模型中，我们选取 Res Net-101 作为模型的特征提取模块用于提取图像特征，从图 3-3 中可以看到，Res Net-101 的整个模型包含 5 个卷积层，其中第一层是 7*7 的卷积，然后是四个卷积层，其中每个卷积层包含了多个“bottleneck design”结构的残差块，网络的最后是一个全局平均池化层、一个全连接层和一个 Softmax 分类层。

对于模型提取图像特征，给定一张多标签图像样本 I，然后选取最后三个卷积层的输出，作为该模块的 feature map，如下所示：

$$X_1, X_2, X_3 = F_{cnn}(I; \theta_{cnn}) \tag{3-1}$$

其中 $F_{cnn}(\bullet;\theta_{cnn})$ 表示整个特征提取网络，$X_1 \in R^{W_1 \times H_1 \times N_1}$、$X_2 \in R^{W_2 \times H_2 \times N_2}$、$X \in R^{W_3 \times H_3 \times N_3}$ 分别三个不同尺度的特征图，θ_{cnn} 表示模型需要学习的参数。

3.2.3 多尺度语义注意力模块

为了使模型更好地提升对图像中小目标的分类性能，提出一个多尺度语义注意力模块通过融合多个尺度的特征图来增强小目标的特征信息，并使用标签语义指导学习图像特征图中特定于类别的特征向量，通过使用注意力机制可以使模型更好地学习到与类别标签相关的图像特征。

首先，得到特征提取模块输出的三个特征图，分别为 X_1，X_2，X_3，并将其同时输入到多尺度语义注意力模块中。然后，将三个尺度的特征图融合，用于增强小目标的特征信息。如下所示：

$$\hat{X} = \omega_1(\omega_2(X_1), \omega_3(X_2), X_3) \tag{3-2}$$

其中 ω_1，ω_2 和 ω_3 是三个卷积核大小不同的卷积层。

在得到多尺度融合特征 $\hat{X}$ 后，本文采用与 SSGRL[34]类似的语义注意力机制使用 d_e 维标签词嵌入向量 $word \in R^{C\times d_e}$ 对图像特征进行特征分解。过程如下所示：

$$E = \Phi_t(word) \in R^{C\times D} \tag{3-3}$$

$$a^i_{(w,h)} = \sigma(\hat{X}_{w,h} \bullet e_i^T) \tag{3-4}$$

$$v_i = \frac{\sum_{w,h} a^i_{w,h} \hat{X}_{w,h}}{\sum_{w,h} a^i_{w,h}} \tag{3-5}$$

公式(3-3)中 Φ_t 为线性转换函数，将词嵌入向量特征维度转换为 D 维。接着本文引入了语义注意机制。该机制使用标签语义向量 E 来引导学习与该类别相关的特征向量。对于特征图中的每个位置，本文使用标签语义向量来计算注意力权重，如公式(3-4)所示。其中，e_i 表示第 i 个类别的标签语义向量，$\hat{X}_{w,h}$ 中的 w, h 表示特征图像素点的位置，$\bullet$ 表示点乘操作。$a^i_{(w,h)}$ 表示位置 (w,h) 与第 i 个类别的之间的注意力权重，$\sigma(\bullet)$ 是 Sigmoid()激活函数。经过语义注意力计算得到所有位置与类别 i 对应的权重 $a^i_{(w,h)}$，然后进行加权平均以获得第 i 个类别的特征向量，如公式(3-5)所示。其中 v_i 表示特定于类别 i 的特征向量。通过多尺度语义注意力模块最终得到一组特定于类别的多尺度语义特征向量 $V = \{v_1, v_2, ...v_c\}$。

3.2.4 多头图注意力模块

在现实世界中同时出现的对象之间可能存在一定的关联和依赖关系，于是本章提出了使用图结构来建模标签之间的依赖关系，并引入了图注意力网络来自适应地挖掘标签之间的依赖关系。在经过多尺度语义注意力模块得到每张图像中特定类别的特征向量 $V = \{v_1, v_2, ...v_c\}$ 之后，我们将这些向量作为图的节点，并引入图注意力模块进行特征交互。充分利用图注意力模块来学习标签之间的相关性，生成最终具有标签相关信息的判别向量 $Z = \{z_1, z_2, ...z_c\}$。详细的模块框架图如图 3-5 所示。